Journal of Machine Learning Research 22 (2021) 1-6 Submitted 7/20; Revised 8/21; Published 8/21

FATE: An Industrial Grade Platform for Collaborative Learning With Data Protection

Yang Liu[1] liuy03@air.tsinghua.edu.cn
Tao Fan[2] dylanfan@webank.com
Tianjian Chen[3] tchenay@connect.ust.hk
Qian Xu[2,3] qianxu@webank.com
Qiang Yang[2,3,⋆] qyang@cse.ust.hk
[1] *Institute for AI Industry Research, Tsinghua University, Beijing, China*
[2] *AI Department of Webank, Shenzhen,China*
[3] *Hong Kong University of Science and Technology, Hong Kong*
[⋆] *Corresponding Author*

Editor: Alexandre Gramfort

Abstract

Collaborative and federated learning has become an emerging solution to many industrial applications where data values from different sites are exploit jointly with privacy protection. We introduce FATE, an industrial-grade project that supports enterprises and institutions to build machine learning models collaboratively at large-scale in a distributed manner. FATE supports a variety of secure computation protocols and machine learning algorithms, and features out-of-box usability with end-to-end building modules and visualization tools. Documentations are available at `https://github.com/FederatedAI/FATE`. Case studies and other information are available at `https://www.fedai.org`.

Keywords: federated learning, collaborative learning, secure multi-party computation, data protection, privacy-preserving

1. Introduction

A major challenge to artificial intelligence (AI) in real-world applications is how to bridge data silos and collaboratively build models while protecting user privacy. Despite the growing awareness of the legal use of data for AI and the value of integrating data silos (Kairouz et al., 2019), there is a lack of practical and high-performance platforms for enterprises to collaborate with each other on a production scale for enterprises to collaborate with each other. Existing open-sourced frameworks are mostly research-oriented and lack industrial-scale implementation. FATE (Federated AI Technology Enabler) is the first production-oriented platform developed by Webank's AI Department. Its goal is to support a collaborative and distributed AI ecosystem with cross-silo data applications while meeting compliance and security requirements.

2. Related Work

In contrast to traditional distributed learning, federated learning (McMahan et al., 2016; Yang et al., 2019) is proposed to tackle data locality and privacy in various cross-device and cross-silo scenarios, (Kairouz et al., 2019; Li et al., 2019) by allowing model updates or intermediate training results instead of raw data to be communicated among participants. Data protection protocols including Homomorphic Encrytion (HE), MultiParty Computation (MPC) and Differential Privacy (DP) (Dwork, 2006) are typically adopted for protecting data in transit. Depending on how data is partitioned, (Yang et al., 2019) categories federated learning into horizontal federated learning (HFL, or sample-partitioned FL) , vertical federated learning (VFL, or feature-partitioned FL) and federated transfer learning (FTL). Multiple open-sourced projects have emerged since then, including TensorFlow Federated [1], LEAF (Caldas et al., 2018), PySyft[2], Baidu's PaddleFL[3] and Clara Training Framework [4].

By "industrial-scale", we mean that FATE provides all the necessary components for production by design, including a truly distributed platform (Figure 1) supporting both standalone and cluster deployment with more than 30 concurrent enterprise participants and billions of concurrent samples. In comparison, PySyft and TensorFlow Federated started as research-oriented projects supporting only standalone simulations of multi-party collaborations. In June 2021, PySyft released version 0.5.0 including an integration with PyGrid to support federated mode [5]. In addition, FATE offers privacy-preserving XGBoost (called Secureboost (Cheng et al., 2021)), FTL (Liu et al., 2018) and a variety of feature engineering tools such as feature binning, feature Information Value (IV) computations which are essential to real-world applications.

3. Overview of FATE

FATE was developed at the AI department of Webank and was open-sourced in January 2019. As of its 1.6 release, FATE has 48 open-source community contributors and more than 3100 github stars. FATE community (FATE github, mailing list and Wechat Subscriptions) now has over 300 corporations and over 100 universities and institutions. In June 2019, FATE joined the family of Linux Foundation under the Apache 2.0 license. Over the years, FATE has been adopted in real-world applications in finance, health and recommender systems, summarized in Table 1. For example, in Ju et al. (2020) Webank and Tencent collaboratively developed a privacy-preserving Stroke Prediction model based on FATE and deployed it on Tencent's cloud server to allow multiple hospitals to select effective features and train models collaboratively.

1. `https://www.tensorflow.org/federated`
2. `https://github.com/OpenMined/PySyft`
3. `https://github.com/PaddlePaddle/PaddleFL`
4. `https://docs.nvidia.com/clara/tlt-mi/clara-train-sdk-v2.0/nvmidl/additional_features/federated_learning.html`
5. `https://github.com/OpenMined/PySyft/releases/tag/0.5.0`

Finance	Federated Data Network (https://fdn.webank.com/), (Zheng et al., 2020)
Medical	(Xiong et al., 2020; Ju et al., 2020)
Recommender systems	(Tan et al., 2020)
Speech Recognition	(Jiang et al., 2021)

Table 1: Summary of industrial applications

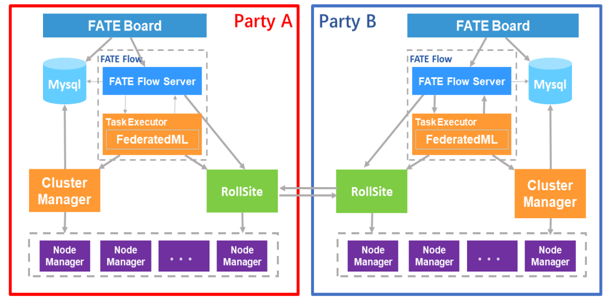

Figure 1: Basic Components of FATE

3.1 Core Features

The basic distributed architecture [6] is shown in Figure 1 . At its core, FATE is built on a library of federated and privacy-preserving machine learning algorithms, called **FederatedML**. Private Set Intersection (PSI) are provided to find the common users among parties. FATE enhances the computation efficiency with a distributed computation framework called **EggRoll**. FATE not only support linear and logistic regression (LR) and deep learning neural networks, but also tree-based algorithms such as XGBoost and transfer learning. FATE interfaces with users through three major components, **FATE-FLow**, a scheduling system that coordinates the execution of algorithmic components, **FATE-Board**, a visualization tool for building and monitoring the pipelines; and **FATE-Serving**, a high-performance inference platform customized with production need. **KubeFATE** [7] is developed by VMware to build FATE on top of Kubernetes in data centers, providing an enterprise-managed solution over distributed infrastructure and across organizations. Both Mac and Linux are supported for either manual or docker deployments. FATE also supports cross-cloud deployment and management through **FATE-cloud**.

3.2 Security And Utility

FATE adopts a security definition in which all parties are *honest-but-curious*. For HFL, FATE assumes a semi-honest server and ensures that the server only learn the aggregated parameters but not any individual's data. For VFL, parties exchange encrypted intermediate results and performed encrypted computations (Yang et al., 2019) so each party only learns the final output, i.e. their local model parameters and their local gradients. (Yang et al., 2019; Liu et al., 2018; Cheng et al., 2021) provide detailed security analysis for VFL, FTL and Secureboost algorithms, respectively. In such VFL implementation where data are feature-partitioned, since training is performed identically as the centralized solution except for the encrypted calculation and communication, FATE guarantees **lossless** performance, meaning the algorithms in FATE provides comparable accuracy as a centralized solution.

6. Full details are available at `https://github.com/FederatedAI/FATE/tree/master/cluster-deploy`
7. https://github.com/FederatedAI/KubeFATE

Table 2: Performance Benchmark

Data Size (in thousands)	# of Features of parties A,B	model	Sklearn	FATE-standalone	FATE-distributed
100	200,20	LR	0.3s	128s	67s
400	1000,1000	LR	20s	4420s	1252s
100	200,20	xgb	6s	198s	206s
400	1000,1000	xgb	26s	2021s	960s
10000	100,100	LR	28s	12754s	2267s
10000	100,100	xgb	1170s	9499s	1112s

4. Performance Benchmark

Using LIBSVM dataset [8], we demonstrate the scalability on training privacy-preserving Logistic Regression (LR) and XGBoost models (xgb). The per-iteration cost for a two-party system is listed in Table 2. For FATE-distributed, 5 CPUs with 80 cores and 256G RAM are used. For FATE-standalone, a 32-core CPU and 128G RAM is used. FATE is computationally heavy due to communication and computation of encrypted data, but FATE-distributed can reduce the overall cost significantly, especially for large-scale training (10 million samples), when the cost for coordination become less dominant and the advantage of parallelization shows. Additional performance benchmarks for implementing advanced algorithms on FATE include XGBoost (Cheng et al., 2021) and FTL (Liu et al., 2018; Jing et al., 2019). In (Zhang et al., 2020), a novel batch encryption algorithm is developed and improve the efficiency by 100× times. In (Sharma et al., 2019) security is enhanced by SPDZ algorithm. In (Liu et al., 2019) efficiency is improved by introducing multiple local update strategies.

5. Conclusions and Future Work

In this paper, we introduced the first industrial-strength federated learning platform FATE. As an open-source software, FATE encourages collaboration among the research and industry community and has been increasingly adopted for business applications. Future work directions include integrating blockchain functionalities into FATE, building light-weight versions of FATE for edge deployment and applications, and building new applications with FATE in industrial scenarios such as computer vision (Liu et al., 2020) and automatic speech-recognition (ASR) (Jiang et al., 2021) to further enable federated AI technologies.

Acknowledgement

This work was partially supported by the National Key Research and Development Program of China under Grant No. 2018AAA0101100.

8. https://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/binary.html

References

Sebastian Caldas, Peter Wu, Tian Li, Jakub Konecný, H. Brendan McMahan, Virginia Smith, and Ameet Talwalkar. LEAF: A benchmark for federated settings. *CoRR*, abs/1812.01097, 2018. URL `http://arxiv.org/abs/1812.01097`.

Kewei Cheng, Tao Fan, Yilun Jin, Yang Liu, Tianjian Chen, Dimitrios Papadopoulos, and Qiang Yang. Secureboost: A lossless federated learning framework. *IEEE Intelligent Systems*, 2021. doi: 10.1109/MIS.2021.3082561. URL `http://arxiv.org/abs/1901.08755`.

Cynthia Dwork. Differential privacy. In *Proceedings of the 33rd International Conference on Automata, Languages and Programming - Volume Part II*, ICALP'06, pages 1–12, Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 3-540-35907-9, 978-3-540-35907-4. doi: 10.1007/11787006_1. URL `http://dx.doi.org/10.1007/11787006_1`.

Di Jiang, Conghui Tan, Jinhua Peng, Chaotao Chen, Xueyang Wu, Weiwei Zhao, Yuanfeng Song, Yongxin Tong, Chang Liu, Qian Xu, et al. A gdpr-compliant ecosystem for speech recognition with transfer, federated, and evolutionary learning. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(3):1–19, 2021.

Qinghe Jing, Weiyan Wang, Junxue Zhang, Han Tian, and Kai Chen. Quantifying the performance of federated transfer learning, 2019.

Ce Ju, Ruihui Zhao, Jichao Sun, Xiguang Wei, Bo Zhao, Yang Liu, Hongshan Li, Tianjian Chen, Xinwei Zhang, Dashan Gao, Ben Tan, Han Yu, and Yuan Jin. Privacy-preserving technology to help millions of people: Federated prediction model for stroke prevention, 2020.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaïd Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konecný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *CoRR*, abs/1912.04977, 2019. URL `http://arxiv.org/abs/1912.04977`.

Qinbin Li, Zeyi Wen, and Bingsheng He. Federated learning systems: Vision, hype and reality for data privacy and protection. *CoRR*, abs/1907.09693, 2019. URL `http://arxiv.org/abs/1907.09693`.

Yang Liu, Tianjian Chen, and Qiang Yang. Secure federated transfer learning. *CoRR*, abs/1812.03337, 2018. URL `http://arxiv.org/abs/1812.03337`.

Yang Liu, Yan Kang, Xinwei Zhang, Liping Li, Yong Cheng, Tianjian Chen, Mingyi Hong, and Qiang Yang. A communication efficient collaborative learning framework for distributed features, 2019.

Yang Liu, Anbu Huang, Yun Luo, He Huang, Youzhi Liu, Yuanyuan Chen, Lican Feng, Tianjian Chen, Han Yu, and Qiang Yang. Fedvision: An online visual object detection platform powered by federated learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(08):13172–13179, Apr 2020. ISSN 2159-5399. doi: 10.1609/aaai.v34i08.7021. URL http://dx.doi.org/10.1609/aaai.v34i08.7021.

H. Brendan McMahan, Eider Moore, Daniel Ramage, and Blaise Agüera y Arcas. Federated learning of deep networks using model averaging. *CoRR*, abs/1602.05629, 2016. URL http://arxiv.org/abs/1602.05629.

Shreya Sharma, Chaoping Xing, Yang Liu, and Yan Kang. Secure and efficient federated transfer learning. *2019 IEEE International Conference on Big Data (Big Data)*, Dec 2019. doi: 10.1109/bigdata47090.2019.9006280. URL http://dx.doi.org/10.1109/BigData47090.2019.9006280.

Ben Tan, Bo Liu, Vincent Zheng, and Qiang Yang. A federated recommender system for online services. In *Fourteenth ACM Conference on Recommender Systems*, RecSys '20, page 579–581, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450375832. doi: 10.1145/3383313.3411528. URL https://doi.org/10.1145/3383313.3411528.

Zhaoping Xiong, Ziqiang Cheng, Xiaohong Liu, Dingyan Wang, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, and Mingyue Zheng. Facing small and biased data dilemma in drug discovery with federated learning. *bioRxiv*, 2020. doi: 10.1101/2020.03.19.998898. URL https://www.biorxiv.org/content/early/2020/03/20/2020.03.19.998898.

Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *CoRR*, abs/1902.04885, 2019. URL http://arxiv.org/abs/1902.04885.

Chengliang Zhang, Suyi Li, Junzhe Xia, Wei Wang, Feng Yan, and Yang Liu. Batchcrypt: Efficient homomorphic encryption for cross-silo federated learning. In *2020 USENIX Annual Technical Conference (USENIX ATC 20)*, pages 493–506. USENIX Association, July 2020. ISBN 978-1-939133-14-4. URL https://www.usenix.org/conference/atc20/presentation/zhang-chengliang.

Fanglan Zheng, Erihe, Kun Li, Jiang Tian, and Xiaojia Xiang. A vertical federated learning method for interpretable scorecard and its application in credit scoring, 2020.