Evaluation of Data Poisoning Attack on Centralized and Federated Learning environments

Safiia Mohammed [1] Sherif Saad [2]

[1]Phd Student @ University of Windsor [2]Associate Professor @ University of Windsor

Abstract

This poster focus on investigating the effectiveness of data poisoning attacks in centralized and federated learning environments. The research utilizes the Flower framework to establish a federated learning setting, which introduces unique challenges and possibilities for malicious actors.
The evaluation involves comparing the impact of data poisoning attacks on two datasets, CIFAR10 and MNIST—the attack success rate used as a metric to evaluate the efficacy of the attacks in both environments. The results indicate that federated learning exhibits higher resistance to data poisoning attacks when applied to the CIFAR10 dataset. However, centralized learning shows a slightly higher resilience level than federated learning when applied to the MNIST dataset.

Centralized and Federated Learning

Machine learning involves supervised and unsupervised learning, with supervised learning relying on labeled data and unsupervised learning using unlabeled data. Security vulnerabilities in supervised learning are a concern.

- **Centralized machine learning** refers to a traditional approach where data is collected and centralized in a single location or server for training and building machine learning models.
- **Decentralized machine learning** addresses privacy concerns by training local models and sharing updates with a central server.

Data poisoning attack in Centralized and Federated Learning

Data poisoning attacks threaten machine learning systems by degrading trained models and concealing them. They evaluate effectiveness using accuracy and attack success rate. Adversaries modify training data, with success varying depending on the adversary's knowledge of the target model. [3]
Although data poisoning attacks on federated learning environments have achieved their objectives, We can notice that the researchers depend on simulated learning environments. However, the effectiveness of these attacks on the production environment is still questionable and requires more investigation. The poisoning attack condected in our research is already proposed in [4, 2]

Threat Model

- **Attacker's Goal:** The attacker's primary objective is to misclassify instances from a class to a different one. The CIFAR10 and MNIST datasets were evaluated for their effectiveness. In the case of CIFAR10, the target is to misclassify instances labeled as class 9 and classify them as class 1. On the other hand, for the MNIST dataset, the goal is to misclassify examples belonging to class 2 and classify them as class 3.
- **Attacker's knowledge:** White-box attack involves attacker with model architecture knowledge and dataset access, while federated learning the attacker may have access to a portion of the dataset.
- **Convex Polytope (CP) data poisoning attack:** Convex Polytope attack generates poisoned data, transforming target instance's feature representation into convex combination using optimization problem, achieving optimal variety of features.This is achieved by solving the following optimization problem to find the optimal variety of features for the poisons.[2]

$$X_p = \underset{\{c_j\},\left\{x^{(j)}\right\}}{\operatorname{argmin}} \frac{1}{2} \frac{\left\| f\left(x_t\right) - \sum_{j=1}^{J} c_j f\left(x^{(j)}\right) \right\|_2^2}{\left\| f\left(x_t\right) \right\|_2^2}$$
$$\text{subject to } \sum_{j=1}^{J} c_j = 1$$
$$\text{and } c_j \geq 0 \forall j,$$
$$\text{and } \left\| x^{(j)} - x_b^{(j)} \right\|_\infty \leq \varepsilon \forall j.$$

- **Evaluate the success of the attack:** The evaluation of the attack is based on the Attack Success Rate (ASR), which represents the percentage of instances where the attack successfully misclassifies the target class into another class.

Poisoning attack on Centralized Environment

The Resnet18 DL model is tested on a pre-trained CIFAR10 dataset, achieving a test accuracy of 88%. Subsequently, the same model is evaluated on a poisoned dataset to assess its performance during training exposure to poisoned data.
The attack is tested with 100 images , a 15% success rate in the poisoning attack is observed. Additionally, we tested the model's performance with different trainset sizes (2500,8000 and 14000 samples).

Summery of the results

MNIST dataset achieves an accuracy of 99%; both are trained with SGD optimization. On the CIFAR10 dataset, when there are 25 poisoned samples on the trainset, the ASR values are 15%, 11%, and 7% for 2500, 8000, and 14000 train sizes, respectively. When the poisoned samples increased to 125, the ASR was 9%, 12% and 11% for 2500, 8000, and 14000 train sizes, respectively. (see Table 1).
On the MNIST dataset(See Table 2), when the trainset size is 2500, the ASR is 15%; for a trainset size of 8000, the ASR decreases to 11%, with a further increase to a trainset size of 14000, the ASR reduces to 7%. Under the Adam o)ptimizer, the ASR drops to 9% for a trainset size of 2500, 6% for a trainset size of 8000, and slightly increases to 10% for a trainset size of 14000. Under the RMSprop optimizer, the ASR is 12% for a trainset size of 2500, 13% for a trainset size of 8000, and 11% for a trainset size of 14000.

Table 1. Attack Success Rate on CIFAR10 dataset

Number of poisoned data in train set	Train-set Size	Attack Success Rate
25	2500	10%
	8000	14%
	14000	12%
125	2500	9%
	8000	12%
	14000	11%

Poisoning attack on Federated Learning Environment

Flower Federated Learning framework: Flower is an open-source framework for developing and deploying federated learning systems, offering a user-friendly interface, tools, and a flexible training process for efficient and scalable implementation [1].
Attack configuration: The experiments conducted on the Flower Federated Learning (FL) platform involved the following two configurations: Three honest clients and one out of the three clients being malicious.
Summery of the results : The test accuracy on the CIFAR10 dataset remains consistent at 88.7%. Regardless of the trainset sizes, the attack success rate (ASR) remains at 3% with 25 poisoned data samples and 4% with 125 poisoned data samples.
On the MNIST dataset, the ASR is consistently 14% regardless of the optimizer or trainset size.
The test accuracy values vary slightly depending on the optimizer and trainset size. Using SGD optimizer, the test accuracies are 97.7%, 97.6%, and 96.3% for trainset sizes 2500, 8000, and 14000, respectively. With the Adam optimizer, the test accuracy values are 97.9%, 96.6%, and 96.6%. Finally, with the RMSprop optimizer, the test accuracy values are 97.99%, 97.9%, and 96.7% for trainset sizes 2500, 8000, and 14000, respectively.

Table 2. CP attack on MNIST dataset with 25 poisoned samples.

Learning Environment	Optimizer	Train-set size	Attack Success Rate	Test accuracy
Centralized Model	SGD	2500	15%	99%
		8000	11%	
		14000	7%	
	Adam	2500	9%	
		8000	6%	
		14000	10%	
	RMSprop	2500	12%	
		8000	13%	
		14000	11%	
Federated Learning (Flower)	SGD	2500	14%	97.7%
		8000		97.6%
		14000		96.3%
	Adam	2500		97.7%
		8000		96.6%
		14000		
	RMSprop	2500		97.99%
		8000		
		14000		96.7%

Discussion

The ResNet18 model achieves an accuracy of 88.7% on the CIFAR10 dataset. In a centralized learning environment, the attack success rate (ASR) for poisoned samples in the train set ranges from 9% to 14%, indicating the relatively low effectiveness of the poisoned attack. However, in the Federated Learning (FL) environment, the ASR remains fixed at 3% or 4%, demonstrating the resilience of FL against poisoned attacks. The impact of increasing the number of poisoned samples in the train set does not necessarily result in a higher ASR or a more significant attack impact.
When comparing optimizers in the centralized environment on the CIFAR10 dataset, the Adam optimizer shows a relatively lower ASR than SGD and RMSprop, indicating its potential to improve model resilience. For the MNIST dataset, regardless of the optimizer or trainset size, the ASR remains consistent at 14%, indicating its higher vulnerability to data poisoning attacks compared to CIFAR10 in the FL environment.

Future Work

In our future work, we aim to investigate the efficacy of various defence mechanisms in countering data poisoning attacks within centralized and federated learning environments. We will focus on developing resilient techniques that can effectively detect and mitigate the adverse effects caused by poisoned samples on model performance. Furthermore, we plan to broaden the scope of our evaluation by including a diverse array of machine-learning models, datasets, and optimization algorithms. This expanded analysis will provide a more comprehensive understanding of the vulnerabilities and potential defence strategies against data poisoning attacks.

References

[1] Daniel J Beutel, Taner Topal, Akhil Mathur, Xinchi Qiu, Javier Fernandez-Marques, Yan Gao, Lorenzo Sani, Kwing Hei Li, Titouan Parcollet, Pedro Porto Buarque de Gusmão, et al.
Flower: A friendly federated learning research framework.
arXiv preprint arXiv:2007.14390, 2020.

[2] Avi Schwarzschild, Micah Goldblum, Arjun Gupta, John P Dickerson, and Tom Goldstein.
Just how toxic is data poisoning? a unified benchmark for backdoor and data poisoning attacks.
In *International Conference on Machine Learning*, pages 9389–9398. PMLR, 2021.

[3] Zhiyi Tian, Lei Cui, Jie Liang, and Shui Yu.
A comprehensive survey on poisoning attacks and countermeasures in machine learning.
ACM Computing Surveys, 55(8):1–35, 2022.

[4] Chen Zhu, W Ronny Huang, Hengduo Li, Gavin Taylor, Christoph Studer, and Tom Goldstein.
Transferable clean-label poisoning attacks on deep neural nets.
In *International Conference on Machine Learning*, pages 7614–7623. PMLR, 2019.