Research Advanced in Federated Learning

Ruixin Gao[a]

School of Electronic and Information Engineering, Beibu Gulf University, Qinzhou, China

Keywords: Federated Learning, Machine Learning, Data Islands.

Abstract: As an emerging machine learning framework, federated learning has received extensive attention in recent years and has been applied to various fields, such as financial, medical, intelligent city, and automatic driving. Different from the traditional centralized machine learning methods, federated learning algorithms allow the model trained locally to update the center server, greatly protecting user privacy and data security. Through extensive literature research and analysis, this paper aims to report on the latest research progress of federated learning. According to the organizational difference in training data distribution, this paper introduces the representative federated learning studies from the aspects of horizontal federated learning, vertical federated learning and transferred federated learning, including their design ideas and basic pipelines. In addition, a discussion about the existing issues and future development directions are given, especially how to reduce the impact of the central server once the subservers are attacked. This is supposed to bring some new insight to the federated learning community.

1 INTRODUCTION

Federal machine learning, or federal learning, is a framework that can assist several businesses in doing machine learning modeling and data utilization while adhering to legal constraints and user privacy and security concerns.As is a distributed machine learning technology, federated learning allows multiple participants to train models and jointly model together without sharing raw data, which can technically break data islands and realize AI collaboration. To this end, federated learning has gradually becoming a hotspot in the machine learning community, attracting more and more attention from the academic and industry fields (Li, 2020; Zhang, 2021; Priyanka, 2021).

Federated learning protects the privacy of users' data through parameter exchange under the encryption mechanism, where users need not upload their original data. Federated learning breaks the data island, allowing participants to cooperate with AI without the data leaves the local area, realizing “knowledge sharing without data sharing”, and improving the effect of their AI models. This distributed machine learning method can effectively use the data scattered on the user's equipment for model training, and improve the accuracy and generalization ability of the model. Federal learning considers information security, data privacy protection and legal compliance in its design, so that it can carry out efficient machine learning on the premise of ensuring information security during big data exchange. Nowadays, federal learning has been applied in various fields, such as medical image analysis, financial data analysis, natural language processing and recommendation systems. For example, the researchers in Google applied federated learning to input predictions from mobile keyboards, namely Google's Gboard system. This method greatly improves the accuracy of the smartphone input method and does not disclose users' private data.

Much effort has been invested in advancing federated learning. According to the organizational difference in training data distribution, existing representative federated learning studies can be divided into the following three categories:

(1) Horizontal Federal Learning (HFL). Considering the scenes that two datasets share few user features, HFL methods divide the dataset along the direction of the user dimension, which aims to train the model using the data with the same user characteristics but different users. Horizontal federation is mainly to

[a] https://orcid.org/0009-0009-0740-5290

Gao, R.
Research Advanced in Federated Learning.
DOI: 10.5220/0012959500004508

In *Proceedings of the 1st International Conference on Engineering Management, Information Technology and Intelligence (EMITI 2024)*, pages 574-579
ISBN: 978-989-758-713-9

solve the problem of insufficient sample of each participant, following the main principle of summarizing the parameters of each local participant to the central server. The central server aggregates all the participant parameters and then broadcast them back to each participant. Each participant constantly interacts with the central server until the final model loss reaches a threshold or the training reaches the specified number of iterations.

(2) Vertical Federated Learning (VFL). If there are two data sets, they will be split based on the vertical dimension of features, eliminating any data that differs between the same users for training purposes. This method is called longitudinal federated learning. Vertical federal information mainly addresses the problem of insufficient data and information dimension.

(3) The Federal Transfer Learning (FTL). Instead of splitting the data when users' attributes in both data sets overlap less, FTL approaches use transfer learning to overcome the data or labeling. We refer to this technique as federated transfer learning.

Focusing on the above three aspects, this paper reports the latest research advanced in the federal learning field. In detail, the representative methods will be introduced in Section 2, including their design ideas, basic framework, key steps, advantages, and disadvantages. The performance of various federated learning methods is also compared in Section 3. In Section 4, a discussion of the existing challenges and future development directions is summarized, which is supposed to bring some new insights for the federated learning field.

2 METHOD

2.1 Horizontal Federal Learning

Horizontal federated learning mainly focuses on how to train models in parallel across multiple participants, while protecting data privacy for each participant. The representative horizontal federated learning is federal average (FedAvg) proposed by Google McMahan et al. in 2016 (McMahan, 2016). During the joint training stage, the cloud center server randomly selects a fixed proportion of clients from the clients to participate in training each time. Each participating client in local training then upload the training gradient parameters to the cloud center server for aggregation after several iteration times, which effectively reduces the communication rounds in the traditional training method. Compared with the previous algorithm, federal average can reduce 10-100 times and speed up the convergence of the model. Based on the federal average algorithm, Li et al. (Li, 2018) proposed FedProx algorithm in 2018, which is designed to solve the Non-IID (non-independent and equally distributed) data problem in federated learning. FedProx dynamically updates the number of times of different clients to optimize the communication efficiency, making the algorithm more suitable for non-independent and equally distributed joint modeling scenarios. The MOCHA algorithm (Smith, 2017) is another representative horizontal federated learning approach with a multi-task learning strategy. This algorithm enables personalized modeling by using a multi-task learning framework to learn independent but related models for each client, improving the ability to process heterogeneous networks. Federated SGD (Liu, 2020) is a simple baseline algorithm for stochastic gradient descent (SGD) training in a federated learning system. During each training round, participants performed SGD updates using local data and sent the updated model parameters to the server for aggregation. These algorithms have their own characteristics and are suitable for different application scenarios and data features.

FedDyn (Durmus, 2021) belongs to the category of horizontal federated learning. Participants can share the model parameters or gradient information for the joint model training, while maintaining the local storage and privacy protection of the data. FedDyn aims to solve the problem of dynamic participation in federal learning, where the number of participants and the data distribution constantly change since participants can usually join or leave the federated network dynamically. The traditional federal learning framework, usually assume that participants are static, that is, the number and identity of participants in the training process are fixed. However, in practical application, participants may for various reasons (such as equipment failure, network disconnection, user exit, etc.) dynamic to join or leave the federal network.

FedDyn algorithm mainly includes initialization, local training, model aggregation, dynamic adjustment, and iterative optimization. First, the server initializes the global model parameters and distributes these parameters to clients participating in federated learning. Each client then trains the model using its data. Different learning rates or regularization strategies may be applied in this process to accommodate the data distribution of the Non-IID. After the training session, the client sends the model updates back to the server. The server is responsible for aggregating these updates. To generate a new global model, the server dynamically adjusts the learning rate or other hyperparameters according to a strategy (such as validation set based performance) after each round of polymerization. The server distributes it to the client side. This process is repeated over multiple rounds until the model converges or reaches the preset number of training rounds. In some cases, to address the Non-IID problem, FedDyn may also include personalized tailoring steps, To ensure that the model is better adapted to the data distribution of each client.

The design idea of FedDyn includes the following aspects: (1) Dynamic participation mechanism. FedDyn algorithm allows participants to dynamically join or leave in the training process. According to the characteristics of different client data and learning progress, FedDyn can dynamically adjust the model parameters and learning strategy when the new participants join or leave, enabling it to adapt to the new data distribution and computing resources and ensuring the continuity and stability of the training process. (2) Model adaptability. FedDyn dynamically adjusts the learning rate to adapt to the change in the number of participants. When the number of participants increases, FedDyn can increase the learning rate to accelerate the convergence of the model; Otherwise, FedDyn reduces the learning rate to avoid overfitting or oscillation of the model when the number of participants decreases. (3) Privacy protection. In the FedDyn algorithm, data exchange and model updates between participants are carried out under the premise of encryption and privacy protection. FedDyn algorithm pays attention to protecting the security of data on edge devices, ensuring that data is processed locally and does not need to be transmitted to the server, thus reducing the risk of data leakage. This can protect users' privacy and data security. (4) Efficiency. By optimizing the model training process, the algorithm reduces unnecessary computing and communication overhead and improves the efficiency of federated learning. The algorithm can dynamically adjust the strategy and task allocation of the model training according to the computing power and data distribution of the participants, so as to maximize the use of computing resources.

2.2 Vertical Federated Learning

Vertical federated learning focuses on how to leverage a subset of features owned by different participants to jointly train a model. In the scenario of vertical federated learning, each participant has a different feature set, but the sample ID may be the same or different. Vertical federated learning aims to protect the feature privacy of each participant, while leveraging the feature information from all participants to improve the model performance. Common vertical federated learning algorithms includes secure federal linear regression (Reza, 2021) and SecureBoost (Cheng, 2021), which are detailed in following sections.

Secure federal linear regression (Reza, 2021) is a vertical federated learning algorithm based on linear regression models, which protects the feature privacy of each participant through model training and parameter exchange in an encrypted state. Security federal linear regression algorithm puts forward various optimization strategies, such as (1) improving the target function to optimize the parameter update rules; (2) compressing model parameters to improve model efficiency; (3) reducing communication rounds and adopting asynchronous updates to reduce communication overhead; (4) applying differential privacy, homomorphic encryption, and other privacy protection technology to strengthen privacy protection. These advances provide strong support for the promotion and application of federated learning in practical applications.

The basic process of secure federal linear regression is mainly as follows. First, the participants (e. g. A and B) initialize the respective model parameters. At the same time, if there is a Coordinator C, the C generates the key pair and distributes the public key to the other participants. Participant A computes its model A and the corresponding loss function A, and then sends this information to Participant B. After receiving the information from Participant A, Participant B calculates the loss function and sends it to Coordinator C (if available). Similarly, Participant B will also calculate its model B and send the relevant information to Participant A. In this process, A and B can choose to add random masks to their gradient information to increase privacy protection. Participant A generates A random number and adds it to gradient A, and then encrypt the result (if the encryption method is used). The encrypted gradient information is sent to the coordinator C for decryption. Participants A and B separately subtract their previously generated random numbers to obtain a global gradient of common training, which is then used to update models A and B separately. During the inference stage, coordinator C inputs the user characteristics and adds the results output from Participants A and B to obtain the final prediction result.

SecureBoost (Cheng, 2021) is based on the lift tree, which shares gradient information across participants in an encrypted way to achieve collaborative training of the model. By improving the construction process of the tree, optimizing the feature selection and division criteria, introducing regularization terms and so on, the classification or regression performance of the SecureBoost algorithm can be effectively improved. SecureBoost has achieved remarkable progress in financial risk control, medical diagnosis, intelligent recommendation and other fields.

Key steps of SecureBoost includes: (1) Entity alignment. This is the first step in the SecureBoost framework, finding a common set of data samples (such as common users) among all participants, and common users can be identified by the user ID. (2) Sample alignment. Under privacy protection, sample alignment of overlapping users with different characteristics between participants. (3) Initialization model. Select an initial model as the starting point, which is usually a basic decision tree model. (4) Model training. Actively send the id and record id of the sample to be marked to the corresponding participant, and ask the next tree search direction (i. e. to the left or right child node). After the passive participant receives the id and the record id to be marked, compare the values of the corresponding features in the sample to be marked with the records in the local search table, get the next direction of the tree search, and send the decision back to the active party. After the active participant receives the search decision of the passive party, go to the corresponding child node. (6) Aggregate statistical value. All passive participants need to correspond to the current node samples and aggregation, so that the gradient addition homomorphic encryption operations, segmentation evaluation will be performed by the active participant. (7) Building a tree model. By repeating the above training process, gradually build the final lifting tree model.

3 EXPERIMENT

3.1 Common Datasets

Common datasets for federated learning include FEMNIST, Shakespeare, Twitter, CelebA, Synthetic Dataset, Reddit, MNIST, Fashion-MNIST, CIFAR10, and CIFAR100, which are widely used for training and evaluating various federated learning algorithms and models. For instance, Shakespeare is a data set collected from the full set of Shakespeare's work, which contains 1,129 users (the characters in the work) and a total of 422,615 samples. Twitter is mainly used for sentiment analysis, containing 660120 users and a total of 1600498 samples. CelebA is an annotated face dataset, which can be used for training on image classification tasks, containing 9,343 users and a total of 200,288 samples. CIFAR100 is a labeled subset of the so-called 80 million miniature image data set, whose samples are visualized in Figure 1.

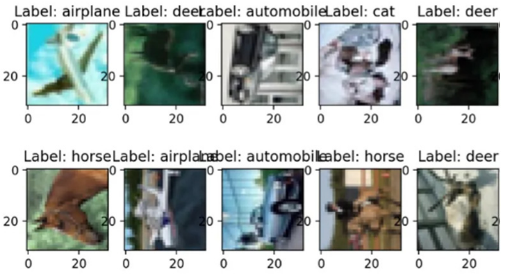

Figure 1: Visualization of training samples in CIFAR100.

3.2 Qualitative Results Comparison

In this section, the results of some representative federated learning algorithms on commonly used datasets are discussed, including federated averaging algorithm, federated optimization algorithm, security federal learning algorithm, stratified federated learning algorithm and personalized federated learning algorithm.

The federated averaging algorithm is usually validated in MNIST, FEMNIST and Shakespeare datasets, whose validation scenarios are usually under a data distribution of non-independent identical distribution (Non-IID) to simulate the data differences between different devices or users in real life. In addition to the federated average algorithm, there are some optimization algorithms for federated learning and can further improve the performance and convergence speed by optimizing the model training process. Federated optimization algorithms are usually validated in CIFAR-10, CIFAR-100 (Color Picture Classification) and Synthetic Dataset (synthetic data set), whose validation scenarios require more refined control of model updates, especially when dealing with large-scale data sets or complex models.

Security federal learning is a method of model training under the premise of protecting data privacy. Stratified federated learning is a method that divides participants into multiple levels for model training. In each level, the model is trained with different participants or algorithms, and the results are then summarized to the previous level to better deal with the problem of unbalanced data distribution and load imbalance. Personalized Federated learning is a method of personalized model training for each participant, where each participant can train the model according to their own data and needs, and then combine the personalized model with the global model to get a better model effect.

4 DISCUSSION

4.1 Existing Challenges

The model update between the subserver and the central server is one of key steps in federated learning design. Considering the fact that the attack on subservers will affect the central server, how to reduce the impact of the central server is an open issue and the following solutions can be taken:

(1) Monitor and protect the network traffic. The central server shall continuously monitor the network traffic from the subservers to guard against security threats.

(2) Quarantine subservers. Deploy subservers in an isolated network environment can reduce their impact on the central servers when they are attacked. Use the virtual private network (VPN) to encrypt the connection between the subserver and the central server to ensure the security of the data transfer.

(3) Access restrictions. By implementing strong password policies to limit access to critical systems and data, only authorized administrators can access to the child and central servers.

(4) Periodic updates and patching. Keep the latest version of operating systems, applications, and security patches for the child and central servers.

(5) Backup of important data. Regular backup of important data on the subserver and the central server can quickly recover the data and reduce losses.

(6) Implementation of security audit and monitoring. Conduct regular security audits and monitoring of sub-servers and central servers to find potential security vulnerabilities and abnormal behaviors.

4.2 Future Applications

Federated learning shows a promising future in many applications fields. (1) In the financial sector, several institutions' joint modeling risk control model can more accurately identify credit risk, and joint fraud. For micro and medium enterprises with scarce credit review data and historical information insufficient precipitation, federal learning can help Banks to ensure the data provider data security and privacy protection, multi-source information, jointly improve the effectiveness of the model. (2) According to retail knowledge, the social media platform is equipped with user preferences, and the bank possesses user purchasing power. Nonetheless, the usual machine learning features of the electric business platform apply, as the model is unable to directly analyze diverse data. Introducing the federal learning on the basis of joint modeling can provide users with more accurate product recommendation services. (3) In medical health, federal learning is very important to enhance the medical industry collaboration level since data privacy and security are more important in the field of health. Federal learning can extend more data sources (such as hospitals, laboratories, etc.), which can effectively protect data privacy and achieve data collaboration across institutions. Meanwhile, federal learning can also improve the processing efficiency of medical data. Traditional centralized methods need to collect all the data, which not only requires a lot of time and resources, but also may encounter inconsistent data and different formats. (4) In the field of autonomous driving and smart home, federal learning can improve the performance of the autonomous driving systems and the intelligence of the smart home systems by using the data collected from various vehicles and home devices to protect users' privacy.

5 CONCLUSIONS

This article systematically reviews the latest research progress in federated learning, and proposes a variety of federated learning algorithms, such as federated average algorithm, hierarchical federated learning, FedDyn, federated stochastic gradient descent algorithm. Meanwhile, federal learning protects the data privacy of participants through encryption technology, differential privacy, secure multi-party computing and other means, making it possible to realize the effective use of data while protecting personal privacy. With the deepening of the research, the application scenario of federal learning is expanding, showing a promising futrue in financial, medical, intelligent city, automatic driving, computer vision, and other fields.

REFERENCES

Cheng, K., Fan, T., Jin, Y., Liu, Y., Chen, T., Papadopoulos, D., Yang, Q., 2021. Secureboost: A lossless federated learning framework. *IEEE Intelligent Systems,* 2021, 36(6):87-98.

Durmus, A., E., A., Zhao, Y., Ramon, M., N., Matthew M., Paul, N., W., Venkatesh, S., 2021. Federated learning based on dynamic regularization. *In arxiv preprint arXiv:2111.04263v2.*

Li, L., Fan, Y., Mike, T., Lin, K., 2020. A review of applications in federated learning. *Computers & Industrial Engineering*, 149: 106854.

Liu, R., 2020. Fedsel: Federated sgd under local differential privacy with top-k dimension selection. *Database Systems for Advanced Applications: 25th International Conference, DASFAA 2020*, Jeju, South Korea, September 24-27, 2020, Proceedings, Part I 25. Springer International Publishing, 2020.

Li, T., 2020. Federated optimization in heterogeneous networks. *In Proceedings of Machine learning and systems,* 2 (2020): 429-450.

McMahan, H., B., Eider, M., Daniel, R., Seth, H., Blaise, A., 2017. Communication-Efficient Learning of Deep Networks from Decentralized Data. *In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 2017.*

Priyanka, M., M., 2021. Federated Learning: Opportunities and Challenges. *In arxiv preprint arxiv:2101.05428.*

Reza, N., Reihaneh, T., Jan, B., David, B., B., 2021. On the Privacy of Federated Pipelines. *In SIGIR'21: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information RetrievalJuly* 2021, 1975-1979.

Smith, V., 2017. Federated multi-task learning. *In Advances in neural information processing systems 30 (2017)*.

Zhang, C., Xie, Y., Bai, H., Yu, B., Li, W., Gao, Y., 2021. A survey on federated learning. *Knowledge-Based Systems*, 216:106775.