MGIA: Mutual Gradient Inversion Attack In Multi-Modal Federated Learning (Student Abstract)

Xuan Liu[1], **Siqi Cai**[2], **Lin Li**[2], **Rui Zhang**[1], **Song Guo**[1]

[1]The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong
[2]School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan 430070
xuan18.liu@connect.polyu.hk, {csiqi, cathylilin}@whut.edu.cn, csrzhang1@comp.polyu.edu.hk, song.guo@polyu.edu.hk

Abstract

Recent studies have demonstrated that local training data in Federated Learning can be recovered from gradients, which are called *gradient inversion* attacks. These attacks display powerful effects on either computer vision or natural language processing tasks. As it is known that there are certain correlations between multi-modality data, we argue that the threat of such attacks combined with Multi-modal Learning may cause more severe effects. Different modalities may communicate through gradients to provide richer information for the attackers, thus improving the strength and efficiency of the gradient inversion attacks. In this paper, we propose the **M**utual **G**radient **I**nversion **A**ttack (MGIA), by utilizing the shared labels between image and text modalities combined with the idea of knowledge distillation. Our experimental results show that MGIA achieves the best quality of both modality data and label recoveries in comparison with other methods. In the meanwhile, MGIA verifies that multi-modality gradient inversion attacks are more likely to disclose private information than the existing single-modality attacks.

Introduction

Federated Learning is a newly proposed privacy-preserving paradigm, which collaboratively trains a global model by exchanging gradients between the parameter server and the clients. Recent studies have shown that private training data can be recovered from gradients called *gradient inversion* attacks (Zhang et al. 2022). These attacks can effectively reconstruct the images in computer vision tasks or the texts in natural language processing tasks (Zhu, Liu, and Han 2019). However, the threat of such attacks combined with Multi-modal Learning is not yet studied, and more serious consequences may arise. In some existing multi-modality federated learning frames for instance HGB (Chen and Li 2022), gradients of different modalities training models are shared separately, which provides fruitful information for gradient inversion attacks. Analysis from the attacker's perspective, during the attack, the recovery of one data attribute can greatly facilitate the inference of other data, and different modalities may communicate through gradients to achieve certain complementarity. To study these problems, we propose the **M**utual **G**radient **I**nversion **A**ttack (MGIA) by exploiting the correlated labels between image and text modalities and the idea of knowledge distillation. The key insight is that the *Teacher* model helps the *Student* model by providing its recovered label distribution, and in turn, the *Student* model feeds its recovered label information back to the *Teacher*. The framework of our MGIA is illustrated in Fig. 1. In our experiments, compared with the representative gradient inversion attacks, MGIA improves the recovery quality for more than 39%, and has much better performance in label recovery when dealing with text and image data together.

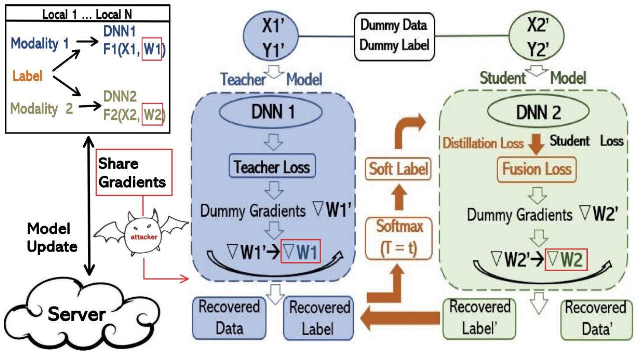

Figure 1: The framework of MGIA. The Teacher model provides extracted labels for the recovery of the Student model.

Methodology

The MGIA algorithm is shown in Algorithm 1. Inspired by knowledge distillation, the modality attack model $F_1(x_1; W_1)$ with better recovery accuracy is set as the *Teacher*, and its recovered soft label guides the *Student* modality attack model $F_2(x_2; W_2)$. We first extract the shared model parameters W_1 and W_2 of two modalities separately and generate the dummy data x'_1, x'_2 and dummy labels y'_1, y'_2 of the above attack models respectively (Line 1 in Algorithm 1).

For the *Teacher* model, we obtain the dummy gradients $\nabla W'_1$ by feeding the corresponding dummy data and label, and iteratively update x'_1 and y'_1 until the distance between $\nabla W'_1$ and ∇W_1 is close enough (i.e., optimization conver-

Algorithm 1: MGIA: Mutual Gradient Inversion Attack

```
1: $x_1', x_2', y_1', y_2' \leftarrow \mathcal{N}(0,1)$
2: **for** $iteration = 1$ to $n$ **do**
3:     $\nabla W_1' = \partial l(F_1(x_1'; W_1), y_1') / \partial W_1$
4:     $D_1 = ||\nabla W_1' - \nabla W_1||^2$
5:     $x_1' \leftarrow x_1' - \eta \nabla_{x_1'} D_1, y_1' \leftarrow y_1' - \eta \nabla_{y_1'} D_1$
6:     $\hat{x}_1 \leftarrow x_1', \hat{y}_1 \leftarrow y_1'$
7: **for** $iteration = 1$ to $n$ **do**
8:     $l_{stu} = l(F_2(x_2'; W_2), softmax(y_2'))$
9:     $l_{dist} = l(F_2(x_2'; W_2)/\tau, softmax(y_1'/\tau))$
10:    $l_{fusion} = \alpha_1 * l_{stu} + (1 - \alpha_1) * l_{dist}$
11:    $\nabla W_2' = \partial l_{fusion} / \partial W_2$
12:    $D_2 = ||\nabla W_2' - \nabla W_2||^2$
13:    $x_2' \leftarrow x_2' - \eta \nabla_{x_2'} D_2, y_2' \leftarrow y_2' - \eta \nabla_{y_2'} D_2$
14:    $\hat{x}_2 \leftarrow x_2', \hat{y}_2 \leftarrow y_2'$
15: $\hat{y} = \alpha_2 * \hat{y}_1 + (1 - \alpha_2) * \hat{y}_2$
16: **return** $\hat{x}_1, \hat{x}_2, \hat{y}$
```

gence). Then, we can approximately recover the input data $\hat{x}_1$ and its label $\hat{y}_1$ (Lines 2-6). For the *Student* model, we use the derived y_1 as the soft label to guide the optimization of dummy gradient $\nabla W_2'$. According to the fused loss function l_{fusion}, we can recover the data $\hat{x}_2$ and label $\hat{y}_2$ in a similar manner (Lines 7-14). Finally, the *Student* model feeds its recovered labels back to the *Teacher* model and MGIA returns the final labels by distributing different weights to the recovered labels from different modalities (Line 15).

Experiments

Our experiments are conducted on popular image datasets (*CIFAR-100*, *STL-10*, *FashionMNIST (FMNIST)*, *Flower Images (FI)*), and we add text descriptions for each image to construct multi-modality data. In the construction of the experimental environment, we choose the multi-modality prediction task as the target model of our attack. Three metrics, i.e., Average Peak Signal to Noise Ratio (*APSNR*), text Recover Rate (*RR*), and label Recover Accuracy (*RA*) are adopted to evaluate the reconstruction of images, texts, and labels respectively. To compare with single-modality gradient attacks, *DLG* (Zhu, Liu, and Han 2019), *Inverting Gradients (IG)* (Geiping et al. 2020), and *GRNN* (Ren, Deng, and Xie 2022) are our baselines. We present two MGIA variant models. ***MGIA (Splicing)*** is the splicing model without any multi-modality interaction, which is used to demonstrate the possibility of recovering multi-modality through single-modality gradient attacks. ***MGIA (No KD)*** is the weighted average model without knowledge distillation, which is used to investigate the effect of knowledge distillation.

As shown in Table 1 and Fig. 2, when applying MGIA, there are improvements in the accuracy of both image data and labels that are previously poorly recovered using single-modality gradient attack methods. Meanwhile, compared to other methods, MGIA can recover images with larger resolution and has a faster convergence rate which demonstrated the positive effect of multi-modality interactions on gradient attacks, as well as MGIA's superiority in the efficiency of data and label recovery.

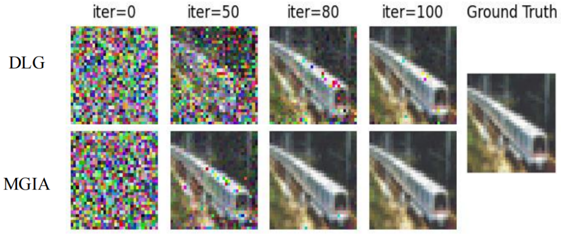

Figure 2: The comparison of convergence rate

Datasets	Methods	APSNR	RR	RA
CIFAR-100	DLG	27.3	-	-
	IG	23.8	-	-
	GRNN	39.1	-	-
	MGIA(Splicing)	9.38	0.92	-
	MGIA(No KD)	28.0	0.94	0.82
	MGIA(Ours)	**43.7**	**0.96**	**0.94**
STL-10	DLG	27.2	-	-
	IG	26.6	-	-
	GRNN	34.6	-	-
	MGIA(Splicing)	10.9	0.82	-
	MGIA(No KD)	24.6	0.98	0.96
	MGIA(Ours)	**39.9**	**1.0**	**1.0**

Table 1: The comparison of Average Peak Signal to Noise Ratio (image), Recover Rate (text), and Recover Accuracy (label) [MGIA (Splicing): resolution: 25×25 px; iterations: 150. Other models: resolution: 32×32 px; iterations: 100.]

Conclusion

In this paper, we explore the effects of multi-modality on gradient attacks and propose a Mutual Gradient Inversion Attack (MGIA). The attack makes use of the correlation between different modalities for accurate data recovery. The extensive evaluation results show that MGIA can effectively and efficiently reconstruct the private training data.

References

Chen, S.; and Li, B. 2022. Towards Optimal Multi-Modal Federated Learning on Non-IID Data with Hierarchical Gradient Blending. In *IEEE INFOCOM*, 1469–1478.

Geiping, J.; Bauermeister, H.; Dröge, H.; and Moeller, M. 2020. Inverting gradients-how easy is it to break privacy in federated learning? In *NeurIPS*, 16937–16947.

Ren, H.; Deng, J.; and Xie, X. 2022. GRNN: Generative Regression Neural Network—A Data Leakage Attack for Federated Learning. *ACM Transactions on Intelligent Systems and Technology*, 13(4): 1–24.

Zhang, R.; Guo, S.; Wang, J.; Xie, X.; and Tao, D. 2022. A Survey on Gradient Inversion: Attacks, Defenses and Future Directions. In *IJCAI*, 5678–5685.

Zhu, L.; Liu, Z.; and Han, S. 2019. Deep Leakage from Gradients. In *NeurIPS*, 14747–14756.