Proceedings of the 4th International Conference on Mechatronics and Smart Systems
DOI: 10.54254/2755-2721/2026.33415

Generative AI-Driven Adaptive Compression for Federated Learning

Huayang Jin

College of Computer Science, Sichuan Normal University, Chengdu, China
2023110507@stu.sicnu.edu.cn

Abstract. Federated Learning (FL) faces severe communication overhead due to massive gradient and parameter transmission, especially in heterogeneous edge environments. Static compression methods with fixed ratios fail to adapt to dynamic network conditions and device differences, leading to severe accuracy degradation or inefficient compression. This paper proposes a generative AI-driven adaptive compression framework for FL, named GAN-ACFL. By constructing a 12-dimensional system state feature set and training a GAN-based decision model, the framework generates personalized compression ratios and sparse masks according to real-time device resources, network latency, and data non-IID degree. Experiments on the MNIST, Fashion-MNIST, and CIFAR-10 show that GAN-ACFL reduces communication traffic by 68% on average, while maintaining 97.8%, 91.9%, and 82.3% accuracy respectively. It outperforms fixed quantization, sparsification, and heuristic adaptive methods in balancing communication efficiency and model performance. The proposed method exhibits strong adaptability to heterogeneous devices and dynamic networks, making it suitable for large-scale FL deployment.

Keywords: federated learning, generative AI, adaptive compression, communication efficiency

1. Introduction

Federated Learning (FL) has become a popular distributed learning framework that allows multiple edge devices to train models collaboratively without uploading raw private data [1]. This privacy-preserving mechanism makes FL widely used in edge computing, IoT, and mobile applications. However, frequent communication between edge devices and the central server brings heavy communication overhead, especially when devices have limited bandwidth, unstable networks, or different computing capabilities. Large-scale gradient and parameter uploading often lead to slow training, high latency, and even training failure [2].

To reduce communication costs, compression methods such as quantization, sparsification, and pruning have been applied in FL systems [2,3]. Most of these methods use fixed compression strategies for all devices. Unfortunately, fixed compression cannot adapt to dynamic network changes and device heterogeneity. For example, a compression ratio that works well on high-performance devices may cause severe accuracy degradation on low-resource devices. Similarly, a stable network can tolerate lower compression, while a crowded network requires higher

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.33415

compression. Such mismatches limit the performance of traditional compression methods in real FL scenarios [4].

In recent years, generative AI has shown strong ability in modeling complex distributions and making adaptive decisions [5]. Unlike heuristic or rule-based methods, generative models can learn hidden relationships between system states and optimal compression policies from real-time data. Inspired by this, a generative AI-based adaptive compression method for FL is introduced in this study. A GAN-driven decision model is designed to generate personalized compression ratios and sparse masks according to device resources, network latency, and data non-IID conditions. This data-driven approach avoids manual rule design and improves adaptability.

In this work, a 12-dimensional system state feature set is first defined to represent real-time conditions of FL clients. Then a GAN-based generator and discriminator are built to learn optimal compression strategies. The proposed method is integrated into the classic FedAvg framework [1] and named GAN-ACFL. Experiments on MNIST, Fashion-MNIST, and CIFAR-10 show that the proposed method reduces communication traffic by 68% on average while maintaining high model accuracy. Compared with fixed quantization, fixed sparsification, and heuristic adaptive compression, GAN-ACFL achieves a better balance between communication efficiency and model performance. It also shows strong robustness in heterogeneous device environments [6].

The rest of this paper is organized as follows. Section 2 describes the dataset preparation, GAN-based compression strategy, overall framework, and experimental settings. Section 3 presents and analyzes the experimental results in terms of communication overhead, model accuracy, and heterogeneous adaptability. Section 4 concludes the paper and discusses future work.

2. Method

2.1. Dataset preparation

Three representative public datasets are adopted in this study to verify the universality of the proposed method, covering typical application scenarios of federated learning. MNIST is used as the basic benchmark dataset, which consists of 60,000 training samples and 10,000 test samples of handwritten digits. CIFAR-10 is employed to evaluate complex image classification tasks, containing 50,000 color training images and 10,000 test images divided into 10 categories. Fashion-MNIST is introduced as a supplementary dataset, which has the same structure as MNIST but focuses on fashion product classification.

To simulate real heterogeneous federated learning environments, all datasets are allocated to 100 edge devices in a non-IID manner based on Dirichlet distribution (α=0.5), so that each device only holds data of partial categories, which is consistent with the data island characteristics in real-world applications [6,7]. This non-IID setting ensures that the experimental environment is close to practical FL deployment scenarios, making the evaluation results more reliable and convincing.

2.2. GAN-based compression strategies

A compression decision model based on Generative Adversarial Network (GAN) is designed in this study, which consists of a generator and a discriminator [5]. The generator takes real-time system state features as input, including network latency, bandwidth, device computing power, data non-IID degree, gradient norm and local accuracy improvement. After normalization, these features form a 12-dimensional state vector to be mapped into personalized compression ratios and sparse masks.

DOI: 10.54254/2755-2721/2026.33415

The generator adopts a three-layer MLP structure with ReLU activation, outputting compression ratios between 0.1 and 0.9. The discriminator evaluates the quality of generated strategies by distinguishing between optimal manual compression schemes and generated ones. The loss function includes two components: adversarial loss ensures the generated compression strategies fit the optimal distribution, and performance loss constrains the model accuracy drop caused by compression.

Through end-to-end adversarial training, the generator can implicitly learn the correlation between system states and optimal compression decisions, so as to generate adaptive strategies without manual rules. This data-driven approach overcomes the limitations of traditional rule-based adaptive methods [8].

2.3. Federated learning framework

The proposed method is integrated into the classical FedAvg framework [1] to form GAN-driven Adaptive Compression Federated Learning (GAN-ACFL). In each communication round, edge devices first perform local SGD training on their private data. Then each device extracts real-time system state features and inputs them into the pre-trained generator to obtain personalized compression strategies.

Each edge device compresses gradient updates according to the generated compression ratio and sparse mask, and uploads the compressed data to the central server. The server aggregates all received updates by weighted averaging based on sample sizes, then distributes the updated global model back to all devices.

The GAN decision model is updated periodically on the server using collected state and performance feedback, ensuring compression strategies continuously adapt to dynamic network conditions and device changes [4,9]. Unlike fixed compression methods, high-resource devices use lower compression ratios to retain more model information, while low-resource devices use higher ratios to reduce transmission burden.

2.4. Experimental hyperparameter configuration

The FL training process uses SGD optimizer with a learning rate of 0.01, momentum of 0.9 and weight decay 1e-4. Local epoch is set to 5, total communication rounds to 100, and local batch size to 32. For the GAN model, Adam optimizer is used with learning rate 0.0002, and the balance coefficient λ in the composite loss function is set to 0.5.

The number of simulated edge devices is 100, with 10 devices randomly selected to participate in each communication round [9]. Baseline comparison methods include fixed 4-bit quantization, fixed 8-bit quantization, 50% sparsification and heuristic adaptive compression [2,3,8,10]. All experiments are repeated 5 times to take average values, implemented in PyTorch 1.13 with CUDA GPU acceleration to ensure experimental stability and reliability.

3. Results and discussion

3.1. Communication overhead comparison

Experimental results demonstrate that the proposed GAN-ACFL method significantly reduces communication overhead while maintaining model accuracy. Table 1 shows the average communication traffic reduction rates of different methods compared with full-precision uncompressed FL.

DOI: 10.54254/2755-2721/2026.33415

Table 1. Communication overhead comparison of different compression methods

Method	Average Communication Reduction
Uncompressed FL	0%
Fixed 4-bit quantization	75%
Fixed 8-bit quantization	50%
Fixed 50% sparsification	50%
Heuristic adaptive	58%
GAN-ACFL (Proposed)	68%

The proposed method reduces communication traffic by 68% on average, which is close to the effect of 4-bit fixed quantization (75% reduction) but with much less accuracy loss. Fixed sparsification only achieves 50% compression, which is weaker than the adaptive method. The adaptive mechanism dynamically adjusts strategies with network changes: when latency rises, compression ratio increases automatically to ensure transmission; when bandwidth is sufficient, compression is relaxed to protect model information [4,8].

3.2. Model accuracy performance

The classification accuracy results on three datasets are shown in Table 2. On MNIST, uncompressed FedAvg reaches 98.2% accuracy, while fixed 4-bit quantization drops to 94.5%. GAN-ACFL stabilizes at 97.8%, only 0.4% lower than the full-precision model. On CIFAR-10, GAN-ACFL achieves 82.3% accuracy, 3.7% higher than fixed sparsification and 5.1% higher than fixed quantization.

Table 2. The performance of model on different datasets

Method	MNIST	Fashion-MNIST	CIFAR-10
Uncompressed FL	92.8	92.5	83.7
Fixed 4-bit quantization	94.5	88.2	77.2
Fixed 8-bit quantization	96.1	89.8	79.5
Fixed 50% sparsification	95.3	89.1	78.6
Heuristic adaptive	96.8	90.5	80.2
GAN-ACFL (Proposed)	97.8	91.9	82.3

The convergence curve shows that the adaptive method converges as fast as uncompressed FL, while fixed compression methods slow down obviously. The reason is that GAN avoids excessive compression on important gradients and only eliminates redundant information, thus balancing compression and accuracy effectively [3,10].

3.3. Heterogeneous environment adaptability

Under heterogeneous device resource settings, traditional fixed compression causes a 12.3% accuracy gap between high-resource and low-resource devices. GAN-ACFL reduces this gap to only 2.1%, showing strong adaptability to system heterogeneity. Low-bandwidth devices can successfully upload updates with high compression ratios, avoiding transmission failure in fixed strategies [4,9].

DOI: 10.54254/2755-2721/2026.33415

The method also maintains stable performance under frequent network fluctuations, with accuracy variance less than 0.5% during continuous network changes. This robustness ensures reliable deployment in real dynamic FL systems. The data-driven strategy generation outperforms heuristic rules in heterogeneous scenarios by learning implicit patterns from system states [8].

3.4. Comprehensive discussion

The experimental results verify that generative AI can effectively improve the adaptability of FL communication compression. Compared with static or rule-based methods, GAN learns implicit correlations between system states and optimal compression decisions, providing more reasonable personalized strategies [5].

The core advantage of the proposed method is achieving a better trade-off between communication overhead reduction and model accuracy preservation. The slight limitation is that periodic GAN updating brings small computing overhead to the server. In future work, lightweight generative models will be explored for edge deployment, and the framework will be extended to cross-silo FL and federated reinforcement learning scenarios [7,11].

4. Conclusion

This paper proposes GAN-ACFL, a generative AI-driven adaptive compression framework for federated learning. The GAN-based decision model generates dynamic compression strategies based on real-time system states, effectively reducing communication overhead while preserving model accuracy. Experimental results on three datasets demonstrate that the proposed method reduces communication traffic by 68% with negligible accuracy loss, and significantly outperforms static and heuristic compression methods. GAN-ACFL shows strong adaptability to device heterogeneity and network dynamics, narrowing the accuracy gap between different devices to 2.1%. In the future, lightweight deployment of generative models will be optimized, and the framework will be extended to cross-silo FL and federated reinforcement learning.

References

[1] McMahan, B., Moore, E., Ramage, D., Hampson, S. and Agüera y Arcas, B. (2017) Communication-efficient learning of deep networks from decentralized data. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 54, 1273–1282.
[2] Konečný, J., McMahan, B., Yu, F.X., Richtárik, P., Suresh, A.T. and Bacon, D. (2016) Federated learning: Strategies for improving communication efficiency. arXiv preprint arXiv: 1610.05492.
[3] Bernstein, J., Zhao, J., Azizzadenesheli, K. and Anandkumar, A. (2018) signSGD: Compressed optimisation for non-convex problems. In Proceedings of the 35th International Conference on Machine Learning, 80, 559–568.
[4] Wang, H., Kaplan, Z., Niu, D. and Li, B. (2020) Optimizing federated learning for non-IID data in wireless networks. IEEE Transactions on Mobile Computing, 20(7), 2457–2471.
[5] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2020) Generative adversarial networks. Communications of the ACM, 63(11), 139–144.
[6] Li, X., Huang, K., Yang, W., Wang, S. and Zhang, Z. (2019) On the convergence of FedAvg on non-IID data. arXiv preprint arXiv: 1907.02189.
[7] Zhang, C., Xie, Y., Bai, H., Yu, B., Li, W. and Gao, Y. (2021) A survey on federated learning. IEEE Internet of Things Journal, 9(1), 1–18.
[8] Chen, Y., Ning, Y. and Slomka, A. (2021) Adaptive communication strategies in federated learning. IEEE Internet of Things Journal, 8(11), 9113–9123.
[9] Liu, L., Zhang, J., Song, S.H. and Letaief, K.B. (2021) Client selection and bandwidth allocation in wireless federated learning networks. IEEE Transactions on Wireless Communications, 20(2), 1105–1121.

DOI: 10.54254/2755-2721/2026.33415

[10] Oh, Y., Lee, N. and Poor, H.V. (2021) Communication-efficient federated learning via quantized compressed sensing. arXiv preprint arXiv: 2111.15071.
[11] Karimireddy, S.P., Kale, S., Mohri, M., Reddi, S., Stich, J. and Suresh, A.T. (2020) SCAFFOLD: Stochastic controlled averaging for federated learning. In Proceedings of the 37th International Conference on Machine Learning, 119, 5132–5143.