Triggers Magic Mirror: A Trigger Reconstruction Method for Backdoor Attacks in Federated Learning

Anonymous submission

Appendix A

Why traditional gradient inversion algorithms cannot be directly applied to backdoor reconstruction?

Traditional gradient inversion studies (e.g., DLG(Zhu, Liu, and Han 2019), iDLG(Zhao, Mopuri, and Bilen 2020)) tend to focus on the gradient inversion effect of a single image. In contrast, trigger reconstruction is influenced by the model training of federated learning. As normal data from clients is continuously trained, the model parameter vectors evolve in a consistent direction. The magnitude of gradient changes (parameter magnitude) gradually decreases, and the extent of parameter updates becomes smaller (As shown in Figure 1). This is why the difference between the target gradient and the generated image gradient increases over time during the gradient inversion process. This results in larger loss values when comparing the target gradient with the generated gradient. Therefore, The traditional gradient descent algorithm can only be successfully inverted when the difference between the target gradient and the generated gradient is not large, meaning that conducting gradient inversion at the early stage of backdoor injection yields the best results.

In contrast to a controlled environment, the majority of sophisticated backdoor attack methods (Zhang et al. 2022; Lyu et al. 2023) in federated learning do not poison the model at the outset of training. Instead, they are deployed during the ongoing training of the target model. Figure 2 clearly shows that poisoning during the early stages of model training allows us to reconstruct clear triggers. However, after reaching 110 iterations, the gradient inversion algorithm struggles to achieve backdoor reconstruction. It is evident that the L_1 and L_2 norms of the backdoor samples remain relatively unchanged throughout the initial stages of poisoning, despite the target model undergoing significant iterations. However, the reconstruction difficulty trend aligns closely with that of a normal model.

The underlying reason is clear: as the target model's training accuracy increases, the gradient of the target model decreases, while the gradient changes of the backdoor samples remain large. This discrepancy causes the learning rate of the gradient inversion algorithm to far exceed the gradient change needed for trigger reconstruction. This prevents the gradient inversion algorithm from converging and results in failed backdoor reconstruction. Therefore, traditional gradient inversion algorithms will fail in backdoor attack scenarios where poisoning occurs at later stages.

Appendix B

How do different learning rates affect the effectiveness of gradient inversion?

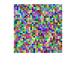
As shown in Figure 3, as the learning rate increases from 0.05 to 0.5, the efficiency of the gradient inversion algorithm's image reconstruction gradually improves. However, when the learning rate is raised to 1, the reconstruction efficiency not only fails to improve further but actually decreases. This can be observed in the reconstruction quality of the trigger in the upper left corner of the image. This occurs because the gradient inversion algorithm uses an optimizer to reconstruct the initial image, with the goal of gradually aligning the reconstructed image's gradient with the target gradient.

To ensure effective reconstruction, the learning rate must be carefully set. A learning rate that is too high may cause the reconstruction gradient to diverge from the target gradient, making it difficult to converge to a smaller value, or even cause the gradient to explode by deviating from the normal optimization path. Conversely, a learning rate that is too low can negatively impact the efficiency of the inversion algorithm's reconstruction. Therefore, the setting of the learning rate theoretically has a significant impact on the efficiency of gradient inversion. In our TMM, we used the L-BFGS optimizer instead of the commonly used Adam optimizer. The reason is that while Adam is generally robust to noisy data, its performance can be affected by extreme outliers or highly noisy datasets. This can cause it to overshoot the optimal reconstruction, resulting in images that are similar in content but different in detail from the original images, or it may fail to reconstruct successfully.

Appendix C

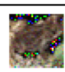
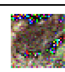
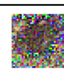
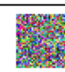
How do different random seeds affect the reconstruction of backdoor samples?

We selected 4 different random seeds for reconstruction and ranked them based on the reconstruction performance. As shown in Figure 4, the iteration count for image reconstruction is significantly affected by the initial samples generated from different random seeds. This is primarily because the gradient inversion algorithm relies on a virtual

Training epochs	40	60	80	100	110	111
Model accuracy	50.74%	53%	54.52%	55.76%	56.33%	56.39%
Reconstructed image						
Backdoor sample gradient L1 norm	1577.33154296875	1667.228759765625	1655.410400390625	1676.8856201171875	1694.2374267578125	1696.2197265625
Reconstructed sample gradient L1 norm	1577.3323974609375	1667.228271484375	1655.4144287109375	1676.8853759765625	1694.2371826171875	1696.162353515625
Difference in L1 norm	0.18741171061992645	0.3238493800163269	0.45534688234329224	0.6734241843223572	1.2251181602478027	22.587656021118164
Backdoor sample gradient L2 norm	568.552001953125	618.8680419921875	591.9664916992188	590.5633544921875	596.1128540039062	596.7151489257812
Reconstructed sample gradient L2 norm	568.5526733398438	618.865966796875	591.9658813476562	590.5646362304688	596.1141967773438	596.612548828125
Difference in L2 norm	8.422342943958938e-06	2.5621189706726e-05	4.891752541880123e-05	0.00010644702706485987	0.00034519220935180783	0.11933054029941559

Backdoor image

Figure 1: Gradient Inversion Becomes Increasingly Difficult as Training Cycles Progress.

Training epochs	20	40	60	80	100
Model accuracy	45.53%	50.74%	53%	54.52%	55.76%
Reconstructed image					
Clean sample gradient L1 norm	887.404296875	725.7664184570312	640.0916748046875	580.51318359375	551.1105346679688
Reconstructed sample gradient L1 norm	887.4052734375	725.762939453125	640.0850219726562	580.5508422851562	2202.3525390625
Difference in L1 norm	0.2739599049091339	0.6243965029716492	1.3496677683425903	1.8629658222198486	2433.47900390625
Clean sample gradient L2 norm	211.7848358154297	137.4261932373047	104.29869079589844	82.72254943847656	72.35628509521484
Reconstructed sample gradient L2 norm	211.7852783203125	137.4254913330078	104.29846954345703	82.7342300415039	1309.88427734375
Difference in L2 norm	1.8080167137668468e-05	9.41141479415819e-05	0.00043952997657470405	0.0007784849731251597	1365.1292724609375

image

Inversion algorithm optimization failed

Figure 2: The Effectiveness of Gradient Inversion Becomes Increasingly Difficult with Delayed Poisoning Start Time.

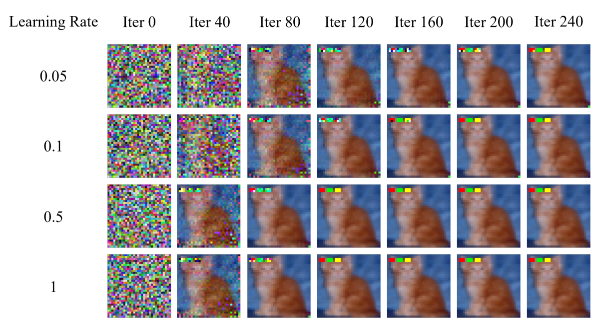


Figure 3: The trigger reconstruction performance of TMM under different learning rate

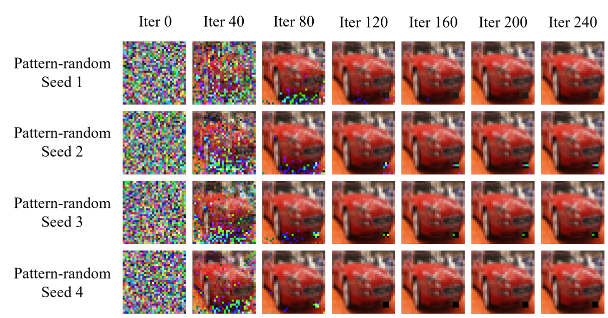


Figure 4: The trigger reconstruction performance of TMM under different random seeds

dataset as the initial sample, which is then gradually refined to reduce the gap with the target image, thereby achieving image reconstruction.

The pixel value differences in virtual data generated by different random seeds introduce uncertainty, resulting in varying distances between the virtual data and the target image. This variability impacts the difficulty of image reconstruction. If the distance is too large, the reconstruction process may slow down, as seen with Seed 1 and Seed 4, both of which successfully reconstructed the trigger, but at different speeds. On the other hand, if the distance is too small, it may lead to triggers reconstruction failure, as with Seed 2 and Seed 3, where some noise points on the trigger could not be reconstructed. Thus, the initial value of virtual data theoretically has a significant impact on the performance of gradient inversion. In practice, we mitigate the effect of different initial values on reconstruction performance by adjusting various initial noise levels.

References

Lyu, X.; Han, Y.; Wang, W.; Liu, J.; Wang, B.; Liu, J.; and Zhang, X. 2023. Poisoning with cerberus: Stealthy and colluded backdoor attack against federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 9020–9028.

Zhang, Z.; Panda, A.; Song, L.; Yang, Y.; Mahoney, M.; Mittal, P.; Kannan, R.; and Gonzalez, J. 2022. Neurotoxin: Durable backdoors in federated learning. In *International Conference on Machine Learning*, 26429–26446. PMLR.

Zhao, B.; Mopuri, K. R.; and Bilen, H. 2020. idlg: Improved deep leakage from gradients. *arXiv preprint arXiv:2001.02610*.

Zhu, L.; Liu, Z.; and Han, S. 2019. Deep leakage from gradients. *Advances in neural information processing systems*, 32.