Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/40/20230640

Research Advanced in Federated Learning

Sirui Li[1,4], Keyu Shao[2] and Jingqi Zhou[3]

[1]Southeast university, Nanjing, Jiangsu province, 211102, China
[2]Nanjing University of Finance & Economics, Nanjing, Jiangsu province, 210023, China
[3]Geely University of China, Chengdu, Sichuan province, 641423, China

[4]213162447@seu.edu.cn

Abstract. With the vigorous development of big data, cloud computing and other fields, it has become a global trend to pay attention to data security and privacy. In order to protect their own data security and privacy, different groups are unwilling to contribute their own data information, making the problem of data islands gradually prominent, which seriously restricts the further development of data-driven artificial intelligence. In order to alleviate the above problems, federated learning has attracted more researchers' attention in recent years. Federated Learning is a collaborative decentralized privacy-preserving technique that makes local data available to multiple parties, which not only can the private data be effectively used to train the model but also the leakage of private data can be avoided. Federated learning has been widely used in practical fields such as the financial industry and the Internet of Things industry. This paper systematically introduces the results of research in the field of federated learning in recent years. Specifically, three structures of federated learning are first introduced, and the differences between these three structures are introduced. Then, the most used datasets in training and validation stage were introduced and the shortcoming of each method were introduced to help advanced understanding of FL. Finally, several unsolved problems were introduced and the future prospects in federated learning domain were proposed.

Keywords: Federated Learning; Machine Learning; Data Privacy.

1. Introduction

The proliferation of big data has pushed artificial intelligence to a new peak. However, training a high-precision learning model requires large-scale high-quality data, involving preprocessing tasks such as data collection, data cleaning, and data labelling. High-quality data requires valuable expert knowledge and a lot of human and material resources, which makes different groups and even different industries unwilling to contribute their own data to each other in order to protect their own data security and privacy. This creates barriers between data sources and inhibits the effective integration and utilization of existing data. Therefore, finding a new data utilization model that allows local raw data to be used effectively is of great significance and has attracted a lot of research attention in recent years.

As a new machine learning paradigm, the emergence of federated learning (FL) provides a solution to the problem of data isolation, which can alleviate the problems of data isolation and privacy protection without exposing the data of all parties. Federated learning is a distributed machine learning model with the characteristics of privacy protection and data encryption, which aims to enable all participating

DOI: 10.54254/2755-2721/40/20230640

organizations to collaboratively provide data for machine learning model training without revealing the core private data of the participants. However, compared with classic distributed learning, federated learning focuses more on learning heterogeneous data sets. The data on different computing nodes may have completely different distributions, and the data scale may differ by several orders of magnitude. At the same time, a certain degree of privacy protection is required for the local data sets of each node.

FedAvg is generally considered to be the first formal exploration of federated learning, whose main contribution is to point out that a large amount of decentralized data is stored on mobile devices while fails to be used due to privacy issues. A federated learning system often includes a central server and multiple clients. The general steps of the training protocol process include: (1) the centre randomly selects a part of nodes in the set of terminal nodes and the selected node downloads the current global model parameters; (3) the selected node uses local data to update the global model parameters; (4) the selected node summarizes the updated model parameters to the centre; (5) the centre aggregates data through a specific algorithm, and updates Global model parameters; (6) iteratively execute the above five steps until the model converges to the expected value. There are three kinds of network frameworks for federated learning, distinguished by the type of used data and the scenario. Horizontal federated learning is suitable for the situation where the data sets have more user feature overlap but less user overlap. Longitudinal federated learning would be used when the features of the data sets used for training have less overlap, but the data is targeted to the same user. Federated transfer learning would be used when researchers use an existing model to extend to a related field but cannot access the data in field.

Focusing on the above three different types of federated learning models, this paper systematically introduces the latest research progress of federated learning technology for the above three different types of federated learning models. Specifically, this paper first summarizes the common training dataset used in federated learning. Secondly, this paper introduces and compares the existing mainstream open-source frameworks of federated learning, and gives the learning and application scenarios of federated learning. The paper also presents representative algorithms and their strengths and weaknesses in different field. Finally, aiming at the problems in the existing work, this paper proposes future challenges and prospects.

2. Federated learning

2.1. Horizontal federated learning

When doing machine learning, traditionally, main server would collect all the data which have same feature from users, and train the forecast model with those data. In many occasions, most enterprises would be reluctant to overtly share the data, because sharing data might lead to a threat to data privacy. With federated learning, users do not need to share their own data publicly, participants would only share the parameters to the main server without specific user data transmission. In the case of horizontal FL, as the data in one participant is limited, in order to increase the size of sample. Researchers would connect all the participants into a federated network. Each participants use the local data to train local model, while the feature of data is similar. As the HFL is the earliest published framework, it has been widely used nowadays. Andrew Hard introduced a method which help google to predict the word that users might input [1]. The data would be processed locally and worker would transmit the parameters of local trained model to the main server. After main server collect all the parameters and weight them, the main server will aggregate the parameters and broadcast the processed parameter to the participants, then the participants would update their local model into the more suitable one. In addition, HFL could also benefit to build modern healthcare systems [2], as the data from patients are private in hospital and it is illegal to transmit the data without permission. However, without access to sufficient data from hospital, it is hard for machine learning model to get reasonable parameters to match the clinical practice, which could limit the performances of machine learning in the medical field. With the help of HFL, the data will be only used locally and could provide a trained model matching local patients. After aggregating each parameter, the server could finally provide a model which is suitable to most

DOI: 10.54254/2755-2721/40/20230640

participants' data. This could keep data away from exposing, as well as increase the data set. In the field of Internet of Things (IoT), the HFL is also an efficient way to help managing IoT network [3]. As there is a great number of devices in factories, the communication between devices would cost a lot of resources when upload data to server when doing machine learning. Guangzhou University proposed a deep reinforcement learning (DRL) algorithm which can help model to make three key parameters in model more reasonable to make the IIoT networks more efficient and cost less in communication. Apart from that, FL could help to deal with quality control, the management of supply line, energy management.

2.2. Vertical federated learning

Because of the differences in the nature of enterprises, the data sets about consumers and potential consumers owned by different organizations have different feature Spaces. These organizations expect to share information to train models to get user's behaving analysis. The model built based on this heterogeneous data is the vertical federated learning model.

Vertical federated learning, also known as feature-based federated learning, divides data vertically (by column) on the basis that the ID of the sample remains unchanged. For vertical federation learning, participants would make entity alignment of the encrypted user ID, during which the system does not expose the user identity. After identifying shared and non-shared entity data, both participants collaboratively train a machine learning model with shared data. The steps of vertical federation learning model training are:

Step 1: In order to help each participant to communicate safely, federated learning network would select a coordinator which creates and distributes several public key pairs to each participant to group the participants in pairs.

Step 2: After finish local training, participants in pairs exchange and encrypt the intermediate results and calculate the loss value with local parameters.

Step 3: Participants send the encrypted gradients with additional masks to coordinator. Meanwhile, one of the participants calculate the value of encryption loss. Coordinator would collect all the result from participants.

Step 4: Coordinator decrypts the result and sends the results back to participants to help them to update local parameters.

Vertical federation learning can change encrypted data to avoid the privacy problems faced by different enterprises when they directly exchange characteristic information. At the same time, by approximating the polynomial, the encrypted operation caused by the exponential operation in the loss and gradient formula is not supported. Vertical federation learning enables all parties to build strong machine learning models based on the distributed characteristics of the same samples, and even improves the performance of vertical federation learning under limited overlapping samples through federated multi-view training [4].

The expectation of vertical federated learning is that both parties providing data are protected and the trained machine learning model is lossless. The practical challenge is that there is no way to build the model with only one party, resulting in the two parties cannot share data, and the loss rate of the model is difficult to approach the expectation. Nowadays, the algorithm model of vertical federation learning is mainly applied in enterprises with high data privacy.

2.3. Federated Transfer Learning

Federated Transfer Learning is a federated learning which is used for mobile terminals proposed by Google in 2016.The WeBank AI team began by studying the financial sector's practices before concentrating on big data collaboration scenarios between institutions and businesses. The solution of "federated transfer learning" was initially put out in order to combine transfer learning with federated learning, so that when there is little overlap between the participants, if the data is integrated into the central body at this time, there will be a large number of blank information, so without splitting the data,

DOI: 10.54254/2755-2721/40/20230640

it overcomes the lack of labels or data, thereby improving the effect of the model. According to [5], the steps of federated transfer learning are as follows:

Step 1: Participants use their own data sources to build local models.

Step 2: Participants use the local model to test their own data to validate the accuracy, and get the parameters. The results are encrypted and sent to the other party to aggregate.

Step 3: The results made by step 2 which is used for the other party to calculate the encrypted gradient and loss value of the model, adds the mask and sends it to the initial party.

Step 4: The enciphered data transmitted from step 3 will be received by each party and then decrypted. After decryption, the data will be transmitted back to the other party, and then the decrypted model information will be used by each party to update its own model. Repeat the step 1 to step 4 until the loss converges.

Distinguishing from horizontal federated learning and vertical federated learning, federated transfer learning does not require the main server as the coordinator among the participants; Federated transfer learning was designed to make the model have the ability to extend to similar situation from existing model. When there is little overlap between the information in each participant's sample and feature space, using the transfer learning algorithm is an efficient way to build the model.

Federated transfer learning can also solve data islands and security and privacy issues. In [6], the DAFedLDA algorithm is an algorithm that uses the mechanism of federated transfer learning to solve data islands and security and privacy issues. Specifically, it can use the federated transfer learning framework. Different members in the network can jointly mine the value of data under the premise of strictly adhering to data privacy, and can transfer supplementary data within the network. In this way, you only need to make minor adjustments to the model to see a significant improvement in accuracy, and it can even compare to the performance of training directly on all data without considering privacy at all. Security is also an important consideration for federated transfer learning, and security protection covers the entire process of training, evaluation, and cross-validation. Safe migration cross-validation mechanism ensures that data can bring performance improvements to members in the federation.

3. Common datasets and performance analysis

When using a federated learning framework, there are multiple devices or data sources involved, so different types of databases are used for training. The researchers would use different type of data sets to test the accuracy in different scenarios. In actual scenarios, private databases are mostly used for local model training, as processing and efficient using those data is the purpose of machine learning. Because such databases contain a lot of sensitive information and private data, using the federated learning network to train the model can keep these private data away from leakage. For the stage of model validation and performance testing, researchers generally use some open-source databases. These two kinds of databank were be introduced in the part, and the shortcoming of each were introduced.

3.1. Open-source database

Common open-source databases include MNIST, CIFAR-10, IMDB, etc. MNIST database is a commonly used handwritten digit recognition database, containing 60000 training samples and 10000 testing samples, which has been widely used in machine learning and deep learning model training and testing. In [7], researchers put forward a new federal study algorithm, matching the average (Matched Averaging) algorithm. Researchers set experiments based on two different datasets (EMNIST and CIFAR-10) to compare the performance of the matched averaging algorithm with the existing federated learning algorithms (including FedAvg and FedProx). The results show that the matching average algorithm can significantly improve the accuracy and convergence speed of the model, because it well balanced the communication and integrity of parameters. However, in the experiment, the algorithm assumes that different devices have similar computing power. In the actual application, due to the heterogeneity among participants, the performance of the algorithm may be degraded. In addition, the algorithm needs to perform multiple model update and average operations in each round, which may lead to high computational complexity and high cost of communication, which would limit the

DOI: 10.54254/2755-2721/40/20230640

scalability of the algorithm. How to improve the scalability of the algorithm without decreasing the performance of the algorithm is still a challenge for federated learning.

Zhang et al. proposed LR-XFL [8], an interpretable federated learning method based on logical inference, which is able to derive accurate global parameters from local parameters without accessing local data at the client. The most suitable logical connector for aggregating the client parameters is automatically determined by LR-XFL based on the characteristics of the client's local data. This transparent design enables domain experts to actively participate in the validation, refinement, and tuning of rules, thus also helping to improve the resulting FL model. In the experiment, the MNIST dataset and COVID-19 CT dataset are used to verify the classification effect and interpretability of the FL model. However, in the experiment, only the availability of logical reasoning in the model and the accuracy of prediction results are considered, but other performance indicators, such as training time and the cost of communication, are not considered.

To sum up, some of the problems of open-source datasets are the scale of datasets is limited, and the label of the data cannot fulfil the target of some experiments. Researchers would use private datasets or collect data from real devices to validate the algorithm and train the model.

3.2. Private database

In the actual application and more specific scenario, the open-source data sets of data label can't satisfy the experiment purpose, the researchers will use the private data sets or public data sets for model training, Andrew et al. [9] designed a recurrent neural network language model to predict the effects of mobile devices the keyboard to a federal study is verified, in the experiments. In the experiment, the authors use a data set from 1000 real users, and these data are divided into multiple groups. Each group is added to the federated learning network for model training to avoid the centralized storage and processing of user data. In addition, federated network would help to avoid the interaction of private data to increase the security of user data. The experimental results show that FL can significantly improve the accuracy of keyboard prediction on mobile devices. However, due to the dataset is from only 1000 users, it may not fully represent the mobile device user in real scenarios. In addition, the authors used only one model in the experiment and did not compare the effects of other models, so it is impossible to determine how the method performs on other models.

Gao et al. [10] proposed a hierarchical heterogeneous horizontal federated learning algorithm for brain-computer interface. The data set used in the training is EEG of multiple subjects, and the federated learning network is built by horizontal federated learning. Because the size of data set used in the experiment is far from the real data, the generalization performance of the algorithm may be affected, and the performance of the algorithm under large-scale data sets has not been verified.

To sum up, private datasets can certainly help to match the parameters to real situation, but one problem is that the access of data could be hard. And the size of private datasets shared for training model is far from the size of real. With the development of security technology, more private data owner could share the data for local model training without the leak of data.

4. Discussion

The model training of federation learning aims to solve the problem of processing different types of data sets by different organizations. Through the above explanation, we find that data privacy has become an unavoidable issue in today's federal learning research. Even if the federated learning model tries its best to protect the security of data privacy information through the intervention of a third party or the method of keeping the data from leaving the local area, the attacker can still obtain the privacy information and model of the local data through the gradient or parameter information. In the face of data attacks such as poisoning attacks, Byzantine attacks, adversarial attacks, etc., the privacy and security of data sets face huge challenges.

On the other hand, the federated learning model itself is trained on a large amount of data, and each model update is accompanied by tens of thousands of parameter training, which leads to high communication costs. In the case of network state, unstable network conditions can lead to higher

DOI: 10.54254/2755-2721/40/20230640

parameter transmission costs. At the same time, in order to strengthen the security and confidentiality of data, some learning models choose to sacrifice some performance to ensure the privacy and security of local data. All these lead to the problem of communication cost.

Based on the existing stage of business data, it is necessary to determine the data ownership and privacy protection, which leads to the problem of data silos to a certain extent, and the data of a single company is difficult to conduct data analysis for direct business transformation. Federated learning is a distributed model that can solve this problem to a certain extent. In the case that the existing framework of federated learning is basically unchanged, the algorithm that integrates blockchain technology and federated learning is the current hot, because it can realize data interaction while protecting private data, and can effectively solve the dilemma of the current application of federated learning.

Our research expectation is that, on the one hand, in order to deal with the challenges brought by different data classification, we hope to carry out personalized processing in terms of devices, data and models to reduce their heterogeneity and ensure that each device gets a high-quality personalized model. On the other hand, we expect the federated learning model to be able to layer data when processing data, and to train its own federated learning model through data sharing when users cannot obtain other parties' private data. With the development of data, federated learning has been applied in more and more ways. Among them, artificial driving and intelligent medical care are particularly prominent. Data cannot be interoperable between different medical institutions, and the amount of data in any hospital is limited, which leads to the formation of many data islands. The application of federated learning model can provide a privacy and security computing environment for the medical field, so that the user's privacy and security can be guaranteed, and the system efficiency and service capability can be improved simultaneously.

5. Conclusion

The emergence of federated learning effectively solves the problem of data islands. This paper provides an in-depth analysis of the definition, characteristics and classification of federated learning. Specifically, according to different data organization and processing ideas, this paper introduces the latest research progress of federated learning from three perspectives: horizontal, vertical, and transfer federated learning, including its design ideas and representative works. In addition, this paper also introduces commonly used data sets for federated learning and reports the experimental results of different methods on common data sets. Finally, this paper points out the problems existing in the existing schemes, and discusses the future challenges and directions worthy of research.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] HARD A, RAO K, MATHEWS R, et al. Federated learning for mobile keyboard prediction[EB/OL]. (2019-02-28) [2021-10-19]. https://arxiv. org/pdf/1811. 03604. pdf.

[2] Rieke, N., Hancox, J., Li, W. et al. The future of digital health with federated learning. npj Digit. Med. 3, 119 (2020). https://doi.org/10.1038/s41746-020-00323-1

[3] Yinghao Guo, Zichao Zhao, Ke He, Shiwei Lai, Junjuan Xia, Lisheng Fan, Efficient and flexible management for industrial Internet of Things: A federated learning approach, Computer Networks, Volume 192,2021,108122,ISSN 1389-1286, https://doi.org/10.1016/j.comnet.2021.108122.

[4] Yan, K., Yang, L., & Xinle, L. (2022) FedCVT: Semi-supervised Vertical Federated Learning with Cross-view Training, ACM Transactions on Intelligent Systems and Technology, 13.4: 64:1-64:16.

[5] Liang Tiankai, Zeng Bi, Chen Guang. A Survey of Federated Learning: Concepts, Technologies, Applications and Challenges [J]. Computer Applications, 2022, 42(12): 3651-3662.

DOI: 10.54254/2755-2721/40/20230640

[6] Wu Xing, Fan Yushun.Decentralized Asynchronous Federated LDA Algorithm for Mining User Needs [J]. Computer Integrated Manufacturing Systems, 2023, 29(04):1055-1068. DOI:10.13196/j.cims.2023.04.001.

[7] Hongyi Wang, Mikhail Yurochkin, Yuekai Sun. (2020). Federated Learning with Matched Averaging. In Proceedings of the International Conference on Machine Learning (ICML). arXiv preprint arXiv:2002.06440v1 [cs.LG].

[8] Yan ci Zhang, Han Yu. (2023). LR-XFL Logical Reasoning-based Explainable Federated Learning. arXiv:2308.12681v1 [cs.AI].

[9] Andrew Hard, Kanishka Rao, Swaroop Ramaswamy. (2019). FEDERATED LEARNING FOR MOBILE KEYBOARD PREDICTION. arXiv:1811.03604v2 [cs.CL].

[10] Gao, D., Ju, C., Wei, X., Liu, Y., Chen, T., & Yang, Q. (2019). HHHFL: Hierarchical heterogeneous horizontal federated learning for electroencephalography. ArXiv: 1909.05784 [Cs, Eess]. Retrieved from http://arxiv.org/abs/1909.05784.