선택적 암호화를 통한 프라이버시 보호 연합 학습

이영한 [1]
[1] 성신여자대학교 융합보안공학과
yhlee@sungshing.ac.kr

Privacy-preserving Federated Learning via Selective Encryption

Younghan Lee[1]
[1]Dept. of Convergence Security Engineering, Sungshin Women's University

Abstract

We introduce a novel method to defend against model inversion attacks in Federated Learning (FL). FL enables the training of a global model by sharing local gradients without sharing clients' private data. However, model inversion attacks can reconstruct the data from the shared gradients. Traditional defense mechanisms, such as Differential Privacy (DP) and Homomorphic Encryption (HE), have limitations in balancing privacy and model accuracy. Our approach selectively encrypts more important gradients, which contain more information about the training data, to balance between privacy and computational efficiency. Additionally, optional DP noise is applied to unencrypted gradients for enhanced security. Comprehensive evaluations demonstrate that our method significantly improves both privacy and model accuracy compared to existing defenses.

1. Introduction

In this paper, we present a novel defensive mechanism designed to protect the training data used to train the global model in federated learning (FL) against model inversion attacks. FL has been increasingly popular over the past years due to its capability of training the global model by sharing local model updates (gradients) without revealing the clients' private dataset which is considered valuable. However, model inversion attacks introduce a significant privacy threat by attempting to reconstruct the private dataset belonging to the client. The attack utilizes the gradients shared with the server to extract the privacy. To mitigate this, various defense mechanisms have been proposed such as differential privacy (DP) and homomorphic encryption (HE). While DP enhances the data privacy in FL setting, it comes with the cost of reduced model performance as the noise is added to the gradient. Recent efforts have focused on optimizing noise levels to balance privacy with model accuracy, though the increasing complexity of model architectures presents limitations for this strategy.
Our proposed method overcome these challenges by leveraging HE to secure shared gradients and maintain the performance of the global model. Instead of applying HE to all gradients, which would be inefficient and overprotection, we selectively encrypt gradients with higher values. Such gradients carry more information about the private data. Moreover, we incorporate DP strategy to enhance the privacy by introducing noise to the unencrypted gradients. Through extensive evaluation compared to state-of-the-art (SOTA) defense mechanisms, our method demonstrates significant improvements in both accuracy and privacy.

2. Background

Homomorphic encryption (HE) [1]: a cryptographic technique that enables computations to be conducted directly on encrypted data, without requiring decryption. It ensures that sensitive information remains confidential throughout the computational process. HE is broadly classified into two categories based on the types of mathematical operations it supports on ciphertext: Partially Homomorphic Encryption (PHE) and Fully Homomorphic Encryption (FHE). PHE allows for only a single operation, either addition or multiplication, on the encrypted data. In contrast, FHE supports both addition and multiplication, facilitating arbitrary computations on encrypted information. This feature of HE makes it particularly useful in cloud-based applications, such as privacy-preserving FL, where encrypted data can be aggregated by the server without revealing the underlying information.

Differential Privacy (DP) [2]: Let ϵ be a positive real number and $0 \leq \delta < 1$. A randomized function M gives (ϵ, δ)-differential privacy if for all datasets D1 and D2 differing on at most one element, and for all $O \subseteq \text{Range}(M)$, it satisfies $\Pr[M(D1) \in O] \leq \exp(\epsilon) \times \Pr[M(D2) \in O] + \delta$.
In definition, we can easily know that ϵ is upper bound on the privacy loss corresponding to mechanism M. When δ is 0, the above definition is called $(\epsilon, 0)$-DP and simply called ϵ-DP, it is very difficult to satisfy the above definition, and the definition is mitigated by adding δ, a number between 0 and 1. In ϵ-DP, the degree of privacy protection is determined by the value of the constant ϵ, and ϵ is called a privacy budget. In other words, if ϵ is 0, perfect privacy can be preserved, and as ϵ increases, privacy is lost.

Model Inversion Attacks: Recent research has identified a new type of attack in federated learning (FL) that poses a threat to client privacy by reconstructing their private datasets. In FL, clients send their local updates (gradients) to the server, which aggregates them to train a global model. As the server has access to all shared gradients while clients await the global model, an honest-but-curious server could potentially reconstruct private data samples through a process known as gradient matching. The effectiveness of the reconstructed data, such as images, is typically evaluated by comparing their similarity to the original samples.

Inverting Gradients (IG) [3] shows that previous attacks only considered the case of a single image in a batch. However, it is possible to reconstruct even with multiple images at high resolution for trained deep networks. They design a new loss function for gradient matching strategy and prove that the data to any fully connected layer can be reconstructed analytically.

Privacy-preserving Federated Learning:
To mitigate the privacy attacks such as model inversion attack, a sophisticated defense mechanism has been developed. This mechanism primarily involves adding noise to the gradients shared with the central server used to train the global model. The key challenge is to implement this defense without significantly compromising the accuracy of the global model. Additionally, it is crucial for this defense strategy to be practical, meaning it should not impose excessive computational overhead. Balancing between these requirements is vital for the effectiveness and feasibility of the defense mechanism.
Soteria [4] emphasizes that privacy leakage in FL largely originates from the data representations embedded within the model updates. To mitigate this issue, the authors propose a perturbation-based defense mechanism that combines noise addition with gradient clipping in a single layer, aiming to minimize privacy leakage while preserving the global model's accuracy. Additionally, the paper offers an analysis of certified robustness and convergence guarantees.
Outpost [5] introduces an innovative defense mechanism for FL, designed to address different levels of privacy risks throughout the FL process. This approach intelligently injects Gaussian noise into the gradients at each iteration, with the noise intensity dynamically adjusted according to the assessed privacy risk. The risk is determined by analyzing the dispersion of model weights in each layer using the Fisher information matrix. Additionally, to reduce computational overhead, Outpost incorporates a decay in the perturbation as the number of iterations increases.

3. Methodology

The process consists of mainly two phases and two steps in each phase.
In the first phase, we selectively encrypt gradients based on their informational significance. We observe that larger gradients play a more crucial role in model updates, and therefore, encryption is applied to both shallow and deep layers where necessary. The optimal percentage of encrypted gradients is identified through iterative experimentation, aiming to balance between defense effectiveness and computational efficiency. This approach is driven by the understanding that gradients with larger magnitudes carry more information about the training data. Hence, we prioritized encryption for those gradients for the better privacy protection.
In the second phase, we further enhance privacy by adding noise to the gradients that are not encrypted.
In this phase, we aim to improve the overall privacy of FL framework by selecting an appropriate noise level represented by delta. Once again, the optimal value of delta is found through iterative experiments. This is important as the higher value of delta can offer better privacy protection, but the accuracy of the global model will decrease. We need to balance the tradeoff between privacy and accuracy. Although this phase is optional and is especially beneficial when minimizing computational overhead is priority, as in cases with fewer encrypted gradients.

4. Experiments

In this section, we evaluate the previously discussed methods, with a particular focus on its first phase. This phase concentrates on the selective application of homomorphic encryption (HE) to gradients, which are prioritized based on their magnitudes. The goal of this approach is to optimize the trade-off between privacy protection and computational efficiency within the encryption and federated learning process.
We assume that the dataset is distributed evenly (IID) across all participating clients. Additionally, it is assumed that the server can receive gradients from the clients and accurately determine their origins. We use CIFAR10 dataset with 500 FL rounds with total of 100 clients. We allow 20 rounds per client with 0.01 learning rate, 0.9 momentum and 0.0005 weight decay. For Soteria, we configure the noise addition to affect 50% of the gradients in a fully connected layer. In Outpost, 80% of the gradients are pruned to zero, and noise is applied to 40% of the remaining gradients. For our method, we encrypt 10% of the largest gradients. The ResNet-50 architecture was employed to train the global model.

We evaluate the effectiveness of our method against model inversion attacks by measuring the similarity of reconstructed images. By utilizing IG attack method, which is described in Background section, we compare the quality of the reconstructed image under various defense mechanisms. The result is shown in Figure 1. The result shows that the image reconstructed under our proposed defense method is blurrier and less similar to the original image compared to other defenses. This indicates our method is most effective in protecting the privacy.
For quantitative analysis, we employ Mean Square Error (MSE) to measure pixel-wise differences and Peak Signal-to-Noise Ratio (PSNR), which calculates the ratio of maximum pixel fluctuation to the MSE between the original and reconstructed images. Our method demonstrates strong defense performance, reflected in higher MSE and lower PSNR values, as shown in Table 2.

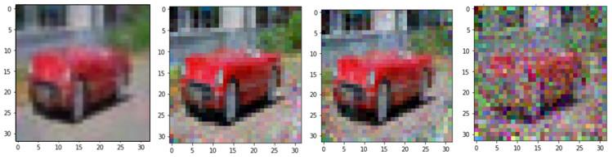

Figure 1. The visual result of reconstructed images using IG attack and the original image.

	Soteria	Outpost	Ours
MSE	0.0368	0.0709	0.4445
PSNR	26.36	23.52	15.54

Table 2. Comparison with state-of-the-art defense mechanisms. Higher MSE and lower PSNR is better

5. Discussion

While the results from our evaluation show that our method outperforms existing defense mechanisms such as Soteria and Outpost in terms of both privacy and accuracy, our experiments were conducted on a limited set of scenarios and models, notably using the ResNet-50 architecture. To ensure the generalizability and robustness of our approach, future work should expand the experiments to include a wider variety of model architectures, datasets, and real-world scenarios. Investigating how our method performs under different types of model inversion attacks, would provide additional insights into its broader applicability. Furthermore, additional research should explore optimizing the selection of gradients for encryption and the balance of DP noise to fine-tune performance.

6. Conclusion

In this paper, we introduced a novel defense mechanism for federated learning designed to protect against model inversion attacks by selectively encrypting the most significant gradients. Our approach reduces the computational overhead associated with homomorphic encryption by focusing on gradients with the highest informational value and, optionally, adding DP noise to the unencrypted gradients to enhance privacy. Through evaluations, our method demonstrated superior performance compared to existing methods, offering a promising solution to the challenge of maintaining both privacy and model accuracy in FL.

References

[1] Gentry, C. "Fully Homomorphic Encryption Using Ideal Lattices." *Proceedings of the 41st Annual ACM Symposium on Theory of Computing*, ACM, 2009, pp. 169–178.

[2] Dwork, C., Roth, A., et al. "The Algorithmic Foundations of Differential Privacy." *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, 2014, pp. 211–407.

[3] Geiping, J., Bauermeister, H., Dröge, H., and Moeller, M. "Inverting Gradients: How Easy Is It to Break Privacy in Federated Learning?" *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 16937–16947.

[4] Sun, J., Li, A., Wang, B., Yang, H., Li, H., and Chen, Y. "Soteria: Provable Defense Against Privacy Leakage in Federated Learning from Representation Perspective." *IEEE Conference on Computer Vision and Patte rn Recognition* (CVPR).

[5] Wang, F., Hugh, E., & Li, B. (2024). More than enough is too much: Adaptive defenses against gradient leakage in production federated learning. *IEEE/ACM Transactions on Networking*.