Dean&Francis
ISSN 2959-6157

A Survey on Optimization of Federated Learning Based on Edge-Cloud Collaboration

Xinyu Chen

School of Electronics and Communication Engineering, Sun Yat-sen University, Shenzhen, China
Corresponding author: chenxy399@mail2.sysu.edu.cn

Abstract:

In order to address issues including resource limitations, data heterogeneity, and computational inefficiency, this study explores the optimization of lightweight federated learning (FL) in edge-cloud collaborative contexts. Due to non-IID data distributions, high communication costs, and restricted device capabilities, traditional federated learning frameworks perform below expectations in edge environments. This study analyzes the background and challenges, elaborates on definitions and theories, emphasizes the classification of federated learning and finally provides several optimization solutions. The objective is to provide strong support for federated learning applications in resource-constrained scenarios while efficiently increasing the operational efficiency of federated learning systems in edge-cloud environments. The results lay the groundwork for future studies on scalable and effective distributed learning systems by highlighting the promise of lightweight FL in applications like smart healthcare and industrial Internet of Things (IoT). In particular, it underscores the need for adaptive mechanisms that dynamically adjust computation and communication strategies based on the capabilities of edge devices.

Keywords: Edge-cloud collaboration; Federated learning; Communication protocols; Optimization Methods.

1. Introduction

Due to the rapid proliferation of Internet of Things (IoT) devices and the extensive integration of artificial intelligence technology, worldwide data generation is increasing at an annual rate above 30%. By 2025, the global count of IoT devices will surpass 41.6 billion, while the market valuation of edge computing will attain $274 billion. The necessity for real-time data processing in fields including intelligent transportation, industrial automation, and smart cities is becoming progressively critical. Federated learning (FL), an innovative paradigm of distributed machine learning, has attracted considerable attention and widespread application in fields such as edge computing and deep learning. By enabling collaborative

ISSN 2959-6157

model training across multiple participants while preserving data privacy, FL effectively mitigates the challenge of data silos [1]. However, the process of federated learning involves multi-party collaborative management, and establishing a distributed, equitable, trustworthy, and secure consensus system is a critical challenge that must be resolved in its application. The data in federated learning comes from different users, each of which may employ different sampling locations, different sampling methods, and different amounts of collected data [2].

Traditional federated learning also faces challenges in the edge-cloud collaboration scenario. The computational resources of edge servers are inherently constrained. As the number of mobile terminals continues to grow rapidly and the complexity of application functionalities increases, the queuing time associated with offloading computing tasks to edge servers for processing has significantly extended, thereby leading to an increase in overall computing delay. Edge devices are typically characterized by weak computing power, small storage, and energy consumption sensitivity. Smart thermostats in smart home scenarios have CPUs with a clock speed of only tens of MHz, a storage capacity of only hundreds of KB, and rely on battery power. In medical imaging diagnosis, CT scanning devices in primary hospitals have preliminary feature extraction capabilities, but are limited by GPU computing power and cannot complete local training of complex models such as ResNet-50. Mobile edge networks are characterized by large bandwidth fluctuations, high latency, and significant packet loss rates. In the industrial Internet of Things scenario, within the coverage of 5G base stations in factory workshops, the upload bandwidth of devices fluctuates dynamically between 5 and 50 Mbps, and the latency ranges from 10ms to 200ms. Traditional federated learning, which uses synchronous aggregation, requires all edge nodes to upload model parameters within a fixed time window, resulting in low-bandwidth nodes becoming system bottlenecks.

The problem of data heterogeneity is particularly prominent in edge-cloud collaboration. There is data heterogeneity on the client side, that is, data from different clients varies in distribution, characteristics, samples, and quantity [3]. For example, in the field of financial risk control, the transaction data of banks in the eastern coastal areas is characterized by high frequency and small amounts, while the data of banks in the central and western regions is mainly low frequency and large amounts. This kind of Non-IID data leads to a 15%-20% drop in the prediction accuracy of global models in local scenarios. Traditional federated learning aggregates models through weighted averages and struggles to adapt to data heterogeneity, forcing systems to increase communication rounds to improve model performance and creating a “precision - efficiency” vicious cycle.

This paper will elaborate on the theoretical foundations, classify the types of federated learning, and provide development recommendations for lightweight federated learning.

2. Theoretical Foundation

2.1 Definition of Federated Learning

To address the issue of data silos and prevent the direct exposure of sensitive data during transmission, Google first introduced Federated Learning (FL) in 2016 [4]. Federated Learning is a collaborative machine learning framework that involves the participation of multiple entities. Through coordinated efforts, these participants jointly train a global model that achieves performance comparable to that of centralized training, while each party retains control. FL integrates advanced privacy-preserving technologies such as homomorphic encryption, differential privacy, and secure multi-party computation to enhance data security and protect user privacy during the collaborative learning process. For instance, differential privacy is applied to add noise during parameter updates to prevent reverse inference attacks. Homomorphic encryption enables direct computations on encrypted data while ensuring the security of intermediate results. Federated learning is now widely used in domains such as intelligent transportation systems, medical imaging analysis, and financial anti-fraud, and it is becoming a vital technological avenue to address issues with data silos and privacy protection [5].

From a technical architecture perspective, federated learning employs a “client-server” collaborative model. Clients train models using local data while transmitting only encrypted gradients or parameter updates to the central server. After aggregating these updates using appropriate algorithms, the central server updates the global model, which is subsequently distributed back to the participating clients for further iterations. Through repeated rounds of optimization, this approach enhances model generalization performance while effectively preventing the disclosure of raw data. The principle of “data remains localized while the model evolves” constitutes a core tenet of federated learning. Under this framework, each participant transmits only gradient updates or model parameters to the central server, which performs aggregation to refine the global model and then disseminates the updated model back to all participants for the next round of training.

2.2 Edge-Cloud Collaboration

An developing computing paradigm called edge-cloud collaboration technology uses cloud and edge computing to gather, process, and analyze different kinds of business data in real time. Low latency, high energy efficiency computing services are achieved through the dynamic allocation of tasks and resources. Edge nodes (edge servers, IoT devices) handle real-time, large volumes of local tasks (video surveillance, industrial sensor data), while the cloud centrally handles non-real-time tasks such as global analysis and long-term storage. The necessity of edge-cloud collaboration stems from the challenges posed by the explosive growth of IoT data. In the traditional cloud computing model, massive amounts of data need to be transferred to the cloud for processing, resulting in increased pressure on network bandwidth and higher latency. In the power grid, edge-cloud collaboration enhances data backup efficiency and security by encrypting monitoring data and uploading it to the cloud, while processing and distributing data storage at the edge.

2.3 Lightweight Federated Learning

Federated learning and edge computing are combined to create lightweight federated learning, which primarily addresses communication and processing power constraints so that devices with limited resources may take part in federated learning. By lowering computational, storage, and communication overhead through weight quantization and channel pruning, lightweight federated learning speeds up the federated learning process on IoT devices and eventually achieves effective training of global models with tolerable accuracy loss.

In communication optimization, federated learning optimizes communication efficiency through methods such as model compression, data compression, and communication scheduling, reducing the amount of data transmission between devices and enhancing transmission speed. Lightweight federated learning employs sparse update and federated averaging algorithms (FedAvg). Sparse updates involve transmitting only non-zero gradients, thereby reducing communication traffic. FedAvg computes a weighted average of parameters from each device, hence reducing the frequency of global model changes. Heterogeneous computing support is essential for achieving lightweight federated learning. Edge devices may utilize diverse hardware architectures, such as ARM processors and NPUs, and thus require deployment across platforms using a unified framework.

3. Classification of Federated Learning

Three forms of federated learning can be distinguished based on variations in data distribution: Horizontal federated learning is applicable to situations when the samples are varied but the characteristics are identical.[6]; Vertical federated learning targets overlapping samples with distinct features, such as credit assessments between banks and e-commerce platforms[6]; Federated Transfer Learning (FTL) is a specific type of federated learning that enhances target domain model performance by utilizing information from source domains when feature and sample spaces have little overlap [7].

3.1 Horizontal Federated Learning

Horizontal federated learning mainly addresses the “isolated data island” problem [8]. It achieves efficient aggregation of data value through collaborative expansion of distributed sample Spaces, especially for cross-agency collaboration scenarios where feature dimensions are highly similar but sample distributions have significant regional or group differences. By training models on local devices and only uploading feature parameters to a central server for aggregation and optimization, horizontal federated learning’s technological core aims to make user data accessible but invisible. All participants must have the same feature space for horizontal federated learning to work. However, the sample sets either have very little or no intersection. Its inherent suitability for vertical businesses with data barriers stems from this fact. The central server combines these local updates using a secure aggregation algorithm, creates global model parameters, and feeds them back to each participant after each participant completes forward and back propagation computations in the local environment. The central server only receives the encrypted gradient information or the amount of model updates. This process forms a closed-loop iteration until the model converges. Lateral federated learning, for example, enables secure data transmission and reliable communication by having participants train the model locally and register it at a key generation center and then use common parameters for authentication and session key negotiation. Under the condition of dispersed user data, the risk user identification model is jointly trained using data from one and two centers to ensure data security.

3.2 Vertical Federated Learning

A distributed learning architecture called vertical federated learning (VFL) enables several users to co-train a model without exchanging the original data, particularly in situations when the feature space is large. In VFL, different participants protect data privacy through encryption

ISSN 2959-6157

technology to ensure that model training and parameter passing are completed without leaving the domain. Longitudinal federated learning achieves cross-domain knowledge fusion by securely aggregating complementary features from different data sources, especially for collaborative scenarios where samples are highly overlapping but feature dimensions are dispersed. Build a virtual feature space that is encrypted so that participants may work together to model features at the feature level without disclosing the underlying data. Although the feature sets are complimentary, longitudinal federated learning necessitates that participants share sample identifiers.

In order to solve the challenge of challenging data sharing and utilization, longitudinal federated learning is employed in cross-border intelligent analysis of energy emissions. Asynchronous network updates and homomorphic encryption techniques are used to ensure the security and effectiveness of multi-party modeling. Vertical federated learning may be used in privacy-protected data frameworks in situations where each participant's training data has unique characteristics but overlapping sample ids, and the common sample ids are computed to accomplish data alignment. To safeguard sample id privacy and prevent intersection information from being revealed during the secret sharing allocation, the ALIGN framework employs homomorphic and exchangeable encryption algorithms throughout the data alignment procedure. When the final alignment results were applied to the model training, experiments showed that for every 10% increase in redundant data, the ALIGN frame could reduce the model training time by approximately 1.3 seconds and ensure a stable accuracy rate of over 85%. Or use encryption technology to create an architecture on isolated data that has the same sample distribution but separate feature distributions. This would allow for cooperative training of intelligent models while protecting data privacy.

3.3 Federated Transfer Learning

Federated transfer learning is a hybrid approach that integrates the principles of transfer learning and federated learning [9]. It enables multiple clients to collaboratively train a shared global model without exchanging raw data, thereby improving the model's generalization and adaptability through mechanisms of model transfer and personalization. In federated transfer learning, a federated aggregation technique based on model parameter averaging is employed to integrate local models from distributed clients into a unified global model, while ensuring the preservation of data privacy. Subsequently, this global model is fine-tuned on each client's local dataset using transfer learning techniques to enhance the performance and detection capabilities of individual local models [10]. By fostering cross-domain expertise, federated transfer learning tackles the problem of collaborative modeling in situations where participants' sample distribution and feature space differ significantly. bridges. The technique's main goal is to identify possible connections between various data domains and create nonlinear mapping relationships between the source and target domains using the little information that is supplied. This mapping is achieved through a common knowledge representation layer - which serves as a hub for cross-domain feature transformation, projecting source domain features into the latent space shared with the target domain while retaining key information about the target task. Its mathematical implementation typically relies on an adversarial training mechanism: the generator network is responsible for transforming the source domain features into a representation that the target domain can understand, while the discriminator network constrains the transformed feature distribution through domain classification loss, making it statistically indistinguishable from the target domain features. Federated transfer learning uses techniques such as gradient inversion layers to automatically optimize the dual objectives of domain adaptation and task prediction during backpropagation, ensuring that the general knowledge accumulated by the model during the pre-training phase in the source domain can be effectively transferred to the target domain, while fine-tuning the mapping parameters through a small number of shared samples, ultimately achieving seamless fusion of cross-domain features and improvement of task performance.

4. Optimization of Lightweight Federated Learning based on Edge-Cloud Collaboration

4.1 Dynamic Communication Optimization

Minimizing the overall volume of data transmitted during the model parameter exchange process or enhancing the transmission speed of model parameters constitutes the primary objective of communication optimization in federated learning. Achieving this can effectively reduce the time required for both local and global model uploads [11]. By enhancing the communication efficiency of federated learning through a hierarchical architecture and integrating it with the real-time perception and dynamic scheduling capabilities of 6G computing power networks, system resources can be dynamically allocated to effectively reduce transmission latency and energy consumption. By putting lightweight models to edge nodes and utilizing the

cloud for global model aggregation, a rational division of computing workloads is accomplished. The edge side is responsible for low-latency local training, while the cloud focuses on complex model updates and multi-node coordination, reducing the computational burden on a single node. Based on the heterogeneity of device computing power, participating clients can be grouped to ensure that low-performance devices prioritize the use of edge resources, while high-performance devices connect to the cloud, maximizing resource utilization. Model parameters may be effectively synced between the edge and the cloud by utilizing the low latency and high bandwidth features of 6G networks, which improves training efficiency overall.

4.2 Multi-modal Data Fusion in Federated Learning

The quantity of accessible training data has a significant impact on machine learning performance. The model produced by machine learning typically performs better the richer data. Data collected by edge devices often contains multimodal features such as text, images, timing signals, etc. Traditional protocols lack synergy in compressing multimodal data. It is recommended to build a multimodal compression framework for joint optimization in combination with feature dimension correlation. The collaborative optimization of multimodal data fusion and federated learning can enhance model classification performance through cross-modal alignment and shared representation [12]. By integrating heterogeneous data sources such as images, text, and sensors, this method provides richer feature representations for federated learning. However, traditional methods face challenges in data alignment, feature extraction efficiency, and privacy protection.

The heterogeneity of multimodal data manifests as significant structural and semantic variations across sources, complicating analysis, reuse, and interpretation [13]. This challenges traditional feature extraction methods to adapt to dynamic data distributions. Within federated learning frameworks, hierarchical feature encoders enable cross-modal alignment. Lightweight architectures combining Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) at edge nodes process image and text data respectively, with automatic feature fusion ratios adjusted through dynamic weight allocation mechanisms [14]. For instance, in medical imaging diagnosis, attentional mechanisms integrate CT texture features with electronic medical record (EMR) textual features, achieving a 12.7% improvement in AUC for pulmonary nodule detection while reducing feature dimensions by 63%. By introducing domain-adaptation layers via transfer learning and leveraging PyTorch's dynamic computational graph capabilities, the framework addresses feature drift caused by parameter differences across healthcare institutions.

In terms of privacy protection, modal federal learning achieves multi-modal data fusion while protecting privacy by performing knowledge distillation using minimal public data or no data on servers, combined with implicit sharing of local model parameters. The gradient aggregation process incorporates a dynamic noise injection strategy that adaptively adjusts noise intensity across training phases to minimize critical parameter perturbations. Integrated with secure multi-party computation technology, this approach enables authenticated parameter updates in ciphertext state, effectively defending against man-in-the-middle attacks initiated by malicious nodes.

4.3 Collaborative Optimization of Edge Computing and Federated Learning for Lightweight Model Architecture

Edge Computing Demonstrates Advantages Across Bandwidth Scenarios [15]. By employing a hyperparameter search algorithm with channel pruning awareness, the model size was reduced to 34% of its original size in breast cancer classification tasks while maintaining key diagnostic metrics within a 2% reduction. PyTorch Mobile achieved dynamic pruning that distilled the core decision-making logic of Transformer models into an LSTM network occupying merely one-tenth of the original volume, with prediction errors controlled below 3%.

Furthermore, by employing techniques such as dynamic hyperparameter optimization, edge-cloud collaborative training, and load balancing algorithms, adaptive resource scheduling and load distribution can be effectively achieved. The integration of edge computing and federated learning enables efficient utilization of computational and communication resources in resource-constrained environments through adaptive control algorithms, achieving distributed machine learning [16]. By integrating the Optuna and Flower frameworks, this solution automatically adjusts learning rates and batch sizes based on varying computational capabilities across edge devices. TensorFlow 2.9 leverages dynamic model slicing technology to automatically segment and load model components based on edge device computing capabilities. In medical imaging diagnosis scenarios, the backbone network of CT image recognition models is deployed on edge gateways, while fine-grained classification modules operate on mobile workstations, effectively reducing overall inference latency. The system initializes node weights based on real-time resource availability, prioritizing task allocation

ISSN 2959-6157

to high-weight nodes through round-robin scheduling. It monitors queue lengths across nodes and distributes new requests to those with the fewest connections. The algorithm dynamically adjusts weights or task routing paths in real-time according to periodic health checks, preventing local overload. By generating unique hash values that map requests to specific nodes, it significantly reduces scheduling overhead.

4.4 Edge Cache Assistance and Joint Computation Scheduling

In mobile edge networks, device offline and network congestion often cause communication disruptions. Deploy edge caching nodes and build a joint scheduling system of "computation-caching-communication". When edge nodes are training locally, cache the intermediate computing results to nearby edge servers. When the network is interrupted, the cache node forwards the cached data to the target node via D2D communication. After the communication is restored, the receiver combines the cache data with the new received parameters through the differential update mechanism. To optimize the caching strategy, a reinforcement learning model is introduced to dynamically adjust the cache content and forwarding path with the aim of minimizing communication delay and computational overhead. It is recommended to integrate the local differential privacy federated learning framework to coordinate user task scheduling, enhancing user device privacy protection while reducing the average total delay cost of the system.

5. Conclusion

In summary, this paper investigated the optimization of federated learning based on edge-cloud collaboration, highlighting key solutions to challenges like resource constraints and data heterogeneity. In the future, with the further evolution of edge computing and federated learning technologies, more efficient compression algorithms and intelligent scheduling mechanisms can be explored, while research on protocol security should be strengthened to promote the widespread application of federated learning in more complex scenarios.

References

[1] T. Li, A. K. Sahu, A. Talwalkar and V. Smith. Federated Learning: Challenges, Methods, and Future Directions. IEEE Signal Processing Magazine, vol. 37, no. 3, pp. 50-60, May 2020.

[2] W. Y. B. Lim et al. Federated Learning in Mobile Edge Networks: A Comprehensive Survey. IEEE Communications Surveys & Tutorials, vol. 22, no. 3, pp. 2031-2063, 2020.

[3] Z. Lu, H. Pan, Y. Dai, X. Si and Y. Zhang. Federated Learning with Non-IID Data: A Survey. IEEE Internet of Things Journal, vol. 11, no. 11, pp. 19188-19209, 1 June1, 2024.

[4] McMahan, H. B. et al. Communication-Efficient Learning of Deep Networks from Decentralized Data. International Conference on Artificial Intelligence and Statistics (2016).

[5] Bao, G., Guo, P. Federated learning in cloud-edge collaborative architecture: key technologies, applications and challenges." J Cloud Comp 11, 94, 2022.

[6] Liu, Yang, et al. "Vertical Federated Learning: Concepts, Advances, and Challenges." IEEE Transactions on Automatic Control 36.7(2024):20.

[7] Saha, Sudipan, and T. Ahmad. Federated transfer learning: Concept and applications. Intelligenza Artificiale 15.1, 2021:35-44.

[8] Kairouz, Edited By: Peter, and H. B. Mcmahan. Advances and Open Problems in Federated Learning. Foundations and Trends in Machine Learning 14.1, 2021.

[9] Huang, W., Li, T., Wang, D., Du, S., Zhang, J., & Huang, T. Fairness and accuracy in horizontal federated learning. Information Sciences: An International Journal, 589, 2022.

[10] Liu, W., Wang, Y., Li, K. et al. Ftmoe. A Federated Transfer Model Based on Mixture-of-Experts for Heterogeneous Image Classification. Cluster Comput 28, 165, 2025.

[11] Yang, Qiang, et al. Federated Machine Learning: Concept and Applications. ACM Transactions on Intelligent Systems and Technology 10.2(2019):1-19.

[12] Zhao, Yuchen, P. Barnaghi, and H. Haddadi. Multimodal Federated Learning on IoT Data. 2021.

[13] Cremonesi, Francesco, et al. The need for multimodal health data modeling: A practical approach for a federated-learning healthcare platform. Journal of biomedical informatics. 104338, 2023.

[14] Yu Q, Liu Y, Wang Y, et al. Multimodal federated learning via contrastive representation ensemble. arXiv preprint arXiv:2302.08888, 2023.

[15] Ye, Yunfan, et al. EdgeFed: Optimized Federated Learning Based on Edge Computing. IEEE Access 8(2020):209191-209198.

[16] Wang, Shiqiang, et al. Adaptive Federated Learning in Resource Constrained Edge Computing Systems. Selected Areas in Communications, IEEE Journal on (J-SAC) 37.6(2019):17.