Proceedings of CONF-CDS 2026 Symposium: Machine Learning and Neural Network Applications in Engineering
DOI: 10.54254/2755-2721/2026.AST36584

A Review of Lightweight Federated Learning Research for Home Edge IoT

Yiwei Zeng

Network Engineering, Shaanxi University of Technology, Hanzhong, China
yzeng233@my.trine.edu

Abstract. With the development of 5G and the widespread adoption of home IoT devices and wearable smart devices, home edge IoT devices generate a large amount of data. This data contains sensitive personal privacy and cannot be uploaded to public servers for training. Distributed machine learning technology, specifically federated training, is designed to protect data privacy, ensuring that data does not leave the local environment while multiple parties collaboratively train and share the model. However, home IoT devices differ significantly from industrial-grade IoT devices, exhibiting substantial heterogeneity, primarily in computing power, memory, power consumption, network connectivity, and in user data distribution, which makes it impossible to deploy federated training directly. This paper systematically reviews lightweight federated learning techniques for home edge IoT, analyzing two technical approaches: "device-aware adaptive lightweight training" and "data distribution-aware weighted aggregation." This paper proposes a co-optimization research framework, incorporating pruning rate, quantization bit width, and aggregation weights into a unified optimization space. The goal is to strike a balance among global model accuracy, training energy consumption, and convergence speed under device resource constraints, thereby providing a theoretical reference for future research.

Keywords: Federated learning, lightweight training, model pruning, model quantization, weighted aggregation

1. Introduction

The widespread adoption of home IoT devices generates large amounts of private data. Federated learning (FL) achieves privacy-preserving collaborative training through local training and parameter aggregation. Existing FL reviews mostly focus on industrial IoT and mobile edge scenarios, but lack a systematic analysis of the coupling between device heterogeneity and non-IID data characteristics in home scenarios [1, 2]. This paper reviews lightweight federated learning techniques for home edge IoT, focusing on device-aware adaptive pruning and quantization, data-distribution-aware weighted aggregation, and their co-optimization. Through a systematic literature review, three technical routes are outlined, providing a theoretical framework for FL deployment in home scenarios and revealing the evolutionary trend from static pre-setting to dynamic adaptation and from quantity weighting to similarity weighting.

DOI: 10.54254/2755-2721/2026.AST36584

2. Device-aware adaptive lightweight training methods

Home edge IoT devices vary greatly in computing power and power capacity: lightweight sensors have only tens to hundreds of KB of memory, while smart gateways have resources close to those of a computer [3, 4]; different devices have different data modalities and exhibit typical Non-IID distributions, and the local model update directions across clients are inconsistent [5, 6]. The superposition of device heterogeneity and data heterogeneity causes standard federated learning to face three bottlenecks in home scenarios: resource-constrained weak devices fall behind [7], non-IID data causes model drift, and existing lightweight strategies (static presets, aggregation based only on data volume) lack collaborative adaptation [8, 9]. To overcome these bottlenecks, this paper explores lightweight training methods from two dimensions: model pruning and model quantization, and further combines the two to achieve collaborative compression. This chapter systematically reviews the research progress of these three technical routes.

2.1. Model pruning techniques

Model pruning reduces model size by removing redundant or unimportant parameters from a neural network, and is one of the most direct ways to reduce on-device storage and computational overhead. Based on the pruning granularity, it can be divided into structured pruning and unstructured pruning.

2.1.1. Structured pruning and unstructured pruning

Structured pruning removes coarse-grained structures such as filters, channels, or layers. After pruning, the model maintains a regular tensor structure and can be deployed without special hardware or sparse computing libraries, which makes it particularly suitable for lightweight home devices that lack dedicated acceleration hardware. Hallaq et al. proposed a structured block pruning method guided by spectrograms, which prunes unnecessary encoder blocks by modeling pre-trained loss through occlusion of spectrograms. This significantly reduces computational and communication overhead while maintaining model accuracy and can be directly deployed on microcontroller platforms such as ARM Cortex-M [10].

Unstructured pruning removes individual weights, theoretically achieving higher compression ratios. However, the resulting sparse matrix requires support from specialized sparse computing libraries (such as cuSPARSE), making it difficult to implement on low-end home devices. The FedPal framework performs pruning during the initialization phase and supports both structured and unstructured modes, introducing a personalized client pruning mechanism that provides differentiated pruning options for devices with different capabilities [11]. In summary, structured pruning is a more realistic option for home edge IoT scenarios. It does not require devices to have sparse computing capabilities, and the pruned model can run efficiently on general-purpose processors.

2.1.2. Dynamic pruning strategy

Traditional static pruning determines the pruning rate either before or after training, which fails to adapt to real-time variations in power consumption, bandwidth, and memory availability of home devices. The core idea of dynamic pruning is to dynamically adjust the pruning ratio according to the current resource status of the device during federated training. EDPrune-FL proposed a dynamic

DOI: 10.54254/2755-2721/2026.AST36584

pruning method for heterogeneous environments, which reduces the client training energy consumption while ensuring model performance, and theoretically analyzed the relationship between pruning rate and convergence [12]. OPALA combines dynamic pruning with adaptive aggregation, and improves the efficiency and accuracy of personalized federated learning by adjusting model complexity [13]. FedCHAP enables the client to flexibly adjust the pruning ratio according to available resources, and ensures the feasibility of aggregation through weight filling and structure alignment [14].

2.2. Model quantization techniques

Model quantization reduces model storage space and computational overhead by reducing the precision of parameters (e.g., from 32-bit floating-point numbers to 8-bit integers or even lower). Unlike pruning, quantization does not change the model's structure but rather alters the representation of numerical values; these two techniques are complementary.

2.2.1. Post-Training Quantization and Quantization-Aware Training

Post-Training Quantization (PTQ) is applied to fully trained models. It is simple to operate, does not require retraining, and is suitable for one-time deployment on resource-constrained devices. However, PTQ may cause significant accuracy loss at low bit widths [15].

Quantization-Aware Training (QAT), on the other hand, simulates quantization errors during training, enabling the model to adapt to low-precision representations and maintain higher accuracy at low bit widths. The additional training overhead is concentrated in the training phase, and the device-side inference burden is not increased [16].

2.2.2. Mixed precision and adaptive quantization

Device heterogeneity requires quantization schemes to be configured according to the different capabilities of each device. Mixed precision quantization allows different layers or devices to use different bit widths, achieving a precision-efficiency trade-off on resource-constrained devices.

FedLQ proposed a layer-wise quantization scheme. The core observation is that different network layers have varying sensitivities to quantization errors. More important layers are assigned high bit widths and less important layers low bit widths, a strategy that maximizes precision preservation while keeping the compression ratio constant. The relationship between quantization level and convergence boundary was theoretically analyzed [17]. LAQ-HC enables each client to adaptively select the quantization level based on its own data quality and communication capabilities [18]. FedQLoRA combines adaptive quantization with low-rank adaptation (LoRA), dynamically adjusting the quantization level according to device resources, significantly reducing training time and memory consumption [19].

The common feature of the above methods is a shift from uniform quantization to adaptive quantization: the quantization bit width is no longer globally fixed but is chosen autonomously by each device based on its own resource status and data characteristics. This autonomy aligns closely with the wide variation in device capabilities typical of home scenarios.

2.3. Co-optimization of pruning and quantization

Optimizing pruning and quantization separately has limitations: pruning alone does not reduce numerical precision and still requires floating-point calculations; quantization alone does not change

DOI: 10.54254/2755-2721/2026.AST36584

the model structure and only slightly reduces computational cost. Combining the two is the optimal solution for edge deployment.

2.3.1. Joint compression framework

In recent years, researchers have begun to integrate pruning and quantization into a unified compression framework. FCP (Full Compression Pipeline) integrates pruning, quantization, and Huffman coding into an end-to-end framework, maintaining competitive accuracy while reducing transmission costs and resource consumption [20].

Other studies have integrated adaptive pruning, quantization, and communication frequency optimization into a hierarchical federated learning framework, dynamically adjusting pruning ratios and quantization levels through joint compression; other works have combined wireless transmission power control and gradient quantization to derive a convergence gap expression that accounts for multi-source errors. The FedSL scheme dynamically prunes on the client side and uses quantized gradient updates, further reducing the training burden on edge devices.

2.3.2. Adaptive adjustment based on device status awareness

The key evolution of the above framework is the shift from statically preset compression rates to dynamic adaptive adjustment based on device status awareness.

Research has proposed an energy-adaptive multidimensional learning control framework for energy-harvesting AIoT systems. This framework determines model complexity and training intensity based on the device's real-time energy status: a lower pruning rate and higher quantization bit width are used when energy is sufficient; automatic reverse adjustment is applied when energy is insufficient to save energy. This approach is also applicable to battery-powered home sensors.

For home scenarios, the system automatically elevates pruning ratios and lowers quantization bit widths to mitigate computational load when battery capacity drops or network performance degrades; when the device is in good condition, the lightweight quantization intensity should be relaxed to ensure accuracy. This adaptive workload reduction mechanism is the key difference between lightweight training in home scenarios and general-purpose scenarios.

2.4. Data distribution-aware weighted federated aggregation algorithms and challenges

Lightweight training on the device side solves the problem of device compatibility, but the ultimate goal of federated learning is to train a globally effective model. In home scenarios, user behavior data is naturally segmented per user, exhibiting a typical Non-IID distribution and significant label differences that cannot be eliminated through random sampling. The core of the aggregation phase is determining the updated weights for each device, which determines the performance and convergence of the global model. This section reviews the evolution of aggregation algorithms from the perspective of data distribution.

2.4.1. Data-quantity-based weighted aggregation methods

The limitations of data-quantity-based aggregation have prompted researchers to explore more reasonable weight allocation schemes. Some reviews have categorized aggregation strategies into three types: data-driven, model-driven, and secure aggregation.

FedProx alleviated the device heterogeneity problem to some extent by adding a proximal term to the local objective function to limit the model's bias across heterogeneous devices [21], but it still

DOI: 10.54254/2755-2721/2026.AST36584

relies on data-quantity-based aggregation and does not fundamentally alter the weight allocation logic. Some works indirectly affect the aggregation weight from the perspective of client selection. Although this can mitigate some biased updates, in home scenarios with a limited number of devices, excessive screening can lead to a low participation rate.

The data-quantity-based method is essentially quantity-first, assuming that more data is better and more representative. This assumption holds true in IID scenarios, but the opposite holds true in Non-IID scenarios.

2.4.2. Weighted aggregation based on data similarity

In recent years, researchers have begun to explore weighted aggregation strategies that account for data quality and distribution similarity.

FedLBW (Federated Learning with Loss-Based Weighting) uses the inverse of the validation loss of the server-side proxy dataset for weighting. A lower validation loss indicates a more reliable update [9]. It significantly outperforms FedAvg in extreme Non-IID scenarios and is robust to client stragglers. FedImp (Federated Impurity Weighting) quantifies device contribution based on the information content of local data. Data with higher information content is considered more valuable for model training [22] and performs well in extreme imbalanced distributions. DDAA (Data Distribution-Aware Aggregation) uses KL divergence to measure the degree of deviation between local and global data distributions and assigns weights [23].

The common trend of the above methods is to shift from quantity-based weighting to quality/similarity-based weighting. The weights no longer depend solely on the amount of data but also on the data quality. This shift is of great significance in mitigating model drift caused by Non-IID data in home scenarios.

3. Conclusions

This paper systematically reviews lightweight federated learning techniques for home edge IoT scenarios, analyzing the core characteristics of this scenario in terms of device heterogeneity and non-IID data distribution, and identifying three key problems: resource-limited devices that lead to low training participation, non-IID data that cause model drift, and the poor adaptability of lightweight strategies to aggregation schemes. To address these problems, this paper reviews two core technical directions. In terms of device-aware adaptive lightweight training, structured pruning is more suitable for home devices. Dynamic pruning and adaptive quantization have evolved from static presets to state-aware dynamic adjustments, and the joint compression of pruning and quantization achieves benefits on both fronts. In terms of data distribution-aware weighted aggregation, methods such as FedLBW, FedImp, and DDAA have achieved a shift from quantity-weighted to quality/similarity-weighted, effectively mitigating model drift caused by non-IID data. The review shows that existing work has accumulated substantial results in both lightweight training and aggregation algorithms, but a systematic understanding of how to use these two as coupled variables for co-optimization has not yet been established. Pruning and quantization alter the distribution characteristics of model parameters, thus affecting the similarity measurement in the aggregation stage; conversely, the allocation of aggregation weights also influences the design of compression strategies. This two-way coupling relationship is the core entry point for future research.

Based on the aforementioned review, future research can be explored in the following directions: The lightweight deployment of large language models on home sensors faces a tension between

DOI: 10.54254/2755-2721/2026.AST36584

compression ratio and accuracy preservation. The challenge of transferring knowledge from large models to micro-architectures has yet to be addressed. User behavior changes with seasons and daily routines, which can lead to conceptual drift. Finally, lightweight continuous learning under the dual constraints of Non-IID and resource limitations remains largely unexplored. Differential privacy, while protecting privacy, leads to decreased accuracy and increased energy consumption. Identifying the Pareto optimal trade-off among these three factors and designing a layered privacy protection scheme tailored to different device capabilities constitute core challenges in system design. Furthermore, FedImp relies on professional metrics such as information entropy to allocate weights, which is difficult for users to understand. A visualized weight allocation basis and an explanation mechanism for end users are needed to establish system credibility. These directions ultimately point to a unified framework for co-optimization—incorporating pruning rate, quantization bit width, and aggregation weights into a unified optimization space, designing alternating optimization algorithms, and analyzing convergence guarantees are the key steps for moving co-optimization from concept to practical application.

References

[1] Kaur, I., & Jadhav, A. J. (2023). Federated learning in IoT: A survey from a resource-constrained perspective. *2023 International Conference on Artificial Intelligence Robotics, Signal and Image Processing (AIRoSIP)*, Yogyakarta, Indonesia.

[2] Myakala, P. K., Naayini, P., & Kamatala, S. (2025). A survey on federated learning for TinyML: Challenges, techniques, and future directions. *Zenodo*.

[3] Ferraz Junior, N., et al. (2025). FedSensor: Federated learning framework for secure sensor-based IoT at the extreme edge. *IEEE Access, 13*, 136945-136969.

[4] Microsoft Azure IoT. (n.d.). Overview of Azure IoT device types. *Microsoft Azure IoT Documentation*. https://learn.microsoft.com/zh-cn/azure/iot/iot-introduction.

[5] Lim, W. Y. B., et al. (2020). Federated learning in mobile edge networks: A comprehensive survey on system design. *IEEE Communications Surveys & Tutorials*.

[6] Fan, B., Ren, Y., & Ma, Y. (2026). Federated learning for smart homes: A review. *ACM*.

[7] Talukder, Z., Lu, B., Ren, S., & Islam, M. A. (2025). Hardware-sensitive fairness in heterogeneous federated learning. *ACM Transactions on Modeling and Performance Evaluation of Computing Systems, 10*(1), 1–31.

[8] Wang, S., Tuor, T., Salonidis, T., et al. (2019). Adaptive federated learning in resource constrained edge computing systems. *IEEE Journal on Selected Areas in Communications, 37*(6), 1205–1221.

[9] Kundroo, M., Singh, T., & Kim, T. (2026). FedLBW: A loss-based weighting strategy for FL on non-IID data. *Expert Systems with Applications, 302*, 130487.

[10] Hallaq, M., Khan, F. M. A., Aboulfotouh, A., et al. (2025). Tiny federated wireless foundation models for resource-constrained devices. *IEEE Internet of Things Journal, 12*(19), 39197–39210.

[11] Wang, H., Liu, Z., Hoshino, K., Zhang, T., Walters, J. P., & Crago, S. (2025). FedPaI: Achieving extreme sparsity in federated learning via pruning at initialization. arXiv:2504.00308.

[12] Chen, G., Sun, F., Zou, W., Li, C., Zhou, Y., Lin, M., et al. (2026). Energy-efficient federated learning with dynamic model pruning for industrial IoT. *IEEE Transactions on Industrial Informatics*.

[13] Veiga, R., Morais, R., Bustincio, R., Bastos, L., Rosario, D., Sargento, S., et al. (2025). OPALA: Optimized pruning adaptive learning approach for federated learning scenarios. *IEEE Symposium on Computers and Communications (ISCC)*, 1–7.

[14] Chang, L.-H., Huang, J.-W., & Lee, T.-H. (2026). Federated learning with dynamic pruning for efficient model aggregation in heterogeneous environments. *IEEE Access, 14*, 8972–8993.

[15] Dataset-aware dynamic pruning strategy with gradient control for heterogeneous federated learning (Doctoral dissertation). (2025). Iowa State University.

[16] Wang, T., Guo, J., Zhang, B., Yang, G., & Li, D. (2025). Deploying AI on edge: Advancement and challenges in edge intelligence. *Mathematics, 13*(11), 1878.

[17] FedLQ: Towards layer-wise quantization for heterogeneous federated clients. (2025). *Computer Networks*.

[18] LAQ-HC: Lightweight adaptive quantization for heterogeneous clients. (2025). *IEEE*.

DOI: 10.54254/2755-2721/2026.AST36584

[19] FedQLoRA: Federated fine-tuning on heterogeneous devices with adaptive quantization and LoRA depths. *IEEE*, 2026.
[20] Colybes, E., Salehi, S., & Schmeink, A. (2026). A full compression pipeline for green federated learning in communication-constrained environments. *IEEE International Conference on Machine Learning for Communication and Networking (ICMLCN)*.
[21] Li, T., Sahu, A. K., Zaheer, M., et al. (2020). Federated optimization in heterogeneous networks. *Proceedings of the 3rd Conference on Machine Learning and Systems (MLSys)*, 429–450.
[22] FedImp: Enhancing federated learning convergence with impurity-based weighting. *IEEE Transactions on Artificial Intelligence, 7*(3), 1652–1665.
[23] DDAA: Data distribution-aware aggregation for federated learning on Non-IID data. *IEEE ICC*, 2024.