Poisoning Attacks to Federated Learning

Jingyu Peng, Libo Zhang, Shucheng Zhang
Duke University

Abstract

In federated learning, model training is divided into several client devices. Each client device has local model with its own data. After the local devices finish training, the center device collects model parameters from clients and build global model with aggregation techniques. However, standard federated learning can be easily attacked by attackers. They can attack the local training datasets and local model parameters so that the global model will get wrong parameters from the local devices and there will be huge drop of testing accuracy. In our project, we aim at implementing federated learning and implementing two kinds of poisoning attack: data poisoning attack and local model poisoning attack, to test the robustness of federated learning. Additionally, we discuss the defense methods against the two kinds of attack.

Objectives

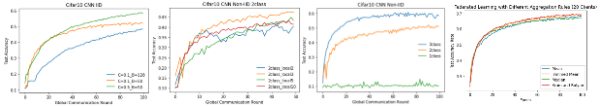
1. Implement Federated Learning and study factors influencing model performance.
2. Implement data poisoning attacks and study effectiveness of the attack and check the vulnerability of Federated Learning.
3. Implement model poisoning attacks and study factors influencing attack effect.
4. Discuss some defense methods against poisoning attacks.

Methods

1. We use FedAvg as the algorithm on IID dataset. Additionally, we use FedProx algorithm on the non-IID dataset. Finally, we use different types of aggregation for robustness and for attack.
2. We use targeted label flipping attacks as the data poisoning attack method. We first turn some of the local clients into malicious participants with poisoned data during training. For malicious participants' training data, we change all labels of source class into labels of target class.
3. We implement model poisoning attack with aim to manipulate the training process to craft model parameters. Malicious parameters involved will make the model unable and increase the testing error rate dramatically.
4. We discuss defend method of poisoning attack based on our experiment and the-state-of-art technique.

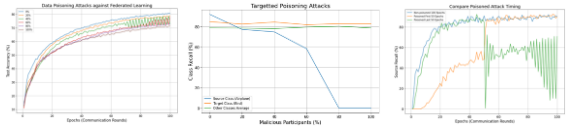
Results

1. Federated Learning



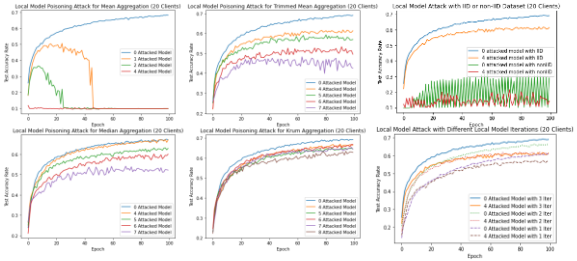
1. Smaller batch size brings better performance. Larger ratio of clients brings better performance.
2. Local update three times per round is best and choice of appropriate local update is necessary.
3. Training on IID data brings better performance compared with training on non-IID data and different extend of non-IID data brings different performance.
4. Different aggregation method brings different performance.

2. Data poisoning attack



1. The test accuracy is lower with malicious percentage increases. When the malicious percentage goes above 80%, we can observe that the testing accuracy significantly decreases about 9%.
2. If the attacks only occur during the early period of training, we can observe that the global model's source class recall can recover in a very short period and can still reach relatively good accuracy compared with the non-poisoned model.

3. Model poisoning attack



1. The result shows our attack increases the error rates significantly as we compromise more work devices. Specifically, in the trimmed mean and the median situations, when we attack 35% local model parameters, the error rate can increase 30% at most, while in the Krum situation, the increase is not obvious.
2. To study the influence of training data type influence on the attack, we used a 20% attack rate on trimmed mean and the models are trained by IID or non-IID data. The result shows training with non-IID data is much easier to be attacked.

Conclusion

1. We find that larger ratios can improve the global model's performance and smaller batch size can improve the global model's performance a lot. We also find local updating three times in one round brings the best performance, while a small or large number brings worse performance. It shows that appropriate choice of local update is necessary. We find training on IID data brings better performance compared with training on non-IID data and different extend of non-IID data brings different performance. The highly extend of non-IID will bring bad performance.
2. Data poisoning attacks affect the global model testing accuracy negatively. Increasing the percentage of malicious participants will decrease the global model testing accuracy. Targeted label flipping attacks can significantly jeopardize the source class recall, while the target class recall and all other classes' average recall are not heavily influenced. Attackers must inject poisoned data to turn some local clients into malicious participants during the last dozens of federated learning epochs (communication rounds).
3. We prove local model poisoning attacks can easily influence the performance of federated learning. To increase the error rates of the learned model, the attacker can compromise more worker devices. On the other hand, using datasets with a less non-IID degree and reducing iterations on local models can make the global more robust. For defense strategy, the byzantine-robust federated learning model performs well under the same degree of attack compared with traditional learning, while ERR and LFR generalized from data poisoning attack only have limited success in this situation.
4. To detect, we may compare local clients' model gradients[2] during federated training process. To defend data poisoning, we could use clustering methods to exclude these malicious gradients when averaging weights to the global model.
5. To defend model poisoning attack, there are two ways:
   1) Byzantine-Robust Aggregation Method[3]: As we mentioned before, it is less likely for an attacker to influence the global model significantly by manipulating only a few local models due to the more advanced aggregation methods (Trimmed mean, median, and Krum).
   2) Error Rate based Rejection (ERR)[3]: This method is based on RONI. It will remove the local models which have a large negative impact on the error rate of the global model. Loss Function based Rejection (LFR)[3]: Similarly, this method which is inspired by TRIM will remove the local model which results in a large loss.

References

[1] McMahan, B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. y. (2017). *Communication-Efficient Learning of Deep Networks from Decentralized Data*. Paper presented at the Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research. https://proceedings.mlr.press/v54/mcmahan17a.html

[2] Tolpegin, V., Truex, S., Gursoy, M. E., & Liu, L. (2020). *Data poisoning attacks against federated learning systems.* Paper presented at the European Symposium on Research in Computer Security.

[3] Fang, M., Cao, X., Jia, J., & Gong, N. (2020). *Local model poisoning attacks to byzantine-robust federated learning*. In *29th {USENIX} Security Symposium ({USENIX} Security 20)* (pp. 1605-1622).