Weekly Report (AUG 20, 2023 - AUG 26, 2023)

Zixun Xiong
Manning College of Information and Computer Sciences, UMASS Amherst
140 Governors Dr, Amherst, MA 01002
zixunxiong@umass.edu

Abstract

This week, I read some papers on security topics in the field of federated learning (FL) and another paper on critical learning period (CLP). First of all, I followed these papers to get the main idea of them. Then, I did some analysis of their pros and cons. Through careful comparison among those papers, I got the basic ideas of modern attacks and defenses in the FL community and a new aspect of CLP.

1 How to Backdoor Federated Learning

This paper [2] introduced how to backdoor the FL system in 2 different scenes using semantic backdoors driven by a model-replacing strategy. It asserted that FL is more vulnerable to model poisoning than data poisoning, making backdoor attacks a powerful attack. [2] demonstrated this assumption in two FL literature: image classification on CIFAR-10 and next-word prediction on a Reddit corpus. It turned out that even in a one-shot attack, in which case only one attacker is selected in a single round of aggregation, the backdoor success rate can reach 100% in these scenes. On the other hand, by only controlling a small fraction of clients (e.g. 1% of all clients), attackers can keep this backdoor for further attacks without reducing the accuracy of the original task: making this attack very stealthy and robust. On the contrary, the data-poisoning task requires a way bigger number of compromised clients to achieve the same backdoor accuracy in the next-world prediction task. [2] also proves that backdoor attack is robust even when Byzantine-tolerant distributed learning (e.g. Krum sampling) and anomaly detection is applied. It also designed a simpler, yet effective train-and-scale technique to evade anomaly detectors that look at the model's weights or its accuracy on the main task.

1.1 Background and Significance

Baseline Model. Past work [1] before this paper simply attacked the FL system by poisoning local data, which needs hundreds or thousands of compromised clients. This is hard to achieve in real-world cases. [1] train its model on backdoor inputs to tell the difference between them and correctly labeled inputs. Moreover, attackers can vary the local learning rate and number of local epochs to maximize the overfitting of backdoor data. This naive method doesn't perform well in FL, since aggregation cancels most of the contributions from backdoor models.

1.2 Challenges and Advantages

[2], however, only need a small fraction of compromised clients to make the joint model to learn an embedded backdoor that can be triggered with certain features while preserving the accuracy on the original task.

Model replacement. Instead of simply modifying local data. [2] used a direct method to cause the final joint model to converge to a model with 2 tasks: (1). learning the original task with the same accuracy; (2). learning the backdoor task as a sub-task with high accuracy. Inspired by this idea, the goal has been shifted to replacing the compromised update from the Equation 5 derived from Equation 1.

$$G^{t+1} = G^t + \frac{\eta}{n}\sum_{i=1}^{m}(L_i^{t+1} - G^t) \tag{1}$$

Equation 1 is the formal description of a global update of FedOPT [4]. Since we want to replace G^{t+1} with malicious models X from compromised clients, we have:

$$X = G^t + \frac{\eta}{n}\sum_{i=1}^{m}(L_i^{t+1} - G^t) \tag{2}$$

By rearranging the equation, we move L_m^{t+1} to the left-hand side, then we get the new update equation (since $L_i^{t+1} \approx G^t$ as the global model converges):

$$\begin{aligned}
L_m^{t+1} &= \frac{n}{\eta}X - (\frac{n}{\eta} - 1)G^t - \sum_{i=1}^{m-1}(L_i^{t+1} - G^t) && (3)\\
&\approx \frac{n}{\eta}X - (\frac{n}{\eta} - 1)G^t && (4)\\
&= \frac{n}{\eta}(X - G^t) + G^t && (5)
\end{aligned}$$

Thus, we only need to change the learning rates of compromised clients to $\gamma = \frac{n}{\eta}$, then the global model will converge to the malicious model. Since the malicious model is trained on both the original task and the backdoor task, we will get a backdoor in the global model.

Semantic Backdoor Attack. Compared with adversarial examples [3] which are aimed at finding boundaries between the model's representations of different classes to produce malicious inputs that are misclassified by the models, semantic backdoor attacks [2] shift these boundaries intentionally so that the model will output wrong classes. Thus, semantic backdoors don't need to change the input during the testing time and are thereby more flexible and powerful than adversarial example attacks. Since semantic backdoors save the trouble of applying physical modifications on targets, all we need to do is just to train malicious local models during the training period. What's more, it can even cause the joint server model to misclassify unmodified inputs (e.g. cars with certain colors as birds).

1.3 Reasults and Experiments

In the experiments, [2] compared the performance between baseline data-poisoning with model replacement backdoors. Researchers found that data poisoning failed in both scenes. Meanwhile, model replacement achieved a high single-shot attack backdoor accuracy without sacrificing much global accuracy($< 0.1\%$).For repeated attacks, the baseline was only successful when the percentage of malicious clients was more than 50% of all clients. Researchers also compared the performance of baseline and model replacement attacks by Pixel-pattern backdoor, finding that model replacement still had excellent performance. On the other hand, the baseline failed completely.

[2] also illustrated how long the backdoor lasts when injected at different phases, finding that the backdoor is most efficient (standing for a relatively long time) when the global model starts to converge.

1.4 Disadvantages of the Paper

As illustrated in the paper, η is critical to the model replacement attack. However, in real-world scenes, the attacker might don't know η directly. Small η will cancel out, big ones will be easier to detect since the norm is growing accordingly. The number of backdoors will also increase the norm of the update.

2 Critical Learning Periods Emerge Even in Deep Linear Networks

References

[1] T. Gu, B. Dolan-Gavitt, and S. Garg, "BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain," arXiv preprint arXiv:1708.06733, 2017.

[2] Bagdasaryan, Eugene, et al. "How to backdoor federated learning." International conference on artificial intelligence and statistics. PMLR, 2020.

[3] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami. Practical black-box attacks against machine learning. In ASIA CCS, 2017.

[4] Reddi, Sashank, et al. "Adaptive federated optimization." arXiv preprint arXiv:2003.00295 (2020).

[5] Feng, Yu, et al. "Fiba: Frequency-injection based backdoor attack in medical image analysis." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[6] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in Proc. of NeurIPS, 2017

[7] T. D. Nguyen, P. Rieger, H. Chen, H. Yalame, H. M ollering, H. Fereidooni, S. Marchal, M. Miettinen, A. Mirhoseini, S. Zeitouni et al., "Flame: Taming backdoors in federated learning," Cryptology ePrint Archive, 2021.

[8] M. S. Ozdayi, M. Kantarcioglu, and Y. R. Gel, "Defending against backdoors in federated learning with robust learning rate," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, 2021, pp. 9268–9276.

[9] Z. Zhang, X. Cao, J. Jia, and N. Z. Gong, "Fldetector: Defending federated learning against model poisoning attacks via detecting malicious clients," in Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022, pp. 2545–2555.

[10] Zhuang, Haomin, et al. "Backdoor Federated Learning by Poisoning Backdoor-Critical Layers." arXiv preprint arXiv:2308.04466 (2023).

[11] Lockdown: Backdoor Defense for Federated Learning with Isolated Subspace Training