Front. Comput. Sci., 2025, 19(1): 191603
https://doi.org/10.1007/s11704-024-3936-0

LETTER

Towards practical data alignment in production federated learning

Yexuan SHI[1,2*], Wei YU[3*], Yuanyuan ZHANG (✉)[1], Chunbo XUE[1,2], Yuxiang ZENG[1], Zimu ZHOU[4], Manxue GUO[3], Lun XIN[3], Wenjing NIE[3]

1 State Key Laboratory of Complex & Critical Software Environment and Advanced Innovation Center for Future Blockchain and Privacy Computing, Beihang University, Beijing 100191, China
2 Zhongguancun Pan Connected Mobile Communication Technology Innovation and Application Research Institute, Beijing 100088, China
3 China Mobile Research Institute, Beijing 100053, China
4 School of Data Science, City University of Hong Kong, Hong Kong 999077, China

1 Introduction

Federated learning has emerged as a promising par-adigm for collaborative model training that facilitates cooperation among multiple parties while ensuring data privacy [1]. Successful alignment of data across parties is crucial for effective federated learning [2]. This alignment involves harmonizing heterogeneous data from different parties to identify shared data for joint model training.

Private set intersection (PSI) is a technique that allows the alignment of common entities between parties without revealing additional information. However, efficiently performing data alignment with PSI in federated learning [3], especially when dealing with highly unbalanced data, remains challenging due to the low efficiency.

In this paper, we propose a new data alignment solution called α-ESF. It incorporates α-indisting-uishability, a client-side privacy requirement metric, and Bucket-ESF, a novel PSI-oriented index, to achieve efficient data alignment. α-ESF operates in two phases: pruning and verification, which can significantly improve efficiency in cases of unbalanced dataset. More details can be found in supplementary material.

2 α-ESF overview

This section presents an overview of α-ESF, a practical framework for efficient data alignment even in case of extreme dataset unbalance.

Methodologies. The efficiency bottleneck of data alignment lies in the computation time of encoding on server-side dataset. Our solution breaks this bottleneck from two aspects.

(i) Exploit indexing to convert per-query server-side encoding overhead to one-off pre-processing. In practice, the server-side usually processes data alignment requests from multiple clients. This allows pre-computing the server-side encoding and sharing the results among clients to amortize the server-side encoding overhead.

(ii) Harness asymmetric privacy requirements to prune server-side dataset for data alignment. Real-world federated learning applications often impose less stringent privacy control on the client-side, which holds the potentials to prune server-side dataset.

Solution workflow. We implement the methodologies above via α-ESF, which consists of a one-off *secure index construction* step, and a two-phase processing framework (*asymmetric privacy based pruning* and *encryption based verification*) for data alignment (see Fig. 1).

- Secure index construction. We devise *Bucket-ESF*, a novel *two-level* PSI-oriented index that allows fast search while complying with privacy constraints. The *first level* partitions the server-side entire dataset into $|B|$ buckets. And the *second level* is an encrypted data structure EncSum Filter, which stores a ciphertext $\mathcal{E}(u)$

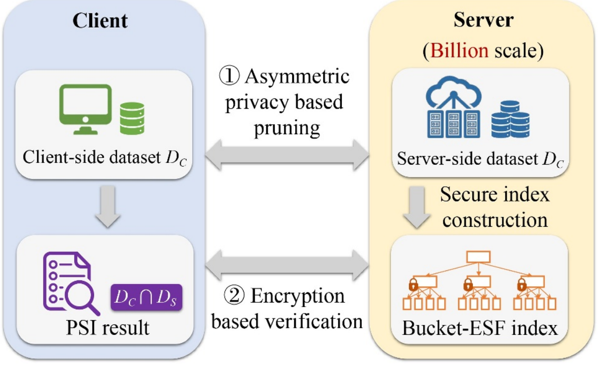
Client
Server
(Billion scale)
① Asymmetric privacy based pruning
Client-side dataset D_C
Server-side dataset D_C
Secure index construction
$D_C \cap D_S$
PSI result
② Encryption based verification
Bucket-ESF index

Fig. 1 Overview of α-ESF framework

Received November 21, 2023; accepted June 26, 2024

E-mail: zyy-buaa@buaa.edu.cn

* These authors contributed equally to this work.

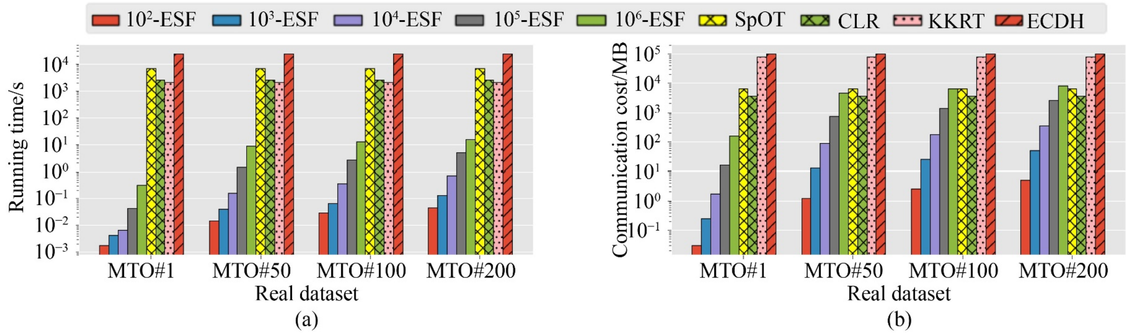

Fig. 2 Experimental results of varying α on real datasets. (a) Running time; (b) communication cost

into the array ESF according to three hash functions.

$$\mathcal{E}(u) = \bigoplus_{z=1}^{3} \mathsf{ESF}[\mathcal{H}_z(u)]. \tag{1}$$

The index construction is a one-off effort that serves multiple data alignment requests.

- Asymmetric privacy based pruning. Given a private dataset D_C held by a client C, we propose to apply α-*indistinguishability* as a metric to quantify the client-side privacy requirement. That is,

$$\forall s \in D_S, \Pr(s \in D_C) \leqslant \frac{1}{\alpha}. \tag{2}$$

On this basis, we design pruning strategies that reduce the server-side dataset for further processing while ensuring the α-indis-tinguishability of the client. This module is the core that detaches the server-side data size from execution time, which notably improves the efficiency of unbalanced data alignment.

- Encryption based verification. We propose a homomorphic encryption based verification procedure upon the pruned Bucket-ESF index. It returns the data alignment result to the client without revealing any extra information to either party. We further propose a ciphertext-compression technique, which can compress multiple elements into one ciphertext to reduce the encryption and communication overhead.

3 Experimental evaluation

Setups. We evaluate our solutions on a real dataset collected from a leading mobile telecom operator (MTO) in China. Our α-ESF is compared with existing PSI based data alignment solutions (KKRT [4], CM [5], SpOT [6], and CLR [7]).

Overall performance. We vary the client-side privacy requirement from 100 to 1 million to evaluate the performance of our α-ESF solutions. As shown in Fig. 2, our α-ESF notably outperforms all the baselines in running time. When the client-side privacy requirement is $\alpha = 10^6$, our α-ESF is 134× faster than the runner-up solution (KKRT). The communication cost of α-ESF is relatively sensitive to the privacy requirements. This is because our solution needs to transmit more EncSum Filters from the server to the client as α increases.

Performance on effectiveness. Figure 3 shows the model performance of vertical federated learning with our data alignment solution. By involving the server-side data via our data alignment solution, the performances of all clients and the federated model increase significantly. The accuracy gain of the data alignment based federated learning is up to 3.93%, which demonstrates the effectiveness of the data alignment in federated learning.

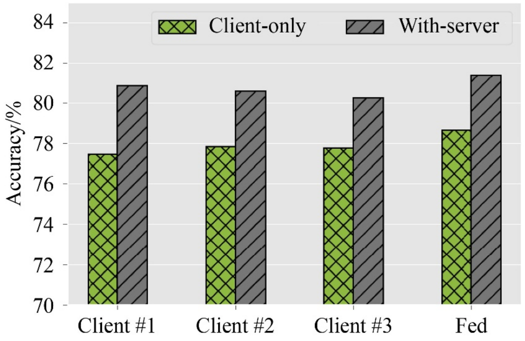

Fig. 3 Model performance with data alignment solution

4 Conclusion

In this paper, we propose α-ESF, an efficient data alignment algorithm for federated learning. It is featured with two core techniques: α-indistinguisha-bility-based server-side dataset pruning, and a novel PSI-oriented index for rapid search on server-side encryption. Extensive experiments show that our α-ESF is significantly faster than prior arts in case of billion-scale dataset unbalance, and it can improve the efficacy of federated learning.

Acknowledgements This work was partially supported by the National Natural Science Foundation of China (Grant Nos. U21A20516, 62076017, and 6233000216), the Beihang University Basic Research Funding (No. YWF-22-L-531), and the CCF-Huawei Populus Grove Fund (CCF-HuaweiDB202310).

Competing interests The authors declare that they have no competing interests or financial conflicts to disclose.

Electronic supplementary material Supplementary material is available in the online version of this article at journal.hep.com.cn and link.springer.com.

References

1. Yang Q, Liu Y, Chen T, Tong Y. Federated machine learning: concept and applications. ACM Transactions on Intelligent Systems and

Technology, 2019, 10(2): 12

2. Zhang K, Song X, Zhang C, Yu S. Challenges and future directions of secure federated learning: a survey. Frontiers of Computer Science, 2022, 16(5): 165817
3. Liu F, Zheng Z, Shi Y, Tong Y, Zhang Y. A survey on federated learning: a perspective from multi-party computation. Frontiers of Computer Science, 2024, 18(1): 181336
4. Kolesnikov V, Kumaresan R, Rosulek M, Trieu N. Efficient batched oblivious PRF with applications to private set intersection. In: Proceedings of 2016 ACM SIGSAC Conference on Computer and Communications Security. 2016, 818−829
5. Chase M, Miao P. Private set intersection in the internet setting from lightweight oblivious PRF. In: Proceedings of the 40th Annual International Cryptology Conference on Advances in Cryptology. 2020, 34−63
6. Pinkas B, Rosulek M, Trieu N, Yanai A. SpOT-light: lightweight private set intersection from sparse OT extension. In: Proceedings of the 39th Annual International Cryptology Conference on Advances in Cryptology. 2019, 401−431
7. Chen H, Laine K, Rindal P. Fast private set intersection from homomorphic encryption. In: Proceedings of 2017 ACM SIGSAC Conference on Computer and Communications Security. 2017, 1243−1255