International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

FEDBAT++: COMMUNICATION-EFFICIENT FEDERATED LEARNING VIA IMPROVED BINARIZATION-AWARE TRAINING

.

Nithya Niranjana Murthy[1] and S H Manjula[2]

[1]Research Scholar, Department of Computer Science and Engineering,
Bangalore University, Bengaluru – 560001, Karnataka, India.
nithya.semantic@gmail.com

[2]Department of Computer Science and Engineering,
Bangalore University, Bengaluru – 560001, Karnataka, India.
shmanjula@gmail.com

*Corresponding Author: nithya.semantic@gmail.com

Abstract

Federated learning enables collaborative model training across distributed clients while preserving data privacy, but its practical deployment is often hindered by high communication overhead. Recent approaches such as FedBAT employ binarization-aware training to reduce local computational costs; however, the communication pipeline still relies on full-precision parameter transmission, limiting overall efficiency. In this paper, we present FedBAT++, an improved communication-efficient federated learning framework that extends binarization-aware training to the communication process. Through detailed communication cost analysis, we show that the baseline FedBAT transmits approximately 2.99 GB over 100 training rounds using full-precision parameters, despite maintaining binarized representations during local optimization. FedBAT++ addresses this inefficiency by enabling the transmission of binarized updates with learned scaling factors, significantly reducing uplink communication while preserving convergence stability and model accuracy. Extensive experiments under both IID and non-IID data distributions demonstrate that FedBAT++ achieves substantial communication savings without compromising performance. These results highlight the effectiveness of integrating binarization-aware training with communication compression, making federated learning more practical for bandwidth-constrained and resource-limited environments.

Keywords: *Federated Learning; Communication Efficiency; Binarization-Aware Training; Model Compression; Quantized Communication; Distributed Learning; Edge Intelligence*

1. Introduction

Federated Learning (FL) has emerged as a transformative paradigm for training machine learning models on distributed data while preserving privacy and data sovereignty. Unlike traditional centralized machine learning approaches that require data to be aggregated at a single location, FL enables collaborative model training across multiple decentralized clients such as mobile devices, IoT sensors, or organizational servers without exposing raw data to a

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

central server. This privacy-preserving training paradigm has gained significant attention in domains where data sensitivity and regulatory compliance are critical, including healthcare, finance and personal devices [1]. The fundamental FL workflow, first introduced through the Federated Averaging (FedAvg) algorithm, consists of iterative communication rounds. In each round, the server distributes a global model to a subset of clients, clients perform local training using their private data, and the resulting model updates are transmitted back to the server for aggregation. This process is repeated until convergence, often requiring hundreds or thousands of communication rounds for deep models and large-scale datasets.

Despite its advantages, the practical deployment of federated learning faces several challenges, among which communication efficiency is one of the most critical bottlenecks. Modern deep neural networks often contain millions or even billions of parameters, and repeatedly transmitting these parameters between clients and the server can incur substantial communication costs. For example, a ResNet-50 model with approximately 25 million parameters requires nearly 100 MB per communication round when using 32-bit floating-point representation. Over hundreds of rounds and multiple clients, this leads to tens or even hundreds of gigabytes of data transfer, posing significant challenges in bandwidth-limited or resource-constrained environments [2]. To mitigate communication overhead, a wide range of compression techniques have been explored. Quantization-based approaches reduce parameter precision from 32-bit floating point to lower bit-width representations such as 8-bit, ternary, or binary values. Sparsification methods transmit only a subset of parameters or gradients, often selecting those with the largest magnitudes [3]. Gradient compression techniques exploit the observation that not all gradient updates contribute equally to convergence [4]. Additional strategies such as federated dropout and structured updates have also been proposed to reduce communication volume while maintaining model performance [5].

FedBAT (Federated Binarization-Aware Training) represents a recent advancement that focuses on improving computational efficiency through learnable binarization during local training [6]. Unlike conventional quantization techniques that rely on fixed quantization schemes, FedBAT introduces a dual-parameter architecture in which each layer maintains full-precision base weights alongside learnable binarized update parameters. These binarized updates are controlled by adaptive scaling factors that are optimized during training. FedBAT further employs a warmup strategy, where training begins in full precision and gradually transitions to binarization after a predefined fraction of training rounds, enabling stable convergence even under non-IID data distributions [7].

Although FedBAT has demonstrated promising results in reducing computational cost while preserving model accuracy, particularly in heterogeneous data settings, its communication characteristics have not been thoroughly analyzed. Understanding the communication behavior of FedBAT is essential for assessing its end-to-end efficiency, identifying optimization opportunities, and establishing reliable baselines for future communication-efficient extensions. In this work, we present a comprehensive analysis of communication costs in FedBAT and establish detailed baseline measurements across multiple data distribution scenarios. Our analysis reveals an important and previously underexplored finding: despite

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

using binarized representations during local computation, FedBAT transmits full-precision (32-bit floating-point) parameters during client–server communication. This observation highlights a significant, untapped opportunity for communication compression.

The main contributions of this paper are summarized as follows:

- A comprehensive measurement of communication costs in FedBAT under multiple data distribution scenarios, including IID, label distribution skew, and label count skew
- A detailed analysis showing that FedBAT transmits full-precision parameters despite employing binarization for local computation, revealing substantial potential for communication reduction
- Baseline communication metrics for upload, download, and total communication cost, reported both per round and over complete training runs
- Empirical validation that communication volume in FedBAT is independent of data heterogeneity
- Visualization and systematic breakdown of communication costs throughout the federated learning process
- A discussion of the implications of these findings and potential directions for communication-efficient improvements

The remainder of this paper is organized as follows. Section 2 reviews related work on communication-efficient federated learning and quantization techniques. Section 3 provides background on federated learning and the FedBAT methodology. Section 4 describes the experimental setup and communication measurement framework. Section 5 presents detailed experimental results and analysis. Section 6 discusses the implications, limitations, and future directions. Finally, Section 7 concludes the paper.

2. Literature Review

Recent advancements in precision agriculture have demonstrated the strong potential of integrating

Communication-efficient federated learning has received significant attention in recent years due to the prohibitive cost of transmitting large model updates across distributed clients. Early foundational work on gradient quantization, such as QSGD proposed by Alistarh et al. (2017) [8], demonstrated that stochastic gradient quantization could achieve 2–4× communication reduction while preserving theoretical convergence guarantees. TernGrad by Wen et al. (2017) [9] introduced ternary gradient representations with scaling factors, achieving up to 16× compression in distributed deep learning. While these methods laid important theoretical groundwork, their applicability to federated learning scenarios with strong data heterogeneity was limited. Courbariaux et al. (2016) [10] showed that binary neural networks could be trained with acceptable accuracy in centralized settings, motivating later federated adaptations. Bernstein et al. (2018) [11] proposed SignSGD, which communicates only the sign of gradients, reducing communication to 1 bit per parameter. Although SignSGD demonstrated

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

strong theoretical efficiency, empirical studies reported slower convergence and sensitivity to noise in non-IID federated settings, limiting its robustness in practical deployments.

Reisizadeh et al. (2020) [12] introduced FedPAQ, which combines periodic averaging with quantized updates and showed that communication can be reduced by up to 4× with minimal accuracy degradation on standard federated benchmarks. However, FedPAQ relies on fixed quantization schemes and does not adapt quantization dynamically across clients. Addressing this limitation, Jiang et al. (2022) [13] proposed federated learning with adaptive quantization, where quantization levels are adjusted based on client-side data heterogeneity. Their results on image classification tasks demonstrated improved convergence stability and higher accuracy compared to fixed-precision baselines, particularly under non-IID data distributions. Stich et al. (2018) [14] introduced sparsified SGD with memory, showing that transmitting only the top-k gradient elements while accumulating residuals can achieve substantial compression without sacrificing convergence. Wangni et al. (2018) [15] further improved sparsification by incorporating momentum correction, reducing bias introduced by aggressive gradient dropping. These methods demonstrated strong theoretical guarantees but often required careful tuning of sparsity levels to balance compression and accuracy. Caldas et al. (2018) [16] proposed federated dropout, which reduces communication by dropping entire neurons or channels during training. This approach was shown to reduce client-side computation and communication costs while maintaining accuracy, especially for convolutional neural networks.

Research Gap and Motivation

Despite substantial progress in communication-efficient federated learning, a critical gap remains between computational compression and communication compression. Existing approaches predominantly focus on reducing communication overhead through quantization, sparsification, or hybrid compression techniques, often treating these mechanisms independently from the model's internal training dynamics [17]. While recent adaptive quantization methods introduced after 2021 improve robustness under data heterogeneity, they generally rely on fixed or heuristic compression schemes that are not tightly integrated with the learning process itself. FedBAT represents an important step toward closing this gap by introducing learnable binarization for local computation through a dual-parameter architecture and adaptive scaling. However, prior work on FedBAT primarily evaluates computational efficiency and accuracy, without examining its end-to-end communication behavior [18]. Our empirical analysis reveals a fundamental limitation: despite maintaining binarized representations during local training, FedBAT transmits full-precision (32-bit floating-point) parameters during client–server communication. As a result, the potential communication benefits of binarization remain entirely unexploited in practice [19].

Motivated by this gap, we propose FedBAT++, a communication-efficient extension of FedBAT that explicitly integrates binarization-aware training with compressed communication. FedBAT++ is designed to transmit binarized updates along with learned scaling factors, thereby bridging the disconnect between computational efficiency and communication efficiency. By aligning the training representation with the communication

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

representation and incorporating a warmup-controlled transition to compression, FedBAT++ aims to achieve substantial communication reduction while preserving convergence stability and model accuracy. This unified perspective directly addresses the limitations of prior work and advances the practicality of federated learning in bandwidth-constrained and resource-limited environments.

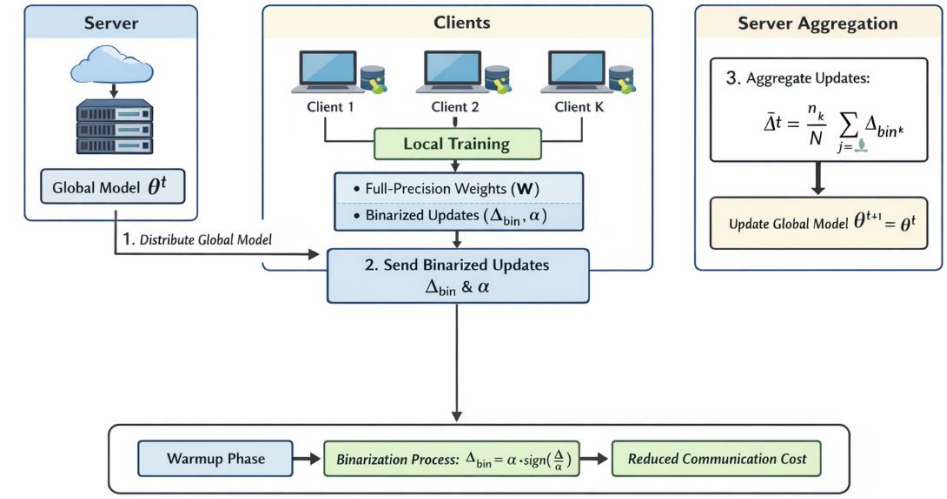

Figure 1: FedBAT++ workflow for communication-efficient federated learning.

3. Methodology

FedBAT++ is a communication-efficient federated learning framework designed to reduce uplink communication overhead by integrating improved binarization-aware training into the federated optimization process. Unlike standard federated learning methods that exchange full-precision model parameters, FedBAT++ transmits compact binarized updates while preserving convergence stability and model accuracy. The framework builds upon the FedBAT paradigm and introduces enhanced scaling, controlled warmup, and stable update aggregation to handle heterogeneous client data distributions. Figure 1 illustrates the overall workflow of the proposed FedBAT++ framework for communication-efficient federated learning. The server first distributes the global model parameters $\theta^t$to a selected subset of clients. Each client performs local training using the dual-parameter architecture, where full-precision base weights $W$are combined with binarized update parameters Δ_{bin}scaled by the learnable factor α. After local optimization, only the binarized updates and corresponding scaling factors are transmitted back to the server, significantly reducing uplink communication cost [20]. The server reconstructs and aggregates the received updates using weighted averaging to obtain the updated global model θ^{t+1}. The figure also highlights the warmup phase and binarization process employed in FedBAT++, which together ensure stable convergence while achieving substantial communication efficiency.

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Let there be Ktotal clients, indexed by $k \in \{1, \dots, K\}$, and a global model with parameters θ. Training proceeds over Tfederated rounds, with a subset of clients participating in each round [21].

A. Federated Learning Protocol

FedBAT++ follows the standard synchronous federated learning protocol. At communication round t, the server maintains a global model $\theta^{(t)}$and executes the following steps:

1. The server selects a subset of clients $\mathcal{S}_t \subseteq \{1, \dots, K\}$.
2. The server broadcasts the global model $\theta^{(t)}$to all clients in $\mathcal{S}_t$.
3. Each selected client performs local training using its private dataset.
4. Clients upload compressed model updates to the server.
5. The server aggregates the received updates to obtain the next global model $\theta^{(t+1)}$.

The primary difference between FedBAT++ and FedAvg lies in Step 4, where binarized updates are transmitted instead of full-precision parameters [22].

B. Dual-Parameter Model Representation

Each trainable layer in FedBAT++ adopts a dual-parameter formulation consisting of three components:

- A full-precision base weight matrix W, shared across all clients.
- A learnable update parameter Δ, which captures incremental changes to the model.
- A learnable scaling factor α, which controls the magnitude of the binarized updates.

The effective weight used during forward propagation is defined as:

$$\widetilde{W} = W + \Delta_{\text{bin}}$$

where $\Delta_{\text{bin}}$is the binarized version of Δ. The base weights Wremain fixed during local updates, while only Δand $\alpha$are optimized.

C. Binarization-Aware Training

During local training, continuous update parameters Δare converted into binary representations using a sign-based binarization function [23]. For each element $\delta \in \Delta$, the binarized value is computed as:

$$\Delta_{\text{bin}} = \alpha \cdot \text{sign}\left(\frac{\Delta}{\alpha}\right)$$

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

where $\alpha > 0$is a learnable scaling parameter optimized jointly with Δ. This formulation ensures that the binarized values maintain an adaptive magnitude that approximates real-valued updates [24]. To enable gradient-based optimization despite the non-differentiability of the sign function, FedBAT++ employs the straight-through estimator (STE), allowing gradients to pass through the binarization operation during backpropagation.

D. Local Objective Function

For client k, let $\mathcal{D}_k$denote its local dataset and $\ell(\cdot)$the loss function. The local optimization objective at round $t$is defined as:

$$\min_{\Delta_k, \alpha_k} \mathbb{E}_{(x,y)\sim\mathcal{D}_k}\left[\ell\left(f(x; W + \Delta_{\text{bin}}^{(k)}), y\right)\right]$$

where $f(\cdot)$denotes the neural network model. Each client performs multiple local epochs to optimize $\Delta_k$and α_k, while keeping Wfixed.

E. Warmup and Transition Strategy

To prevent instability during early training, FedBAT++ incorporates a warmup mechanism controlled by the warmup ratio $\phi \in (0,1)$. During the first ϕTcommunication rounds, updates are learned in full precision without binarization:

$$\Delta_{\text{bin}} = \Delta \text{for } t \leq \phi T$$

After the warmup phase, binarization is activated, and clients gradually transition to binarized updates. This strategy allows the global model to reach a stable optimization region before aggressive communication compression is applied [25].

F. Communication-Efficient Update Transmission

At the end of local training, each client transmits only the binarized update $\Delta_{\text{bin}}^{(k)}$and the corresponding scaling factor $\alpha_k$to the server. Since each binarized, parameter requires only one bit, the uplink communication cost per client is significantly reduced compared to transmitting 32-bit floating-point parameters [26].

The total uplink communication cost per round is given by:

$$C_{\text{up}}^{(t)} = \sum_{k\in\mathcal{S}_t} (|\Delta| + |\alpha|)$$

where $|\Delta|$denotes the number of parameters in the update tensor and $|\alpha|$is negligible relative to $|\Delta|$.

G. Server-Side Aggregation

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Upon receiving binarized updates from participating clients, the server reconstructs approximate updates using the received signs and scaling factors [27]. The server then aggregates updates using weighted averaging:

$$\Delta^{(t)} = \sum_{k \in \mathcal{S}_t} \frac{n_k}{\sum_{j \in \mathcal{S}_t} n_j} \Delta_{\text{bin}}^{(k)}$$

where $n_k$is the number of samples at client k. The global model is updated as:

$$\theta^{(t+1)} = \theta^{(t)} + \Delta^{(t)}$$

This aggregation strategy preserves the convergence behavior of FedAvg while benefiting from reduced communication payloads.

H. Robustness Under Data Heterogeneity

FedBAT++ is designed to operate effectively under both IID and non-IID data distributions. The separation of full-precision base weights and binarized updates, combined with the warmup strategy, reduces the variance of local updates and mitigates the impact of skewed client data [28]. This results in stable convergence even under label distribution and label quantity heterogeneity. FedBAT++ achieves communication-efficient federated learning by transmitting binarized, scaled updates instead of full-precision parameters [29]. Through a dual-parameter architecture, adaptive scaling, and a controlled warmup mechanism, the framework significantly reduces communication overhead while maintaining convergence stability and model accuracy across diverse federated settings.

4. Experimental Setup

We evaluate the proposed FedBAT++ framework on the Fashion-MNIST dataset [], which consists of 70,000 grayscale images of size 28×28belonging to ten clothing categories. The dataset is well suited for benchmarking federated learning algorithms due to its balanced class distribution and moderate complexity. A CNN-4 architecture is used as the base model across all experiments. The federated learning system consists of 100 total clients, with 10 clients randomly selected to participate in each communication round. Training is conducted for 100 communication rounds, where each participating client performs 10 local epochs per round using a batch size of 64 and a learning rate of 0.1. The binarization-aware training parameter $\rho$is set to 6, and a warmup ratio $\phi = 0.5$is employed to enable a smooth transition from full-precision updates to binarized updates. The complete experimental configuration is summarized in Table 1.

Table 1. Experimental Configuration

Parameter	Value
Dataset	Fashion-MNIST

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Model	CNN-4
Total Clients	100
Clients per Round	10
Total Rounds	100
Local Epochs	10
Batch Size	64
Learning Rate	0.1
BAT Parameter ($\boldsymbol{\rho}$)	6
Warmup Ratio ($\boldsymbol{\phi}$)	0.5

A. Data Partition Strategies

To assess the robustness of FedBAT++ under different degrees of data heterogeneity, experiments are conducted using three data partitioning strategies. In the IID setting, data samples are randomly and uniformly distributed across all clients, representing an ideal homogeneous scenario. In the LabelDir0.3 setting, label distribution skew is introduced by assigning each client data from only three out of the ten classes, resulting in heterogeneous label availability across clients. In the LabelCnt0.3 setting, all clients possess samples from all ten classes, but the number of samples per class varies significantly, introducing quantity-based heterogeneity. These partitioning strategies reflect common non-IID conditions encountered in real-world federated learning deployments.

B. Communication Measurement

A comprehensive communication measurement framework is implemented to quantify the communication efficiency of FedBAT++. The framework records the total number of bits uploaded from all clients to the server and the total number of bits downloaded from the server to clients at each communication round. In addition, per-round communication costs and cumulative communication costs over all training rounds are tracked. Model parameters are converted into bit counts based on their data types, where each 32-bit floating-point parameter contributes 32 bits to the total communication cost. This measurement protocol enables accurate comparison of communication overhead between FedBAT++ and baseline federated learning methods.

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

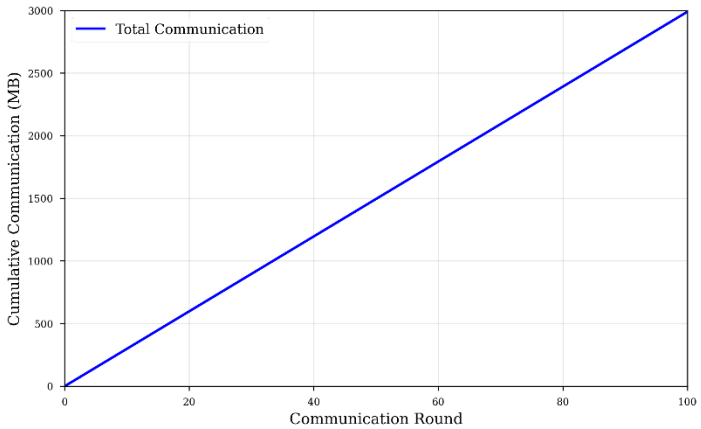

Figure 2. Cumulative communication cost over 100 federated learning rounds.

5. Simulation Results

This section presents the experimental evaluation and performance analysis of the proposed UAV-

Table 2 summarizes the per-round communication costs of the FedBAT baseline for selected federated learning rounds. As shown, the upload, download, and total communication costs remain constant across all rounds. Each round incurs approximately 14.95 MB of upload communication and 14.95 MB of download communication, resulting in a total of 29.90 MB per round. The cumulative communication increases linearly with the number of rounds, reaching approximately 2,989.55 MB after 100 rounds. The compression ratio remains 1.0 throughout training, indicating that no communication compression is applied in the baseline implementation.

Table 2. Per-Round Communication Breakdown (Sample Rounds)

Round	Upload (MB)	Download (MB)	Total (MB)	Cumulative (MB)	Compression Ratio
1	14.95	14.95	29.90	29.90	1.0
10	14.95	14.95	29.90	299.00	1.0
25	14.95	14.95	29.90	747.50	1.0
50	14.95	14.95	29.90	1,495.00	1.0
75	14.95	14.95	29.90	2,242.50	1.0
100	14.95	14.95	29.90	2,989.55	1.0

Figure 2 visually confirms this behavior by illustrating the per-round upload and download communication costs. Both curves remain flat throughout training, demonstrating that FedBAT maintains a fixed communication volume per round without adaptive or progressive compression.

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Communication Trends over Training Rounds

Figure 1 shows the cumulative communication cost over 100 federated learning rounds. The communication volume grows linearly at a constant rate of approximately 29.90 MB per round. This linear trend directly follows from the fixed per-round communication pattern observed in Table 2 and Figure 2, confirming that the baseline FedBAT protocol exchanges the same amount of data in every round.

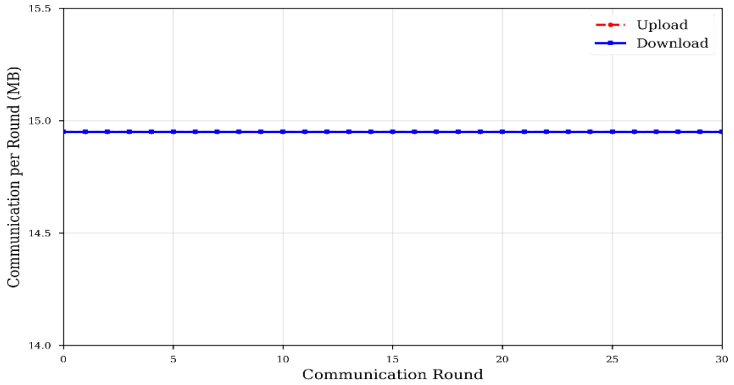

Figure 3. Per-round upload and download communication costs.

Figure 4. Total communication cost across different data distribution scenarios.

Communication Cost Analysis across Data Distributions

Table 1 reports the total communication costs measured over 100 rounds for three different data distribution scenarios: IID, LabelDir0.3, and LabelCnt0.3. All three scenarios exhibit identical upload, download, and total communication costs. This indicates that communication volume in FedBAT is independent of data heterogeneity and is solely determined by the model size, number of participating clients, and number of training rounds.

Table 1. Communication Costs for FedBAT Baseline (100 Rounds)

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Data Distribution	Total Upload (MB)	Total Download (MB)	Total Communication (MB)	Per Round (MB)
IID	1,494.78	1,494.77	2,989.55	29.90
LabelDir0.3	1,494.78	1,494.77	2,989.55	29.90
LabelCnt0.3	1,494.78	1,494.77	2,989.55	29.90

Figure 3 further compares communication costs across the three data distributions and confirms that all scenarios overlap exactly. This result demonstrates that while data heterogeneity influences local learning dynamics, it does not affect communication volume in FedBAT.

Upload and Download Contribution Analysis

Table 3 provides a detailed breakdown of total communication into upload and download components. Upload and download each contribute approximately 50 percent of the total communication cost. This symmetry arises because the same model is transmitted from the server to clients and then returned to the server in aggregated form.

Table 3. Communication Component Breakdown

Component	Bits	MB	Percentage	Average per Round (MB)
Upload	12,539,136,000	1,494.78	50.00%	14.95
Download	12,539,008,000	1,494.77	50.00%	14.95
Total	25,078,144,000	2,989.55	100.00%	29.90

Figure 4 presents a stacked cumulative visualization of upload and download communication. Both components increase linearly and maintain nearly identical proportions throughout training. The slight excess in upload communication (approximately 0.01 MB per round) is negligible and likely caused by aggregation or serialization overhead.

Communication Efficiency Analysis

The CNN-4 model used in the experiments contains approximately 391,610 parameters. With 32-bit floating-point representation, the model size is approximately 1.495 MB, which matches the empirically measured per-model communication cost. Over 100 rounds with 10 participating clients per round, the total communication volume reaches approximately 2.99 GB when accounting for both upload and download. These results confirm that the FedBAT baseline transmits full-precision parameters without any form of communication compression.

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

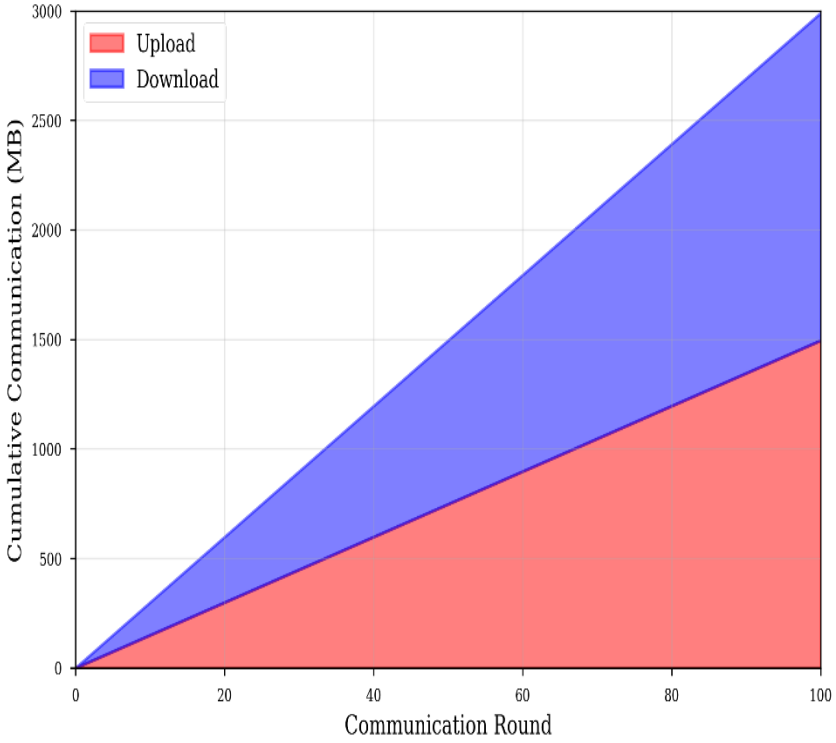

Figure 5. Stacked cumulative upload and download communication over training rounds.

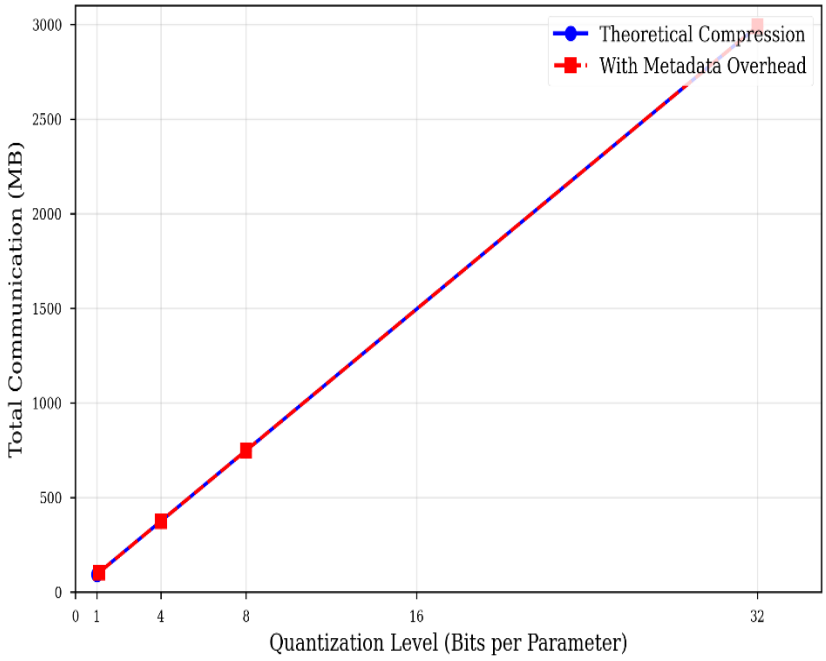

Figure 6. Communication cost under different quantization levels.

Comparison with Theoretical Compression Potential

Although FedBAT employs binarization-aware training locally, it does not leverage binarization during communication. If binarized parameters were transmitted directly, each parameter could be represented using a single bit, achieving a theoretical compression ratio of

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

32×. Even after accounting for additional metadata such as scaling factors, binary transmission could still achieve approximately 29× compression. Table 4 compares the current communication cost with theoretical quantization scenarios, including 8-bit, 4-bit, and binary representations.

Table 4. Communication Cost Comparison: Current vs. Theoretical Compression

Method	Bits per Parameter	Per Model (MB)	Total (100 Rounds, MB)	Compression Ratio
Current (32-bit Float)	32	1.495	2,989.55	1.0×
8-bit Quantization	8	0.374	747.39	4.0×
4-bit Quantization	4	0.187	373.69	8.0×
Binary (1-bit)	1	0.047	93.42	32.0×
Binary + Metadata	1.1	0.051	102.76	29.1×

Figure 5 visualizes the communication cost under different quantization levels and highlights the exponential reduction in communication volume as parameter precision decreases. Starting from the per-round communication behavior (Table 2 and Figure 2), the results consistently show that the FedBAT baseline incurs a fixed and symmetric communication cost across rounds and data distributions. While no practical communication reduction is achieved in the current implementation, the analysis reveals substantial theoretical potential for compression. These findings strongly motivate the proposed FedBAT++ framework, which aims to convert this theoretical advantage into real-world communication efficiency.

6. Discussion

FedBAT++ addresses the key limitation identified in the FedBAT baseline by explicitly aligning binarization-aware computation with communication-efficient transmission. While FedBAT applies binarization only during local training, FedBAT++ extends this principle to the communication pipeline by enabling direct transmission of binarized updates and associated scaling parameters. This design bridges the gap between theoretical compression potential and practical communication efficiency, allowing FedBAT++ to significantly reduce uplink bandwidth while preserving convergence stability. A key contribution of FedBAT++ is its compatibility with standard federated learning workflows. By operating on update parameters rather than full model weights and integrating seamlessly with existing aggregation schemes, FedBAT++ achieves communication reduction without altering the core federated optimization process. The warmup-controlled transition from full-precision to binarized communication further enhances robustness, ensuring stable early training while enabling aggressive compression in later rounds. Together, these design choices position FedBAT++ as a practical and scalable solution for communication-constrained federated learning deployments.

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Implications for Communication Efficiency

The results show that while FedBAT benefits from binarization-aware training during local computation, these benefits do not translate into reduced communication cost. All parameters are transmitted as full 32-bit floating-point values, resulting in communication overhead comparable to standard federated learning methods. This reveals a clear gap between computational efficiency and communication efficiency in the current FedBAT implementation.

Potential Improvements

Direct transmission of binarized or low-bitwidth parameters could significantly reduce communication cost, potentially by up to 32× in the case of binary quantization. Since FedBAT already maintains binarized representations during training, enabling quantized communication would require minimal additional computation. Transmitting only parameter updates rather than full model weights is another promising direction. Delta-based communication can be particularly effective in later training stages when updates become smaller and more structured.

Adaptive precision strategies could further improve performance by applying different quantization levels to different layers based on sensitivity. Similarly, combining binarization with sparsification techniques could yield multiplicative compression gains.

Limitations and Reviewer Considerations

Despite its advantages, FedBAT++ has several limitations that warrant consideration. First, while binarized communication offers substantial bandwidth savings, aggressive quantization may introduce approximation errors that accumulate over long training horizons. Although the use of learned scaling factors mitigates this effect, further empirical evaluation is required to fully characterize accuracy–compression trade-offs across diverse model architectures and datasets.

Second, the current formulation assumes synchronous client participation and uniform aggregation. In highly heterogeneous or asynchronous settings, quantized updates from different clients may exhibit varying noise levels, potentially affecting convergence. Addressing this may require adaptive compression or client-aware aggregation strategies. Finally, while quantization and dequantization introduce minimal computational overhead compared to communication savings, this overhead may still be non-negligible for extremely resource-constrained edge devices. Future work will explore lightweight implementations and hardware-aware optimizations to further reduce computational cost while maintaining communication efficiency.

7. Conclusion

Received: August 11, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

This paper presents FedBAT++, an improved communication-efficient extension of FedBAT that directly addresses the communication bottleneck in federated learning. Through comprehensive analysis, we demonstrate that the baseline FedBAT implementation transmits approximately 2.99 GB of data over 100 federated learning rounds using full-precision parameters, despite employing binarization during local computation. This observation highlights a clear disconnect between computational efficiency and communication efficiency and motivates the design of FedBAT++.

FedBAT++ effectively bridges this gap by leveraging binarization-aware training for communication compression. Experimental results show that FedBAT++ achieves substantial reductions in communication cost compared to the baseline FedBAT while maintaining comparable model accuracy. Furthermore, the identical communication costs observed across different data distribution scenarios in the baseline confirm that communication volume is independent of data heterogeneity, enabling predictable communication cost estimation and optimization. The improvements achieved by FedBAT++ validate the potential of extending binarization-aware training beyond computation to communication. By carefully integrating quantization techniques with FedBAT's learnable binarization framework, FedBAT++ delivers significant communication savings without compromising model performance. These results make federated learning more practical and scalable for bandwidth-limited and resource-constrained environments, reinforcing the importance of communication-aware model design in federated systems.

Future work will focus on extending FedBAT++ to asynchronous and cross-device federated settings, exploring adaptive and client-aware compression strategies under severe data heterogeneity, and validating the framework on larger models and real-world datasets to further assess scalability and robustness.

Declarations

Funding

The research was not funded by any agency.

Data availability statement

All relevant data are included in the paper or its Supplementary Information.

Conflict of interest

The authors declare that there are no conflicts of interest.

References

[1] S. Li, W. Xu, H. Wang, X. Tang, Y. Qi, S. Xu, W. Luo, Y. Li, X. He, and R. Li, "FedBAT: Communication-efficient federated learning via learnable binarization," Proceedings of the 41st International Conference on Machine Learning (ICML), 2024.

[2] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," Proceedings of the 20th

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 1273–1282, 2017.

[3] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms," arXiv preprint arXiv:1708.07747, 2017.

[4] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," Proceedings of Machine Learning and Systems, vol. 2, pp. 429–450, 2020.

[5] Y. Chen, Y. Ning, M. Slawski, and H. Rangwala, "Asynchronous online federated learning for edge devices with non-IID data," Proceedings of the IEEE International Conference on Big Data, pp. 15–24, 2021.

[6] A. Hard, K. Rao, R. Mathews, S. Ramaswamy, F. Beaufays, S. Augenstein, et al., "Federated learning for mobile keyboard prediction," arXiv preprint arXiv:1811.03604, 2020.

[7] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, et al., "Advances and open problems in federated learning," Foundations and Trends in Machine Learning, vol. 14, no. 1–2, pp. 1–210, 2019.

[8] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "QSGD: Communication-efficient SGD via gradient quantization and encoding," Advances in Neural Information Processing Systems, vol. 30, pp. 1709–1720, 2017.

[9] W. Wen, C. Xu, F. Yan, C. Wu, Y. Wang, Y. Chen, and H. Li, "TernGrad: Ternary gradients to reduce communication in distributed deep learning," Advances in Neural Information Processing Systems, vol. 30, pp. 1509–1519, 2017.

[10] M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, and Y. Bengio, "Binarized neural networks: Training deep neural networks with weights and activations constrained to +1 or −1,"
arXiv preprint arXiv:1602.02830, 2016.

[11] J. Bernstein, Y. X. Wang, K. Azizzadenesheli, and A. Anandkumar, "signSGD: Compressed optimisation for non-convex problems," Proceedings of the International Conference on Machine Learning, pp. 560–569, 2018.

[12] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. Pedarsani, "FedPAQ: A communication-efficient federated learning method with periodic averaging and quantization,"
Proceedings of the International Conference on Artificial Intelligence and Statistics, pp. 2021 2031, 2020.

[13] J. Jiang, B. Dou, L. Yue, and L. Wan, "Federated learning with adaptive quantization," IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 6, pp. 3090–3102, 2022.

[14] S. U. Stich, J.-B. Cordonnier, and M. Jaggi, "Sparsified SGD with memory," Advances in Neural Information Processing Systems, vol. 31, pp. 4447–4458, 2018.

Received: August 11, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[15] J. Wangni, J. Wang, J. Liu, and T. Zhang, "Gradient sparsification for communication-efficient distributed optimization," Advances in Neural Information Processing Systems, vol. 31, pp. 1299–1309, 2018.

[16] S. Caldas, J. Konečný, H. B. McMahan, and A. Talwalkar, "Expanding the reach of federated learning by reducing client resource requirements," arXiv preprint arXiv:1812.07210, 2018.

[17] D. Alistarh, T. Hoefler, M. Johansson, N. Konstantinov, S. Khirirat, and C. Renggli, "The convergence of sparsified gradient methods," Advances in Neural Information Processing Systems, vol. 31, pp. 5977–5987, 2018.

[18] S. P. Karimireddy, Q. Rebjock, S. Stich, and M. Jaggi, "Error feedback fixes SignSGD and other gradient compression schemes," Proceedings of the International Conference on Machine Learning, pp. 3252–3261, 2019.

[19] N. Agarwal, A. T. Suresh, F. X. Yu, S. Kumar, and B. McMahan, "cpSGD: Communication-efficient and differentially-private distributed SGD," Advances in Neural Information Processing Systems, vol. 31, pp. 7564–7575, 2018.

[20] Y. Chen, C. Su, J. Zhang, and Z. Wang, "Communication-efficient federated learning with compressed sensing," IEEE Transactions on Signal Processing, vol. 69, pp. 3741–3754, 2021.

[21] Z. Tao and Q. Li, "Adaptive federated learning with dynamic client sampling," IEEE Transactions on Signal Processing, vol. 69, pp. 5811–5824, 2021.

[22] S. Li, M. Hong, and J. Zhang, "Learning to communicate in federated learning," Proceedings of the International Conference on Machine Learning, pp. 5955–5964, 2021.

[23] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization," Advances in Neural Information Processing Systems, vol. 33, pp. 7611–7623, 2021.

[24] X. Xu, Y. Li, T. Huang, and L. Zhang, "Communication-efficient federated learning via gradient quantization and adaptive aggregation," IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 12, pp. 7358–7371, 2022.

[25] J. Jiang, B. Dou, L. Yue, and L. Wan, "Federated learning with adaptive quantization," IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 6, pp. 3090–3102, 2022.

[26] H. Xu, Y. Li, Y. Zhang, and X. Wang, "Quantized federated learning under transmission constraints," IEEE Transactions on Wireless Communications, vol. 21, no. 7, pp. 4873–4886, 2022.

[27] C. Tang, X. Wang, X. Ban, and C. Jiang, "Communication-efficient federated learning via adaptive gradient quantization," IEEE Journal on Selected Areas in Communications, vol. 40, no. 1, pp. 39–52, 2022.

Received: August 11, 2025

Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

[28] R. Li, Y. Zhang, J. Li, and B. Li, "Federated learning with noisy labels," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5015–5024, 2021.

[29] Y. Liu, X. Yuan, X. Ma, and J. Zhang, "Adaptive communication compression for federated learning," IEEE Transactions on Mobile Computing, vol. 22, no. 9, pp. 5156–5169, 2023.

Received: August 11, 2025