

Communication-Efficient and Robust Federated Training of Vision Models on IoT Camera Devices

Chengjun Li

haojuti91876@163.com

Henan University of Science and Technology

Research Article

Keywords: federated learning, privacy-preserving vision, gradient sparsification, non-IID data, edge computing, image classification

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11109342/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Communication-Efficient and Robust Federated Training of Vision Models on IoT Camera Devices

Chengjun Li

*School of Software, Henan University of Science and Technology,
Luoyang, China.

Corresponding author(s). E-mail(s): haojuti91876@163.com;

Abstract

Internet-of-Things camera networks generate massive volumes of visual data, yet privacy regulations increasingly prohibit the centralized collection of raw imagery. Federated learning (FL) offers a natural solution by training vision models directly on devices so that raw images never leave the camera. However, two practical obstacles limit the deployment of federated vision training at the edge: the uplink communication bottleneck caused by transmitting full model updates every round, and the accuracy degradation induced by non-IID visual data distributions across heterogeneous cameras. This paper presents a communication-efficient and robust federated training framework for vision models that combines (1) Top-K gradient sparsification to compress uplink model updates, (2) FedProx proximal regularization to mitigate client drift under non-IID data, and (3) GroupNorm normalization, which we show is essential for stable federated vision training where BatchNorm statistics are corrupted by long-tailed client data volumes. In addition to studying each component in isolation, we analyze their joint deployment under data heterogeneity—a configuration largely absent from prior work. Experiments simulate 50 camera clients on CIFAR-10 and MNIST with a lightweight CNN (0.62M parameters). The framework achieves 78.9% test accuracy on CIFAR-10 with a 67% reduction in uplink communication under non-IID conditions, while the joint FedProx+sparsification configuration recovers +3.89% accuracy compared with sparsification alone under strong heterogeneity. All experiments are reproducible with a fixed random seed.

Keywords: federated learning, privacy-preserving vision, gradient sparsification, non-IID data, edge computing, image classification

1 Introduction

Networked cameras—in smart cities, industrial inspection, retail analytics, and home surveillance—now constitute one of the largest sources of visual data in the world [15]. Training vision models on this data conventionally requires aggregating it in a central cloud, a practice that is increasingly constrained by privacy legislation such as the GDPR and by the sheer cost of transporting raw video. Federated learning (FL) [1, 2] resolves this tension: models are trained collaboratively while the training data never leaves the camera, making FL a canonical paradigm for *privacy-preserving vision* [10, 11].

Despite this promise, two challenges dominate federated training of vision models on IoT camera devices. First, *communication efficiency*: a lightweight CNN contains hundreds of thousands of parameters, and uploading full model updates from every participating camera each round can saturate the uplink of resource-constrained devices over hundreds of rounds [3, 4]. Second, *data heterogeneity*: cameras observe different scenes, viewpoints, lighting, and class distributions, producing non-IID local datasets that cause client drift and slow convergence for standard FedAvg [8, 12].

This paper presents a unified framework addressing both challenges for vision models, with the following contributions:

- (i) A Top-K gradient sparsification scheme applied to CNN model updates, reducing uplink communication while retaining the dominant gradient directions (Section 3).
- (ii) FedProx proximal regularization to bound client drift under non-IID camera data (Section 3).
- (iii) A joint deployment analysis of both techniques under data heterogeneity—showing empirically that sparsification and proximal regularization compose without degrading one another (Section 4), a configuration not covered by prior work.
- (iv) A systematic comparison of BatchNorm and GroupNorm in federated vision training, demonstrating that GroupNorm is essential for convergence when clients exhibit long-tailed data volumes (Section 3.4).
- (v) Full reproducibility: fixed seeds, a public codebase, and both CIFAR-10 [21] and MNIST [22] evaluations with 50 simulated camera clients.

2 Related Work

Communication-efficient FL. Reducing the communication cost of federated training has been studied through quantization [3, 6], sparsification [4, 5, 7], and infrequent aggregation. Our Top-K scheme follows the sparsification line: only the largest-magnitude components of the model update are transmitted together with a binary mask, and we account for mask overhead explicitly in the communication model.

Non-IID robustness. FedAvg [1] degrades under heterogeneous client data. FedProx [8] adds a proximal term to the local objective to keep local models near the global consensus; FedMA [9] and LG-FedAvg [10] tackle heterogeneity through architecture matching and local representation learning, respectively. We adopt FedProx because it is agnostic to the model architecture and composes naturally with sparsification.

Federated learning for vision. Prior work has applied FL to object detection platforms [11], face and medical image analysis, and multimodal perception [15]. Our work contributes to this line by studying the *training-system* side of federated vision: how communication compression and heterogeneity regularization interact in realistic camera-network conditions.

Normalization in federated learning. Hsieh et al. [12] identified BatchNorm as a major source of divergence in decentralized learning over skewed data and recommended GroupNorm [14] as a remedy; FedBN [13] instead keeps BN parameters local. We adopt GroupNorm, which requires no running statistics and is therefore robust to client data heterogeneity by construction; Section 3.4 provides direct empirical evidence in the federated setting.

3 Methodology

3.1 System Model

We consider a camera network of N IoT devices coordinated by a cloud aggregation server (Fig. 1). Device k holds a local visual dataset $\mathcal{D}_k$ of size n_k , sampled from a device-specific distribution P_k . The global objective is

$$\min_w F(w) = \sum_{k=1}^N \frac{n_k}{n} F_k(w), \quad F_k(w) = \mathbb{E}_{(x,y) \sim P_k} [\ell(w; x, y)]. \quad (1)$$

Consistent with Eq. (1), our server aggregates client updates weighted by their local data sizes (the standard FedAvg weighting), which matters in practice because n_k follows a long-tail distribution across cameras.

3.2 Top-K Gradient Sparsification

After E local epochs, client k computes its update $\Delta_k = w_k^t - w_{t-1}$. Instead of transmitting all d parameters ($32d$ bits), the client transmits only the $K\%$ components with the largest absolute value, together with a d -bit binary mask identifying their positions:

$$\text{comm bits per client} = K\% \cdot d \cdot 32 + d. \quad (2)$$

Sparsifying the *update* rather than the weights is crucial: after local training, most parameters change only slightly while a few change substantially, so the update vector is naturally sparse (Fig. 3).

3.3 FedProx Proximal Regularization

Each client minimizes the regularized objective

$$h_k(w) = F_k(w) + \frac{\mu}{2} \|w - w_{t-1}\|^2, \quad (3)$$

where μ bounds how far a local model may drift from the global consensus. Under non-IID data this suppresses client divergence and improves convergence stability.

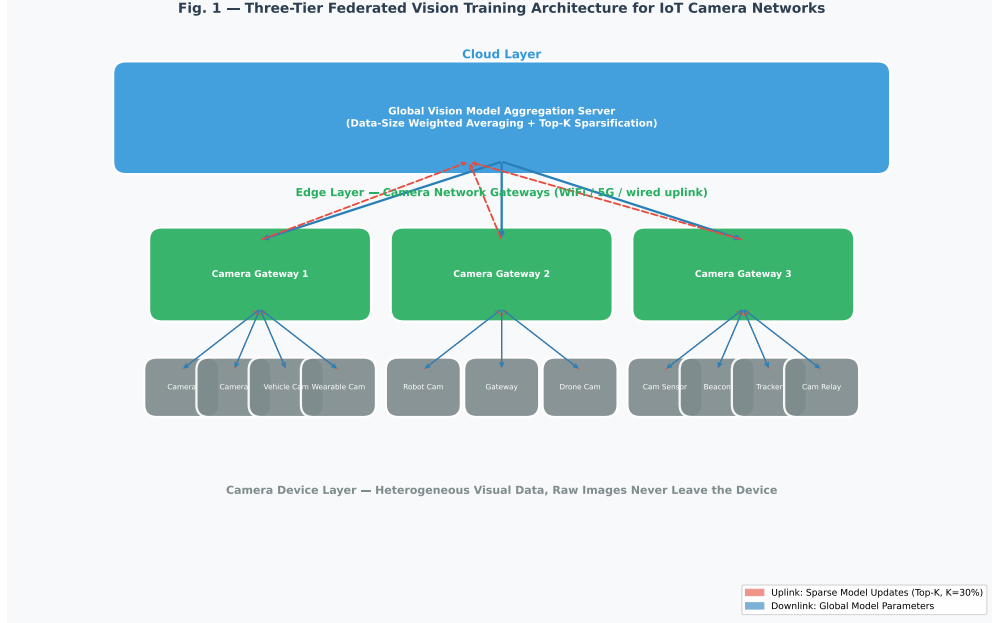


Fig. 1 Three-tier federated vision training architecture for IoT camera networks. Cameras upload only sparse model updates (Top-K, $K = 30\%$); raw images never leave the devices.

3.4 GroupNorm for Federated Vision

BatchNorm maintains running mean/variance statistics that are updated on local client batches. When client data volumes follow a long-tail distribution, clients with very few samples produce noisy BN statistics; these noise-corrupted statistics are then averaged into the global model every round, degrading convergence. GroupNorm normalizes within feature groups and maintains *no* running statistics, making the model state fully determined by the weights—a property well suited to weight-averaging FL. Section 4 reports the empirical magnitude of this effect.

3.5 Joint Deployment

The full pipeline per round (Fig. 2): server selects M clients $\rightarrow$ broadcasts w_{t-1} $\rightarrow$ each client trains E epochs with the FedProx objective $\rightarrow$ each client Top-K sparsifies its update $\rightarrow$ server aggregates the sparsified updates weighted by n_k . A natural concern is whether sparsification amplifies drift under non-IID data, and whether the proximal term compensates; we answer this empirically in Section 4.

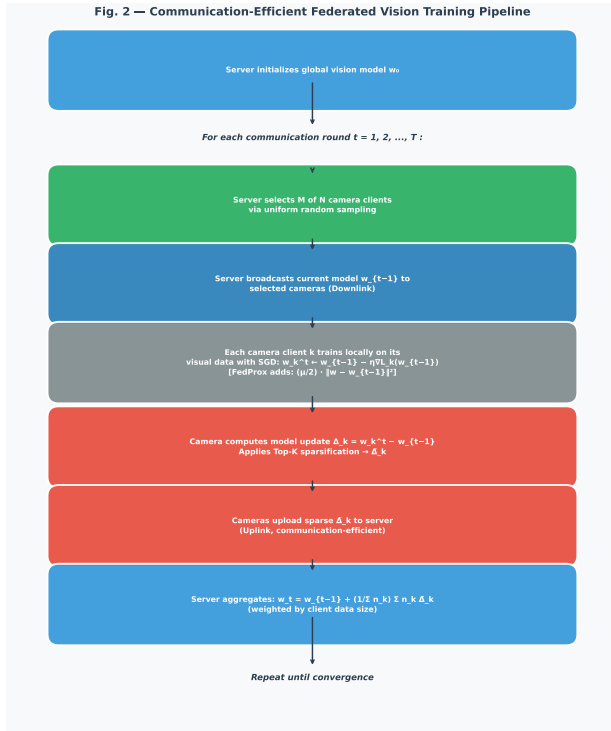


Fig. 2 One communication round of the proposed federated vision training pipeline.

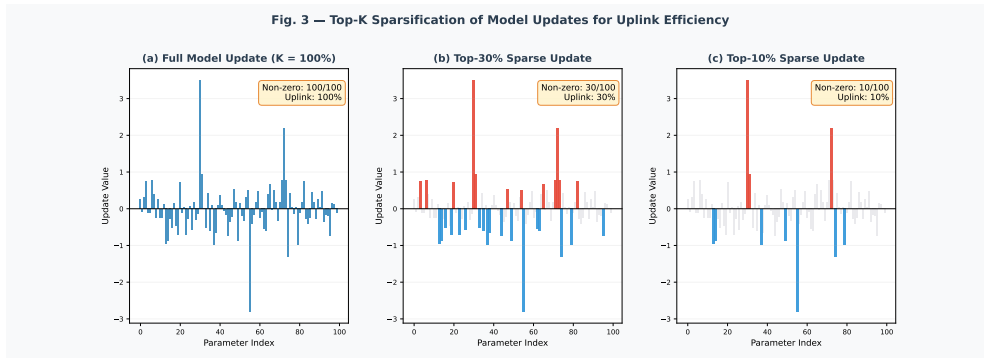


Fig. 3 Top-K sparsification of a model update: (a) $K = 100\%$, (b) $K = 30\%$, (c) $K = 10\%$. Non-zero components are highlighted; uplink cost scales with the retained fraction.

Fig. 4 — Model Convergence (CIFAR-10)

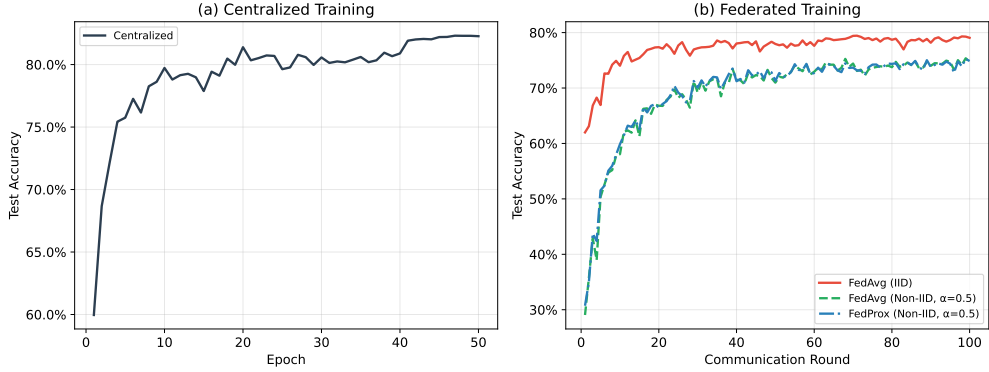


Fig. 4 Model convergence on CIFAR-10: (a) centralized training; (b) FedAvg (IID), FedAvg (non-IID, $\alpha = 0.5$), and FedProx (non-IID, $\alpha = 0.5$).

4 Experiments

4.1 Setup

Datasets. CIFAR-10 [21] (60,000 32×32 RGB images, 10 classes) serves as the primary benchmark; MNIST [22] (70,000 grayscale images) validates the joint experiment on a second vision dataset. **Model.** A lightweight CNN with three conv blocks (Conv–GroupNorm–ReLU–MaxPool, 32/64/128 channels) and two fully-connected layers (2048–256–10); 620,810 parameters on CIFAR-10, representative of an edge-deployable vision model. **FL configuration.** $N = 50$ clients, $M = 10$ selected per round, $E = 5$ local epochs, batch size 32, SGD with learning rate 0.01 and momentum 0.9, $T = 100$ rounds, seed 42. Client data is partitioned with a Dirichlet distribution ($\alpha \in \{0.1, 0.5, 1.0\}$, smaller α = more heterogeneous) for label heterogeneity, and with a long-tail skew (skew = 2.0) for data-quantity heterogeneity. The centralized baseline uses the same model with StepLR decay ($\times 0.1$ at epoch 40). Uplink latency is modeled with a shared 1 Gbps camera uplink. All experiments run on CPU with fixed seeds for reproducibility.

4.2 Convergence Analysis

Fig. 4 shows convergence on CIFAR-10. Centralized training reaches 82.3% after 50 epochs, establishing the attainable upper bound for this model capacity. FedAvg with IID labels converges to 79.1%. Under non-IID conditions ($\alpha = 0.5$) FedAvg degrades to 74.7%, while FedProx ($\mu = 0.01$) recovers to 74.9%—confirming that the proximal term mitigates client drift on camera-scale visual data.

4.3 Communication–Accuracy Trade-off

Fig. 5 and Table 1 report the effect of Top-K sparsification (IID setting). At $K = 10\%$ uplink communication drops by 87% with 77.0% accuracy; the sweet spot $K = 30\%$ retains 78.9% accuracy with 67% communication reduction; full upload achieves 79.1%.

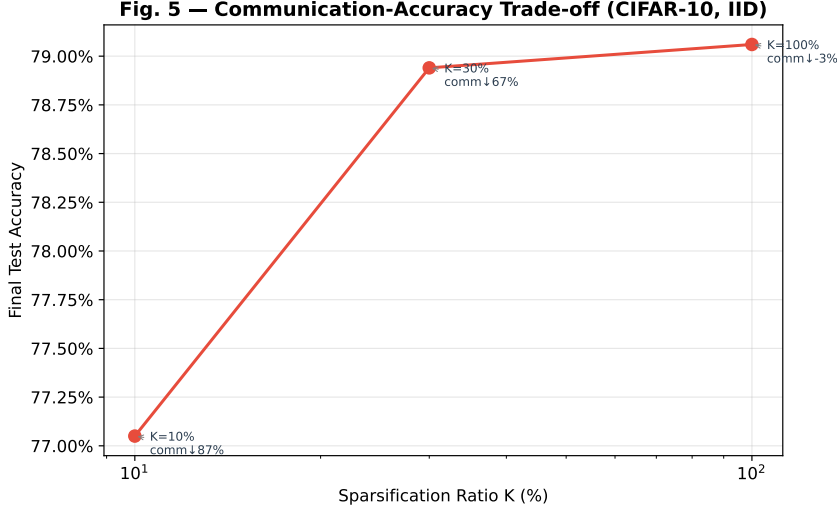


Fig. 5 Communication-accuracy trade-off under Top-K sparsification (CIFAR-10, IID).

Table 1 Sparsification performance on CIFAR-10 (IID, 100 rounds). Compressed settings include the 1-bit-per-parameter mask overhead; the baseline assumes full upload without a mask.

K	Test accuracy	Total comm. (MB)	Comm. reduction	Uplink time (s)
10%	77.05%	325.9	86.9%	2.61
30%	78.94%	822.6	66.9%	6.58
100% (baseline)	79.06%	2483.2	—	19.87

4.4 Non-IID Robustness

Fig. 6 sweeps the Dirichlet concentration $\alpha \in \{0.1, 0.5, 1.0\}$. FedAvg accuracy declines from 76.6% to 59.8% as heterogeneity increases. FedProx recovers 64.3% at $\alpha = 0.1$ (a +4.55% gain over FedAvg) and 74.9% at $\alpha = 0.5$ (+0.18%), with the largest gains under the strongest heterogeneity; near IID ($\alpha = 1.0$) the proximal term is no longer needed and accuracy is comparable (76.0%, -0.58% vs FedAvg), consistent with FedProx’s design for heterogeneous settings.

4.5 Joint Deployment: Sparsification $\times$ Proximal Regularization

Fig. 7 and Table 2 answer the central question of this study: *does Top-K sparsification amplify drift under non-IID data, and does FedProx compensate?* Under strong heterogeneity ($\alpha = 0.1$) on CIFAR-10, FedAvg with $K = 30\%$ sparsification reaches 61.1%; combining sparsification with FedProx raises this to 65.0% (a +3.89% recovery), recovering most of the full-upload FedProx accuracy (64.3%) while retaining the 67% communication saving. Under moderate heterogeneity ($\alpha = 0.5$), where FedProx’s isolated benefit is small, the two techniques compose without interference (74.4% vs

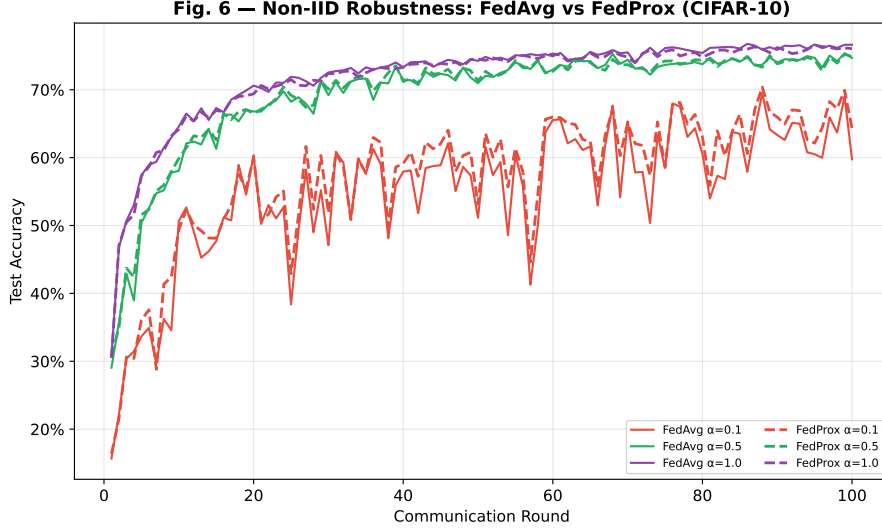


Fig. 6 Non-IID robustness on CIFAR-10: FedAvg (solid) vs FedProx (dashed) across Dirichlet concentration α .

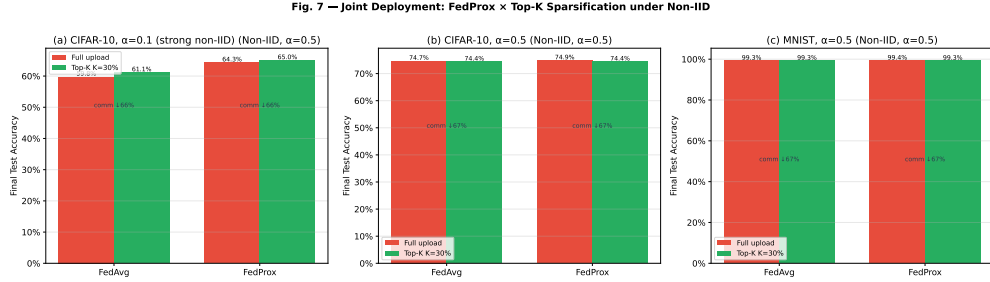


Fig. 7 Joint deployment under non-IID data: FedAvg vs FedProx, each with full upload or $K = 30\%$ sparsification. (a) CIFAR-10, $\alpha = 0.1$; (b) CIFAR-10, $\alpha = 0.5$; (c) MNIST, $\alpha = 0.5$.

74.4%). The same pattern holds on MNIST (99.3% $\rightarrow$ 99.3%), confirming that the two techniques compose across datasets and heterogeneity levels.

4.6 Ablation and the Role of GroupNorm

Table 3 traces the incremental contribution of each component: centralized upper bound (82.3%) $\rightarrow$ FedAvg IID (79.1%) $\rightarrow$ non-IID degradation (-4.37 pp) $\rightarrow$ FedProx recovery ($+0.18$ pp) $\rightarrow$ sparsification at negligible accuracy cost.

We additionally compare BatchNorm and GroupNorm under identical FL conditions: with BatchNorm, FedAvg peaks at 53.1 around round 50 and then *diverges*; with GroupNorm the identical protocol converges monotonically (Fig. 4). This corroborates the analysis of Section 3.4 and motivates GroupNorm as the default for federated vision training with heterogeneous client volumes.

Table 2 Joint deployment matrix under non-IID data (α denotes Dirichlet concentration). “Full” = full model upload; “K=30%” = Top-K sparsification.

Dataset	α	Optimizer	Mode	Accuracy	Comm. (MB)	Reduction
CIFAR-10	0.1	FedAvg	Full	59.75%	2483.2	—
CIFAR-10	0.1	FedAvg	K=30%	61.13%	856.5	65.5%
CIFAR-10	0.1	FedProx	Full	64.30%	2483.2	—
CIFAR-10	0.1	FedProx	K=30%	65.02%	854.9	65.6%
CIFAR-10	0.5	FedAvg	Full	74.69%	2483.2	—
CIFAR-10	0.5	FedAvg	K=30%	74.40%	822.5	66.9%
CIFAR-10	0.5	FedProx	Full	74.87%	2483.2	—
CIFAR-10	0.5	FedProx	K=30%	74.42%	822.5	66.9%
MNIST	0.5	FedAvg	Full	99.34%	1563.4	—
MNIST	0.5	FedAvg	K=30%	99.32%	517.9	66.9%
MNIST	0.5	FedProx	Full	99.41%	1563.4	—
MNIST	0.5	FedProx	K=30%	99.32%	517.9	66.9%

Table 3 Ablation study on CIFAR-10 (non-IID, $\alpha = 0.5$ unless noted).

Method	Accuracy	Note
Centralized (upper bound)	82.27%	50 epochs, StepLR
FedAvg (IID)	79.06%	baseline FL
FedAvg (non-IID, $\alpha = 0.5$)	74.69%	−4.37% due to heterogeneity
+ FedProx ($\mu = 0.01$)	74.87%	+0.18% from proximal term
+ Top-K ($K = 30\%$, non-IID)	74.42%	comm. ↓ 66.9%

4.7 Visual Analysis of the Trained Model

Fig. 8 provides qualitative evidence for the CV community: (a) Top-K masks applied per parameter tensor at a uniform retention rate, visualized for the convolutional layers—retained updates in early layers form structured patterns clustered along specific filter channels, indicating that the dominant gradient energy concentrates in a few feature channels; (b) a PCA projection of the global model’s features on 2,000 test samples shows well-separated class clusters; (c) the confusion matrix highlights the expected CIFAR-10 confusions (e.g., cat/dog), confirming that the federated model learns semantics rather than spurious artifacts.

5 Discussion

Privacy implications. In our framework a camera never uploads raw imagery: the only artifacts exchanged are model updates (compressed by up to 67%). For reference, uploading a single CIFAR-10 image costs 3,072 bytes, while a client’s $K = 30\%$ update costs $\approx 803\text{K}$ bytes per round—the privacy surface is fundamentally different. Gradient leakage attacks [18] remain a residual risk and are a natural direction for future work (e.g., adding differential privacy [19]).

Fig. 8a — Top-K sparse update masks across CNN parameter tensors

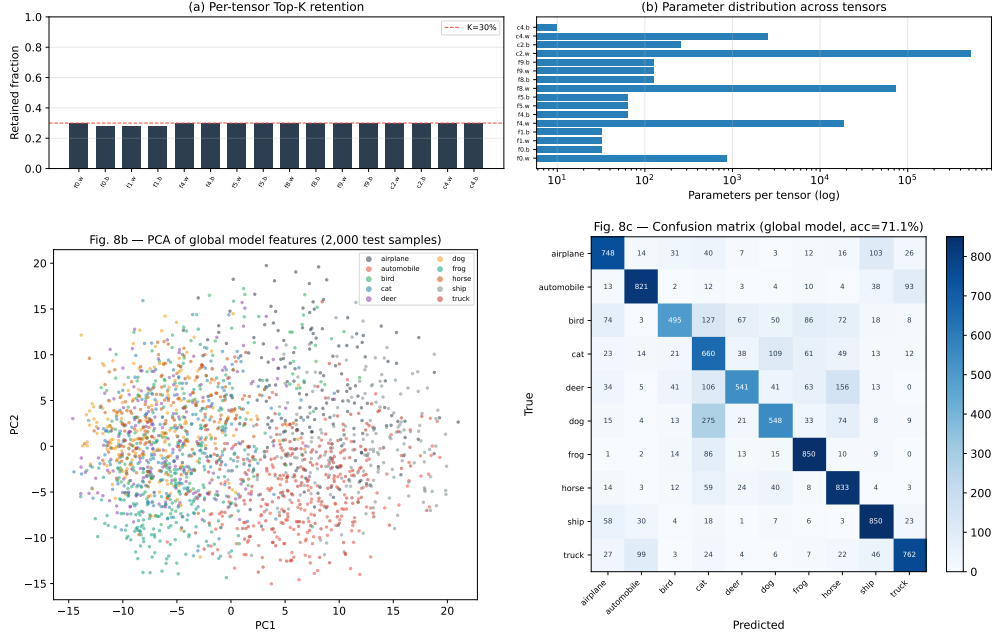


Fig. 8 Visual analysis of the federated model on CIFAR-10: (a) per-tensor Top-K retention and sparse update masks; (b) PCA of global-model features (2,000 test samples); (c) confusion matrix.

Limitations. The study is simulation-based: clients are simulated partitions rather than physical cameras, datasets are small by modern standards, and the uplink model is idealized. Extending to CIFAR-100/ImageNet-scale vision tasks, real camera deployments, and per-device bandwidth constraints are natural next steps.

6 Conclusion

We presented a communication-efficient and robust federated training framework for vision models on IoT camera devices, integrating Top-K gradient sparsification, Fed-Prox proximal regularization, and GroupNorm normalization. On CIFAR-10 with 50 simulated camera clients, the framework retains 78.9% accuracy while cutting uplink communication by 67%, and its joint sparsification+FedProx configuration recovers +3.89% accuracy under strong non-IID data compared with sparsification alone. The results support the practical feasibility of privacy-preserving vision model training at the network edge.

Declarations

Funding. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing Interests. The author declares no competing interests.

Ethics approval and consent to participate. Not applicable.

Consent for publication. Not applicable.

Author Contributions. Chengjun Li conceived the study, designed and implemented the experiments, analyzed the results, and wrote the manuscript.

Data Availability. The CIFAR-10 and MNIST datasets are publicly available. The source code and experimental configurations are available from the corresponding author upon reasonable request.

Code Availability. Reproducible scripts with fixed random seeds (seed 42); code available from the corresponding author upon request.

References

- [1] McMahan B, Moore E, Ramage D, Hampson S, Agüera y Arcas B (2017) Communication-efficient learning of deep networks from decentralized data. In: Proc. AISTATS
- [2] Kairouz P, McMahan HB, Avent B, et al (2021) Advances and open problems in federated learning. Found Trends Mach Learn 14(1–2):1–210
- [3] Konečný J, McMahan HB, Yu FX, Richtárik P, Suresh AT, Bacon D (2016) Federated learning: strategies for improving communication efficiency. arXiv:1610.05492
- [4] Lin Y, Han S, Mao H, Wang Y, Dally WJ (2018) Deep gradient compression: reducing the communication bandwidth for distributed training. In: Proc. ICLR
- [5] Stich SU, Cordonnier JB, Jaggi M (2018) Sparsified SGD with memory. In: Proc. NeurIPS
- [6] Reisizadeh A, Mokhtari A, Hassani H, Jadbabaie A, Pedarsani R (2020) FedPAQ: a communication-efficient federated learning method with periodic averaging and quantization. In: Proc. AISTATS
- [7] Rothchild D, Panda A, Ullah E, et al (2020) FetchSGD: communication-efficient federated learning with sketching. In: Proc. ICML
- [8] Li T, Sahu AK, Zaheer M, Sanjabi M, Talwalkar A, Smith V (2020) Federated optimization in heterogeneous networks. In: Proc. MLSys
- [9] Wang H, Yurochkin M, Sun Y, Papailiopoulos D, Khazaeni Y (2020) Federated learning with matched averaging. In: Proc. ICLR
- [10] Liang PP, Liu T, Ziyin L, et al (2020) Think locally, act globally: federated learning with local and global representations. arXiv:2001.01523
- [11] Liu Y, Huang A, Luo Y, et al (2020) FedVision: an online visual object detection platform powered by federated learning. In: Proc. IAAI

- [12] Hsieh K, Phanishayee A, Mutlu O, Gibbons P (2020) The non-IID data quagmire of decentralized machine learning. In: Proc. ICML
- [13] Li X, Jiang M, Zhang X, Kamp M, Dou Q (2021) FedBN: federated learning on non-IID features via local batch normalization. In: Proc. ICLR
- [14] Wu Y, He K (2018) Group normalization. In: Proc. ECCV
- [15] Nguyen DC, Ding M, Pathirana PN, et al (2021) Federated learning for internet of things: a comprehensive survey. *IEEE Commun Surv Tutor* 23(3):1622–1658
- [16] Bonawitz K, Eichner H, Grieskamp W, et al (2019) Towards federated learning at scale: system design. In: Proc. MLSys
- [17] Yang Q, Liu Y, Chen T, Tong Y (2019) Federated machine learning: concept and applications. *ACM Trans Intell Syst Technol* 10(2):1–19
- [18] Zhu L, Liu Z, Han S (2019) Deep leakage from gradients. In: Proc. NeurIPS
- [19] Wei K, Li J, Ding M, et al (2020) Federated learning with differential privacy: algorithms and performance analysis. *IEEE Trans Inf Forensics Secur* 15:3454–3469
- [20] He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proc. CVPR
- [21] Krizhevsky A (2009) Learning multiple layers of features from tiny images. Tech report, Univ. of Toronto
- [22] LeCun Y, Cortes C, Burges CJC (1998) The MNIST database of handwritten digits. <http://yann.lecun.com/exdb/mnist/>