Cost-Effective Vertical Federated Learning for Multi-Platform Collaborative Recommendation

Wenjie Li*, Shutao-Xia#
Shenzhen International Graduate School, Tsinghua University
liwj20@mails.tsinghua.edu.cn, xiast@sz.tsinghua.edu.cn

I. INTRODUCTION

Recommendation systems are now ubiquitous, capturing user behaviors across various service platforms that reflect diverse user interests. Using this multi-platform data collectively can achieve comprehensive user modeling. However, intrinsic data isolation, privacy laws (GDPR [1]), and commercial confidentiality make direct data sharing impossible, a problem well-known as the "***isolated data island problem***" [2]. To tackle this problem, vertical federated learning [3]–[6] has been proposed and explored in various recommendation tasks [7]–[11]. However, due to the necessity of conducting cross-agency data intersection [12] before training and the distributed nature of the split model, most of existing works suffer from the following three challenges:

- ***(C1)*** **Diminished Training Data Scope**: Aligned users for dissimilar businesses are often limited and constitute only a small portion of the user population. This reduced training set size can increase the risk of overfitting and result in low-quality embeddings and hidden representations, especially in sparse high-dimensional recommendation datasets.
- ***(C2)*** **Limited User Group Benefits**: The intrinsic field missing in passive parties makes it infeasible for a federated model to train on or make predictions for unaligned users. Thus, vanilla VFL can only bring benefits to aligned users, largely undermining the practicability of VFL. If a participant has more unaligned users, or places greater emphasis on unaligned users in their business, joining the federation is not cost-effective.
- ***(C3)*** **Costly Federated Inference**: The inference process of VFL models incurs extra time costs (due to cross-agency feature transmission and security enhancement operations) and poses new system design challenges (arising from inconsistent network conditions and computational capabilities of different parties). These challenges make it difficult for a federated inference system to meet the high throughput and real-time latency requirements of advertising systems (million-wise peak QPS, 100$\sim$100ms processing time per request [13], [14]). These obstacles may render the federation infeasible or excessively costly for participants.

Our Contribution To address these challenges, we have integrated cutting-edge machine learning techniques such as

*This series research is supported by Tencent Rhio-Bird talent project and Meituan research collaboration project. #PhD Supervisor.

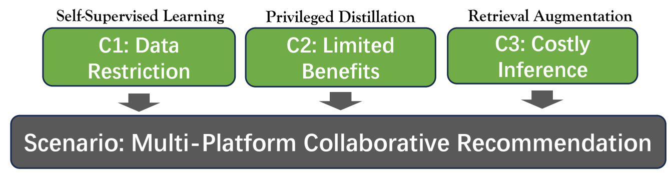

Fig. 1. The overview structure of my research.

self-supervised learning [15], ***privileged distillation*** [16], and ***retrieval augmentation*** [17], all tailored for recommendation systems under a VFL setting, to enhance VFL's practicality. Our methods show encouraging outcomes on both public and industry datasets.

Basic Learning Framework We focus on the two-party VFL setting [5], [6], where an *active party* A, holding the labels and some attributes, collaborates with a *passive party* B, who provides additional attributes to train a distributed federated model for a specific task. The federated model consists of the *bottom model* held by each party and the *top model* held by the active party: $\hat{y}_{fed} = g_A(f_A(\mathbf{x}_A), f_B(\mathbf{x}_B))$. where f denotes the bottom model and g denotes the top model, $\mathbf{x}$ denotes inputs from parties. More details can be found in [15].

II. PROGRESS A: EXTENDING DATA SCOPE WITH SELF-SUPERVISED LEARNING

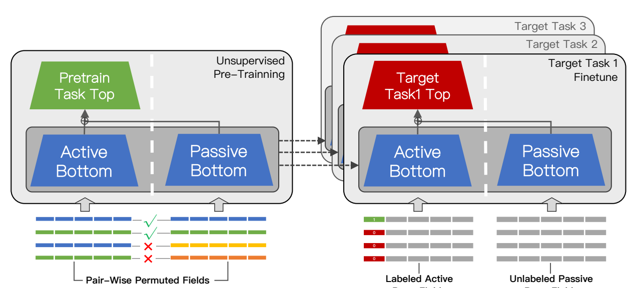

Fig. 2. The overall framework of VFL-MPD.

This work aims to tackle challenge (C1). We argue that massive historical records with outdated labels [18] in advertising systems could be useful for representation learning. By incorporating these massive unlabeled data in VFL, we can compensate for the shortage of overlapped data.

Therefore, we developed the first VFL-tailored self-supervised task, matched pair detection (**MPD**), to utilize these unlabeled data. We use MPD to learn a pre-training splitNN model and employ its bottom model to initialize downstream task models. Intuitively, the MPD task is to learn a binary classifier to distinguish whether input attributes from two parties match. All original overlapped samples are positive samples, and we construct the negatives by frequency-based random sampling [19]. MPD has a intrinsic connection to ***maximizing mutual information***, that is:$g_A(\mathbf{h}_A, \mathbf{h}_B) = PMI(\mathbf{x}_A, \mathbf{x}_B) - \log k$. This reveals that the top model implicitly models the point-wise mutual information (PMI) of the observed input pairs, with a shifted constant $\log k$. Such a learning principle strongly supports the MPD pre-training task in learning effective cross-party representations. Our experiments, conducted on two industry datasets from Tencent and one simulated public dataset, validate MPD's superiority, with a 2 to 10 thousandths AUC improvement compared to naive self-training.

III. Progress B: Reducing Inference Cost with Privileged Distillation

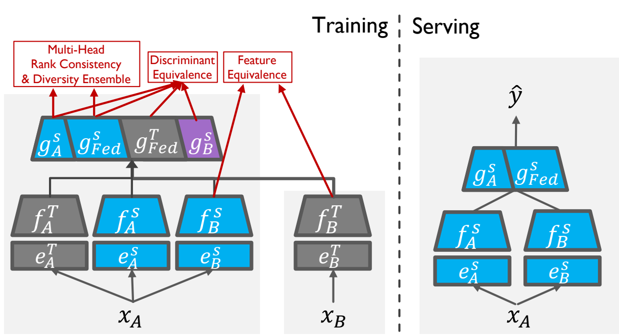

Fig. 3. The overview of the joint privileged learning framework.

In order to jointly tackle challenges (C1)–(C3), we investigate a lightweight and practical problem setting, Semi-VFL (Vertical Semi-Federated Learning), which utilizes the full sample set during training and achieves standalone local inference. It is a relaxed setting where the active party cannot achieve real-time inference for distributed models but can for local models. There are two key challenges to achieving Semi-VFL: ***1) effective passive party fields-free inference*** and ***2) integrating distribution bias*** *between overlapped and non-overlapped sets*. To address these challenges, we propose the two-stage Joint Privileged Learning framework (**JPL**). The first stage is federated teacher training, extracting knowledge from the full attribute set on overlapped data. The second stage is joint privileged distillation, where the model jointly uses all data to learn an input-restricted student model aimed at efficient local serving. Specifically, JPL consists of learning components and objectives:

- **Learning components**: The model is composed of multiple classifier heads and a cross-domain feature encoder. The former is responsible for capturing discriminative patterns from different input signals (e.g., from a-side only, b-side only, or both). The latter is designed to learn cross-party feature correlations, thus alleviating the field missing problem in the inference stage.
- **Learning objectives**: In terms of tackling field missing, we adopt a prediction-based Discrimination Equivalence and contrastive-based Feature Equivalence on the cross-party feature encoder. For cross-set bias, we use multi-head ranking consistency regularization and multi-head Diversity Ensembling. Readers can refer to [16] for more details.

As a consequence, JPL consistently outperforms basic federated distillation approaches [15], [20] on various data settings on two public Click-Through Rate (CTR) prediction datasets.

IV. Progress C: Achieving Full Set User Benefit with Retrieval Augmentation

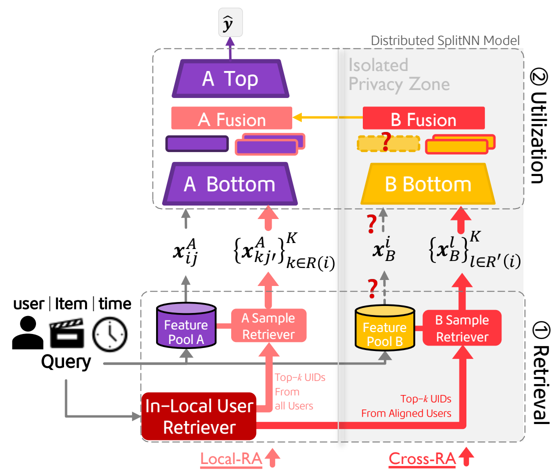

Fig. 4. The overall framework of ReFer.

In order to tackle challenges (C1) and (C2), we propose a retrieval-enhanced approach named **ReFer** (as depicted in Figure 4). We focus on the Fully Vertical Federated Recommendation (**Fully-VFR**) problem, which is similar to Semi-VFL but assumes that participants are capable of conducting online federated serving. The design of ReFer revolves around achieving two types of retrieval augmentation (RA) strategies in a distributed and privacy-preserving manner: ***1) cross-party RA for field missing*** and ***2) in-local RA to mitigate cross-group user bias***. Specifically, ReFer is two-staged:

- **Federated Retrieval Augmentation:** We design a federated retriever that enhances each active party's sample with K relevant samples from both parties. The retriever is designed hierarchically with a two-stage user-item structure to ensure privacy and efficiency in the VFL environment.
- **Federated Retrieval Utilization:** The fusion modeling process learns a retrieval-oriented fusion representation for the query sample and uses it to promote better predictions.

As a result, experiments on both sequential and non-sequential CTR prediction tasks show that ReFer achieves the best AUC performance over baselines in 9 VFL scenarios and is beneficial for all user groups.

V. Future Directions

In the future, we plan to further extend this work to more types of recommendation scenarios, exploring the possibility of combining it with generative recommendation models and large language models (LLMs), and to further investigate their security and privacy concerns.

References

[1] P. Voigt and A. Von dem Bussche, "The eu general data protection regulation (gdpr)," *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, vol. 10, no. 3152676, pp. 10–5555, 2017.

[2] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *TIST*, vol. 10, no. 2, pp. 1–19, 2019.

[3] Y. Liu, Y. Kang, T. Zou, Y. Pu, Y. He, X. Ye, Y. Ouyang, Y.-Q. Zhang, and Q. Yang, "Vertical federated learning," *arXiv preprint arXiv:2211.12814*, 2022.

[4] K. Wei, J. Li, C. Ma, M. Ding, S. Wei, F. Wu, G. Chen, and T. Ranbaduge, "Vertical federated learning: Challenges, methodologies and experiments," *arXiv preprint arXiv:2202.04309*, 2022.

[5] P. Vepakomma, O. Gupta, T. Swedish, and R. Raskar, "Split learning for health: Distributed deep learning without sharing raw patient data," *arXiv preprint arXiv:1812.00564*, 2018.

[6] I. Ceballos, V. Sharma, E. Mugica, A. Singh, A. Roman, P. Vepakomma, and R. Raskar, "Splitnn-driven vertical partitioning," *CoRR*, vol. abs/2008.04137, 2020.

[7] L. Yang, B. Tan, V. W. Zheng, K. Chen, and Q. Yang, *Federated Recommendation Systems*, pp. 225–239. Cham: Springer International Publishing, 2020.

[8] M. Huang, H. Li, B. Bai, C. Wang, K. Bai, and F. Wang, "A federated multi-view deep learning framework for privacy-preserving recommendations," *arXiv preprint arXiv:2008.10808*, 2020.

[9] F. Fu, H. Xue, Y. Cheng, Y. Tao, and B. Cui, "Blindfl: Vertical federated machine learning without peeking into your data," in *Proceedings of the 2022 International Conference on Management of Data*, pp. 1316–1330, 2022.

[10] Y. Hu, D. Niu, J. Yang, and S. Zhou, "Fdml: A collaborative machine learning framework for distributed features," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2232–2240, 2019.

[11] T. Chen, X. Jin, Y. Sun, and W. Yin, "Vafl: a method of vertical asynchronous federated learning," *arXiv preprint arXiv:2007.06081*, 2020.

[12] D. Peterson, P. Kanani, and V. J. Marathe, "Private federated learning with domain adaptation," *arXiv preprint arXiv:1912.06733*, 2019.

[13] J. Shen, B. Orten, S. C. Geyik, D. Liu, S. Shariat, F. Bian, and A. Dasdan, "From 0.5 million to 2.5 million: Efficiently scaling up real-time bidding," in *ICDM*, pp. 973–978, IEEE, 2015.

[14] Y. Yuan, F. Wang, J. Li, and R. Qin, "A survey on real time bidding advertising," in *Proceedings of 2014 IEEE International Conference on Service Operations and Logistics, and Informatics*, pp. 418–423, 2014.

[15] W. Li, Q. Xia, J. Deng, H. Cheng, J. Liu, K. Xue, Y. Cheng, and S.-T. Xia, "Semi-supervised cross-silo advertising with partial knowledge transfer," *arXiv preprint arXiv:2205.15987*, 2022.

[16] W. Li, Q. Xia, H. Cheng, K. Xue, and S.-T. Xia, "Vertical semi-federated learning for efficient online advertising," *arXiv preprint arXiv:2209.15635*, 2022.

[17] W. Li, Z. Wang, J. Wang, S.-T. Xia, J. Zhu, M. Chen, J. Fan, J. Cheng, and J. Lei, "Refer: Retrieval-enhanced vertical federated recommendation for full set user benefit," in *SIGIR*, 2024.

[18] X. Wu, Q. Liu, J. Qin, and Y. Yu, "Peerrank: Robust learning to rank with peer loss over noisy labels," *IEEE Access*, vol. 10, pp. 6830–6841, 2022.

[19] O. Levy and Y. Goldberg, "Neural word embedding as implicit matrix factorization," *NeurIPS*, vol. 27, 2014.

[20] Z. Ren, L. Yang, and K. Chen, "Improving availability of vertical federated learning: Relaxing inference on non-overlapping data," *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2022.