ENHANCEMENT OF OBJECTIVE FUNCTION IN IMAGE RECOVERY ATTACKS UNDER GRADIENT COMPRESSION CONDITIONS IN FEDERATED LEARNING

Hoang Van Phi, Dao Thi Nga*
Le Quy Don Technical University

ARTICLE INFO

Received: 21/3/2025
Revised: 05/6/2025
Published: 05/6/2025

KEYWORDS

Deep leakage from gradients
Image recovery attack
Distributed machine learning
Data security
Gradient compression

ABSTRACT

Image recovery attacks pose a significant privacy threat in distributed machine learning systems, even when gradient compression is employed. These attacks exploit gradient information to reconstruct original training data, raising serious concerns about data confidentiality. This study presents an improved method based on Deep leakage from gradients to enhance image recovery accuracy under compressed gradient conditions. The proposed method introduces gradient masking to selectively retain significant gradient components and features a key innovation in the integration of Total Variation and L6-norm regularization terms to enhance image smoothness and mitigate artifacts. Experimental evaluations on MNIST and CIFAR-100 datasets reveal that the improved method significantly outperforms traditional Deep Leakage From Gradients and Highly Compressed Gradient Leakage Attack methods, particularly under extreme compression rates. By reducing visual distortions while preserving structural details, the proposed method provides valuable insights for enhancing data security in distributed learning and developing robust defenses against gradient compression attacks.

CẢI TIẾN HÀM MỤC TIÊU TRONG CÁC TẤN CÔNG KHÔI PHỤC ẢNH DƯỚI ĐIỀU KIỆN NÉN GRADIENT TRONG HỌC LIÊN KẾT

Hoàng Văn Phi, Đào Thị Ngà
Trường Đại học Kỹ thuật Lê Quý Đôn

THÔNG TIN BÀI BÁO

Ngày nhận bài: 21/3/2025
Ngày hoàn thiện: 05/6/2025
Ngày đăng: 05/6/2025

TỪ KHÓA

Lộ dữ liệu từ gradients
Tấn công khôi phục ảnh
Học máy phân tán
Bảo mật dữ liệu
Nén gradient

TÓM TẮT

Các cuộc tấn công khôi phục ảnh đặt ra một mối đe dọa nghiêm trọng đối với quyền riêng tư trong các hệ thống học máy phân tán, ngay cả khi sử dụng nén gradient. Những cuộc tấn công này khai thác thông tin gradient để tái tạo dữ liệu huấn luyện ban đầu, gây ra những lo ngại đáng kể về bảo mật dữ liệu. Nghiên cứu này giới thiệu một phương pháp cải tiến dựa trên Deep Leakage From Gradients nhằm nâng cao độ chính xác khôi phục ảnh dưới điều kiện gradient nén. Phương pháp đề xuất áp dụng kỹ thuật mặt nạ gradient để chọn lọc và giữ lại các thành phần gradient quan trọng, đồng thời có một cải tiến chủ chốt trong việc tích hợp các hệ số điều chuẩn tổng biến thiên và L6-norm nhằm cải thiện độ mượt của ảnh và giảm thiểu hiện tượng méo. Các đánh giá thực nghiệm trên bộ dữ liệu MNIST và CIFAR-100 cho thấy phương pháp cải tiến vượt trội so với Deep Leakage From Gradients truyền thống và phương pháp tấn công Highly Compressed Gradient Leakage Attack, đặc biệt ở mức nén cực đoan. Bằng cách giảm thiểu biến dạng hình ảnh trong khi vẫn bảo toàn các chi tiết cấu trúc, phương pháp đề xuất cung cấp những hiểu biết quý giá nhằm nâng cao bảo mật dữ liệu trong học máy phân tán và phát triển các cơ chế phòng thủ mạnh mẽ chống lại các cuộc tấn công dựa trên nén gradient.

DOI: https://doi.org/10.34238/tnu-jst.12360

* Corresponding author. *Email: daothinga@lqdtu.edu.vn*

1. Introduction

The rapid advancement of distributed machine learning systems has significantly enhanced data processing capabilities, yet it has concurrently introduced critical security vulnerabilities [1]. Although these systems exchange gradients rather than raw data to preserve privacy, recent research has demonstrated that gradients can be exploited to reconstruct the original training data [2], [3]. The Deep Leakage from Gradients (DLG) method [4], depicted in Figure 1, facilitates such attacks by optimizing the gradient difference loss $L_{\text{grad\_diff}} = \|g_t^c - g_d\|^2$ to align dummy gradients g_d with shared gradients g_t^c. To reduce communication overhead, gradient compression is widely implemented [5], [6]; however, this practice further complicates attack scenarios by reducing the fidelity of the transmitted gradient information. Traditional DLG methods struggle to maintain image quality under compression, frequently yielding artifacts and distortions. Even the improved Highly Compressed Gradient Leakage Attack (HCGLA) [7], which optimizes $L_{\text{grad\_diff}} = \|g_t^c - g_d^c\|^2$, fails to adequately address the challenges posed by extreme gradient compression percentages.

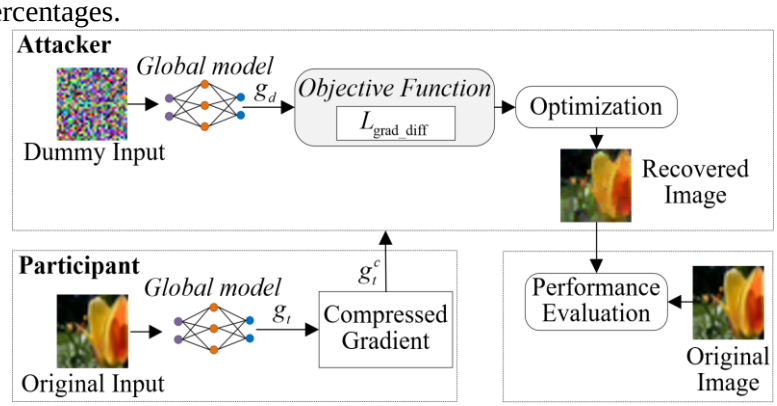

Figure 1. *Framework of DLG and HCGLA*

The deterioration in image recovery quality under gradient compression conditions is attributable to multiple factors, including substantial information loss during compression [6], [8], noise amplification during optimization [9], [10], and convergence issues in recovery algorithms [2], [7]. Furthermore, conventional approaches lack advanced regularization mechanisms, a shortfall that is particularly detrimental when handling complex datasets [4], [9]. To overcome these limitations, the proposed method introduces gradient masking to selectively retain the most informative gradient components [7]. In addition, it incorporates Total Variation (TV) and L6-norm regularization terms to enhance image smoothness and mitigate reconstruction artifacts [11]. These methodological enhancements aim to improve the accuracy of image recovery attacks under gradient compression, thereby providing valuable insights for bolstering data security and optimizing efficiency in distributed machine learning systems.

The remainder of this paper is organized as follows: Section 2 details the proposed methodology, section 3 presents experimental results and evaluation metrics, and section 4 concludes with discussions on implications and directions for future research.

2. Proposed method

To improve the effectiveness of image recovery attacks under conditions of high gradient compression, this study introduces gradient masking in conjunction with a key innovation, the integration of TV and L6-norm regularization terms.

2.1. Gradient masking

In gradient compression scenarios, the quality of gradient information varies significantly across components. To address this, the proposed method incorporates a gradient masking

technique designed to retain only the most informative gradient components while discarding less significant ones. This strategic filtering process enhances image recovery accuracy by preserving key information for optimization.

Gradient significance analysis: A threshold parameter θ is introduced to classify gradient components based on their magnitude. Components satisfying the condition $|q| \geq \theta$ are identified as significant and retained, while those with $|q| < \theta$ are excluded from the reconstruction process. This approach draws inspiration from Lin *et al.* [8], whose work highlighted the efficiency of gradient compression in reducing communication overhead in distributed systems.

Selective loss computation: The loss function is adapted to concentrate solely on regions with significant gradients. The revised masked loss function is defined as follows:

$$L_{\text{grad\_diff}} = \|(g_t^c - g_d^c) \odot \mathbf{M}\|^2 \quad (1)$$

Where g_t^c and g_d^c represent the compressed gradients of the original and dummy images respectively. The binary mask matrix $\mathbf{M}$ assigns values of 1 to significant gradients and 0 otherwise, while $\odot$ denotes the Hadamard (element-wise) product [12], [13].

Targeted parameter optimization: During the optimization phase, updates are applied exclusively to parameters corresponding to significant gradient locations. Parameters associated with negligible gradients remain unchanged, ensuring that computational resources focus only on meaningful data.

By selectively emphasizing critical gradient components, this method achieves multiple benefits. It effectively reduces the search space dimensionality, accelerates convergence, and improves the quality of recovered images under gradient compression. The proposed gradient masking strategy enhances robustness by focusing on gradient regions that encode the most informative features, improving reconstruction fidelity even in highly compressed environments.

2.2. Integration of regularization terms

To enhance the quality of recovered images, the proposed method incorporates TV and L6-norm regularization into the loss function. These regularization techniques are widely recognized in image processing for improving visual clarity and reducing noise artifacts [14], [15].

TV regularization: The TV term is designed to minimize local intensity variations by encouraging consistency between neighboring pixels [14]. This process effectively suppresses noise while preserving sharp edges and structural details. The TV regularization term is defined as follows:

$$L_{TV} = \sum_{i,j} \sqrt{\left(x_{i+1,j} - x_{i,j}\right)^2 + \left(x_{i,j+1} - x_{i,j}\right)^2} \quad (2)$$

Where $x_{i,j}$ represents the pixel value at position (i, j) of the dummy image being reconstructed.

L6-norm regularization: The L6-norm regularization term is designed to penalize extreme pixel values by enforcing a higher-order sparsity constraint [16]. This process stabilizes the optimization and prevents over-amplification of specific pixel intensities, thereby reducing reconstruction artifacts. The L6-norm regularization term is defined as follows:

$$L_{sixnorm} = \sum_i (x_i')^6 \quad (3)$$

where x_i' represents the pixel intensity at position i.

Overall loss function: The final objective function combines these regularization terms with the primary gradient matching loss, ensuring a balance between image fidelity and noise reduction. The complete loss function is expressed as:

$$L_x = L_{grad\_diff} + \alpha_{TV}.L_{TV} + \alpha_{sixnorm}.L_{sixnorm} \quad (4)$$

where α_{TV} and α_{sixnorm} are hyperparameters that control the contribution of each regularization term and are optimally tuned via grid search to balance the trade-off between noise reduction and detail preservation.

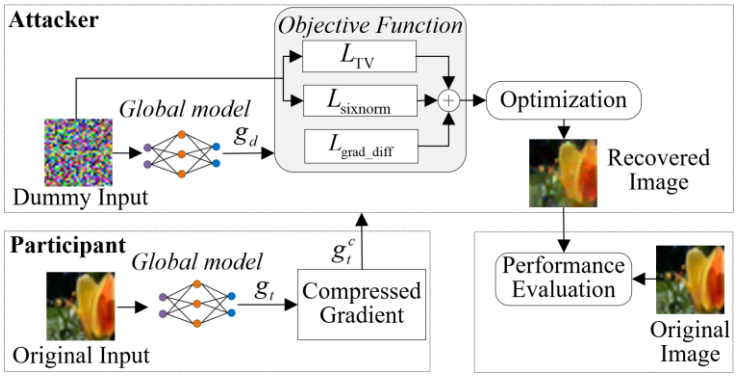

Figure 2. *Framework of proposed method*

As depicted in Figure 2, the proposed method integrates gradient masking with TV and L6-norm regularization techniques. The framework highlights the sequential process of gradient significance analysis, selective loss computation, and targeted parameter optimization, ensuring efficient convergence and improved reconstruction quality under gradient compression conditions.

3. Experiment and evaluation

3.1. Experimental setting

The experimental evaluation was performed on the MNIST and CIFAR-100 datasets using the LeNet architecture, in alignment with previous studies [4]. MNIST was employed as a benchmark for simpler scenarios, whereas CIFAR-100 was utilized to assess performance on more complex, color image data. All experiments were implemented in PyTorch and executed on an NVIDIA Quadro T1000 GPU. The dummy images were initialized with uniform noise, and the LBFGS optimizer was used with a learning rate of 1.0 for 300 iterations per recovery attempt. To simulate realistic distributed learning environments, gradient compression percentages were varied from 80% down to 0.1%. The proposed method was rigorously compared against DLG and HCGLA, with the regularization coefficients optimized via grid search over the range from 10-4 to 10-10 to achieve a balance between gradient matching precision and image smoothness. Performance was quantitatively evaluated using Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index Measure (SSIM), while qualitative analysis focused on visual fidelity and artifact reduction, following a similar evaluation approach outlined in [7]. Specifically, MSE quantifies the average squared difference between the pixels of the original and recovered images, where lower values indicate a closer resemblance. PSNR measures image quality based on the ratio of the original signal to noise, with higher values reflecting better reconstruction quality. SSIM evaluates the structural similarity between the original and recovered images, with values closer to 1 indicating higher structural and visual resemblance.

3.2. Results and Evaluation

3.2.1. Attack results on CIFAR-100

The experimental results on the CIFAR-100 dataset clearly show that the proposed method outperforms existing approaches across various gradient compression percentages, as evidenced by both quantitative metrics and visual assessments. At a 50% compression percentage, it achieves an MSE of 0.0016, which is substantially lower than the MSE values of 0.27471 for DLG and 0.00604 for HCGLA. Furthermore, it attains a high PSNR of 37.82992 dB and an SSIM of 0.98906, as detailed in Table 1. Under more extreme compression conditions, such as a 1% compression percentage, the method records an MSE of 0.00901 and an SSIM of 0.56110. In

stark contrast, DLG fails entirely under such conditions and HCGLA experiences significant degradation, yielding a PSNR of 7.71094 dB and an SSIM of 0.04104. These results underscore the enhanced reconstruction fidelity and robustness of the proposed method in scenarios characterized by severe gradient compression.

Table 1. *Attack results on CIFAR-100 under gradient compression percentages*

Methods	Full gradient			80%			60%			50%		
	MSE (↓)	PSNR (↑)	SSIM (↑)	MSE (↓)	PSNR (↑)	SSIM (↑)	MSE (↓)	PSNR (↑)	SSIM (↑)	MSE (↓)	PSNR (↑)	SSIM (↑)
DLG	$3.46*10^{-6}$	54.61222	0.99976	0.03754	14.25507	0.37049	0.20990	6.77982	0.05665	0.27471	5.61123	0.03210
HCGLA	$3.31*10^{-6}$	54.79621	0.99977	$4.53*10^{-6}$	53.44262	0.99957	0.00022	36.60564	0.98174	0.00604	22.18687	0.72289
Proposed Method	**$7.95*10^{-7}$**	**60.99852**	**0.99994**	**$2.45*10^{-6}$**	**56.11692**	**0.99981**	**$4.15*10^{-5}$**	**43.81782**	**0.99672**	**0.00016**	**37.82992**	**0.98906**

	20%			10%			1%			0.1%		
	MSE (↓)	PSNR (↑)	SSIM (↑)	MSE (↓)	PSNR (↑)	SSIM (↑)	MSE (↓)	PSNR (↑)	SSIM (↑)	MSE (↓)	PSNR (↑)	SSIM (↑)
DLG	0.32830	4.83728	0.01966	0.33922	4.69509	0.01551	0.33558	4.74203	0.01499	0.35443	4.50466	0.00178
HCGLA	0.14136	8.49658	0.11053	0.31548	5.01023	0.01610	0.16940	7.71094	0.04104	0.18172	7.40586	0.01468
Proposed Method	**0.00184**	**27.32841**	**0.88635**	**0.00357**	**24.47342**	**0.79773**	**0.00901**	**20.45186**	**0.56110**	**0.02539**	**15.95388**	**0.21914**

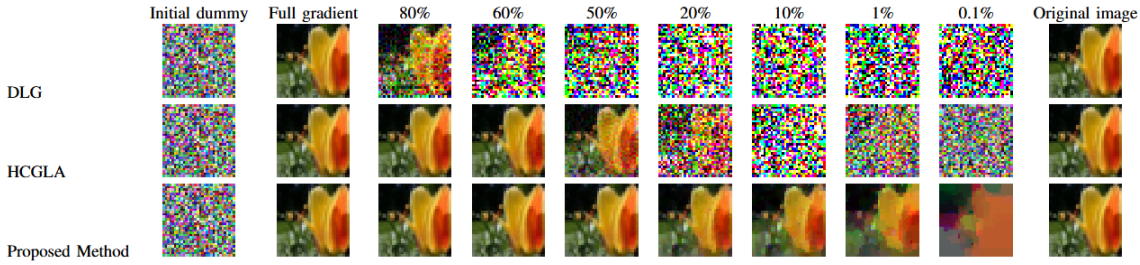

Figure 3. *Recovered CIFAR-100 images under gradient compression percentages*

Visual analysis in Figure 3 shows that the proposed method effectively preserves structural details and maintains accurate color distributions even at compression levels as high as 80%. In contrast, the DLG approach produces largely unrecognizable images due to significant noise amplification. Although HCGLA exhibits a slight improvement over DLG, it still suffers from significant artifacts when compression percentage exceeds 50%. Additionally, the convergence curve presented in Figure 4 underscores the stability of the proposed method, achieving near-optimal MSE and SSIM values within 150 iterations. Notably, the final MSE of the proposed method is 0.00016, substantially lower than HCGLA's 0.00843, emphasizing its superior overall reconstruction performance.

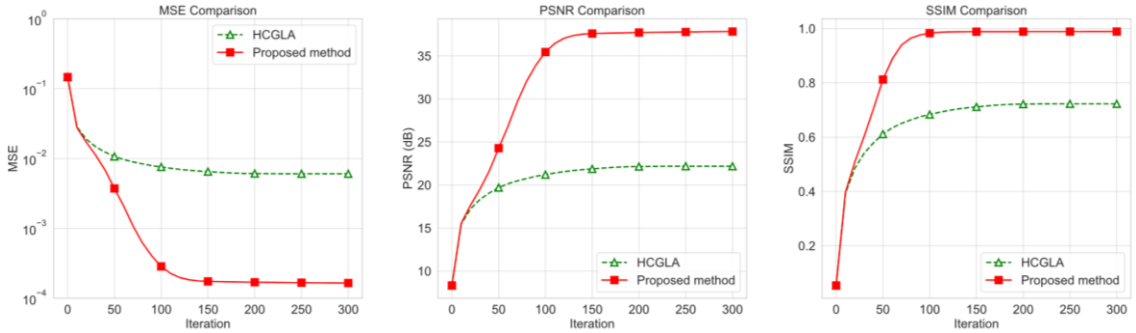

Figure 4. *Performance evaluation at 50% compression on CIFAR-100*

3.2.2. Attack results on MNIST

The proposed method exhibits remarkable robustness on the MNIST under diverse gradient compression percentages, outperforming DLG and HCGLA in both quantitative metrics and visual recovery quality. At 50% compression, as shown in Table 2, the method achieves near-perfect reconstruction with an MSE of $3.06*10^{-10}$, PSNR of 95.13612 dB, and SSIM of 1.0, surpassing DLG, which has an MSE of 0.024 and an SSIM of 0.37429, as well as HCGLA, which has an MSE of $3.74*10^{-8}$ and an SSIM of 0.99997. Notably, even under extreme

compression at 0.1%, the method maintains a high SSIM of 0.07065, compared to HCGLA's 0.00660 and DLG's failure, demonstrating its ability to preserve structural integrity in highly constrained settings.

Table 2. *Attack results on MNIST under gradient compression percentages*

Methods	Full gradient			80%			60%			50%		
	MSE (↓)	PSNR (↑)	SSIM (↑)	MSE (↓)	PSNR (↑)	SSIM (↑)	MSE (↓)	PSNR (↑)	SSIM (↑)	MSE (↓)	PSNR (↑)	SSIM (↑)
DLG	$2.76*10^{-8}$	75.59343	0.99994	0.00076	31.18187	0.67541	0.00716	21.45221	0.46109	0.02414	16.17204	0.37429
HCGLA	$1.74*10^{-8}$	77.58341	0.99998	$1.87*10^{-8}$	77.28295	0.99998	$2.11*10^{-8}$	76.74818	0.99999	$3.74*10^{-8}$	74.27115	0.99997
Proposed Method	**$9.73*10^{-11}$**	**100.1185**	**1.0**	**$1.54*10^{-10}$**	**98.13414**	**1.0**	**$2.12*10^{-10}$**	**96.73004**	**1.0**	**$3.06*10^{-10}$**	**95.13612**	**1.0**

	20%			10%			1%			0.1%		
	MSE (↓)	PSNR (↑)	SSIM (↑)	MSE (↓)	PSNR (↑)	SSIM (↑)	MSE (↓)	PSNR (↑)	SSIM (↑)	MSE (↓)	PSNR (↑)	SSIM (↑)
DLG	0.301445	5.20792	0.08600	0.46157	3.35763	0.04038	0.48339	3.15697	0.00409	0.51480	2.88359	0.00706
HCGLA	$8.48*10^{-7}$	60.71647	0.99936	0.00017	37.74665	0.94132	0.28272	5.48636	0.06752	0.30845	5.10817	0.00660
Proposed Method	**$1.35*10^{-9}$**	**88.68839**	**1.0**	**$6.46*10^{-9}$**	**81.89783**	**1.0**	**0.01024**	**19.89740**	**0.60730**	**0.07903**	**11.02196**	**0.07065**

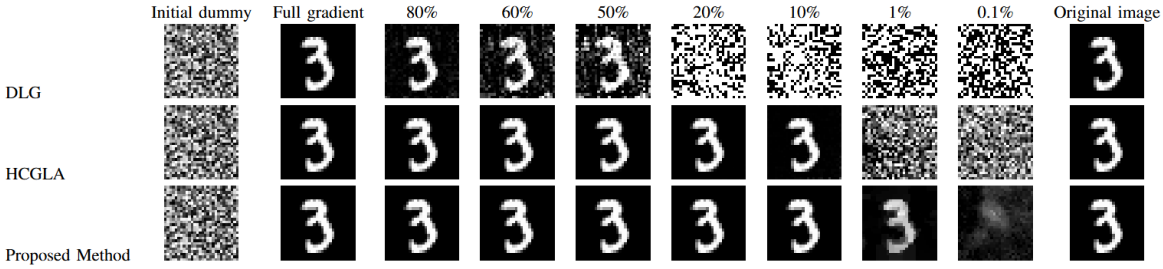

Figure 5. *Recovered MNIST images under gradient compression percentages*

Visual comparisons in Figure 5 underscore the method's superiority, as the recovered MNIST digits retain sharp edges and clear, recognizable shapes even at compression levels as high as 10% and 1%. In contrast, both DLG and HCGLA yield outputs that are either distorted or entirely unrecognizable. Moreover, the convergence analysis depicted in Figure 6 further highlights the stability of the proposed method, achieving an MSE of $6.46*10^{-9}$ and an SSIM of 1.0 at a 10% compression level after 300 iterations. This performance is significantly better than that of HCGLA, which reports an MSE of 0.00017 and an SSIM of 0.94132.

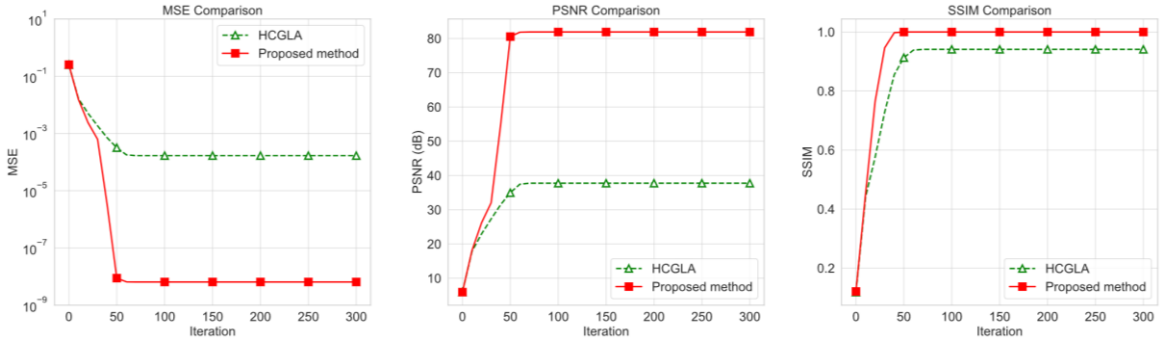

Figure 6. *Performance evaluation at 10% compression on MNIST*

3.3. Performance comparison

The proposed method consistently outperforms DLG and HCGLA across all compression percentages, achieving 2 to 3 orders of magnitude lower MSE and 10 to 30 dB higher PSNR. This advantage is particularly evident at a 0.1% compression percentage, where SSIM reaches 0.22 on CIFAR-100 and 0.07 on MNIST. Such robustness arises from the integration of TV-L6 regularization, which enhances structural detail preservation by balancing pixel smoothness and intensity distribution. While DLG fails beyond 10% compression and HCGLA deteriorates significantly, the proposed method effectively reconstructs key features like edges and color patterns with minimal artifacts, improving both visual fidelity and data protection in privacy-critical environments.

4. Conclusion

This study introduces an enhanced image recovery attack method that effectively addresses critical limitations inherent in existing techniques under gradient compression conditions. By integrating gradient masking with TV-L6 regularization, the proposed approach achieves robust reconstruction accuracy even at extreme compression levels and significantly outperforms both DLG and HCGLA. These findings underscore the urgent need for more advanced privacy protection mechanisms in federated learning systems. Future research should explore adaptive regularization strategies and adversarial training techniques to further mitigate reconstruction risks, particularly in heterogeneous or non-IID data environments. This work not only advances our understanding of gradient-based privacy vulnerabilities but also lays a solid foundation for the development of stronger defense mechanisms, ultimately promoting safer and more efficient distributed learning frameworks.

REFERENCES

[1] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *International Conference on Artificial Intelligence and Statistics*, 2016, pp. 1273-1282.

[2] J. Geiping, H. Bauermeister, H. Dröge, and M. Moeller, "Inverting Gradients - How easy is it to break privacy in federated learning?," *Advances In Neural Information Processing Systems*, vol. 33, pp. 16937-16947, 2020.

[3] B. Zhao, K. R. Mopuri, and H. Bilen, "iDLG: Improved Deep Leakage from Gradients," *arXiv preprint arXiv:.02610*, 2020.

[4] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems: Curran Associates Inc.*, 2019, Art. no. 1323.

[5] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, "QSGD: Communication-Efficient SGD via Gradient Quantization and Encoding," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 1707-1718.

[6] A. F. Aji and K. Heafield, "Sparse Communication for Distributed Gradient Descent," *ArXiv*, vol. abs/1704.05021, 2017.

[7] H. Yang, M. Ge, K. Xiang, and J. Li, "Using Highly Compressed Gradients in Federated Learning for Data Reconstruction Attacks," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 818-830, 2023.

[8] Y. Lin, S. Han, H. Mao, Y. Wang, and W. J. Dally, "Deep Gradient Compression: Reducing the Communication Bandwidth for Distributed Training," *ArXiv*, vol. abs/1712.01887, 2017.

[9] W. Wei *et al.*, "A Framework for Evaluating Gradient Leakage Attacks in Federated Learning," *ArXiv*, vol. abs/2004.10397, 2020.

[10] J. Jeon, K. Lee, S. Oh, and J. Ok, "Gradient inversion with generative image prior," *Advances in neural information processing systems*, vol. 34, pp. 29898-29908, 2021.

[11] A. Mahendran and A. Vedaldi, "Understanding Deep Image Representations by Inverting Them," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5188-5196.

[12] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436-444, 2015.

[13] X. Zhang, *Matrix Analysis and Applications*. Cambridge University Press, 2017.

[14] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, pp. 259-268, 1992.

[15] X. Zhang, M. Burger, and S. Osher, "A Unified Primal-Dual Algorithm Framework Based on Bregman Iteration," *Journal of Scientific Computing*, vol. 46, pp. 20-46, 2010.

[16] S. P. Boyd and L. Vandenberghe, "Convex Optimization," *IEEE Transactions on Automatic Control*, vol. 51, pp. 1859-1859, 2010.