Evaluation of Data Poisoning Attack on Centralized and Federated Learning environments

Safiia Mohammed.[*]
School of Computer Science
University of Windsor
Windsor ON, Canada
mohamm7d@uwindsor.ca

Abstract

This paper investigates 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.

1 Background

1.1 Centralized and Federated Learning

Machine learning encompasses supervised and unsupervised learning, where supervised learning relies on labeled data, and unsupervised learning operates with unlabeled data. Recent research has focused on the security vulnerabilities of supervised learning, particularly regarding poisoning attacks.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.Federated learning has emerged as a solution to address privacy concerns, where clients train local models and share model updates with a central server to create a global model. This decentralized approach ensures data privacy and security without directly sharing clients' private data.

1.2 Data poisoning attack in Centralized and Federated Learning

Data poisoning attacks have emerged as a significant threat to machine learning systems, compromising the integrity and performance of trained models. These attacks aim to degrade the target model's performance and may also attempt to conceal the attack or control the level of abnormality introduced. The effectiveness of such attacks is evaluated using metrics like accuracy and attack sucess rate. Adversaries can conduct poisoning attacks by modifying the training data. The attack's success depends on the adversary's knowledge of the target model, ranging from complete visibility in white-box attacks to limited visibility in black-box attacks, where certain model aspects are not fully known. These vulnerabilities in machine learning models present a new challange for cybersecurity research Tian et al. [2022]

[*] Ph.D student in computer science.

5th Deep Learning Indaba Conference (DLI 2023).

Data poisoning attacks aim to compromise data integrity. These attacks involve manipulating or injecting malicious data into a dataset during the training phase to influence the performance or behaviour of machine learning models. Previous studies discussed different data poisoning attack strategies and objectives.
Ramirez et al. [2022] in their research focuses on data poisoning attacks involving label-flipping. The adversary intentionally alters the labels of some instances belonging to a specific class and assigns them the label of another class during the training process. The attacker's objective is to compromise the integrity of the targeted machine-learning model by significantly reducing its overall accuracy and potentially causing the misclassification of specific samples. These attacks target machine learning classifiers for malware detection using mobile exfiltration data. The results show that the accuracy of the models is reduced after the attack.

Tolpegin et al. [2020] considered a data poisoning attack where a subset of participants in federated learning (FL). The adversary aims to manipulate the learned parameters so that the final global model (M) shows high errors for specific classes. The effectiveness of the attack is evaluated on two datasets: CIFAR-10 and Fashion-MNIST. The results indicate that as the number of participants increases to 50 in CIFAR-10, the loss increases to 70.5%. Similarly, in the Fashion-MNIST dataset, with an increased number of participants, the loss increases to 58.9%.

Another attack Shi et al. [2022] assumes a federated learning system with n clients, including m, controlled by an attacker. The attacker aims to decrease the global model's accuracy without focusing on specific classes. Experiments were conducted on Fashion-MNIST and CIFAR-10 datasets. Experimental results indicate that federated learning is susceptible to adversarial samples, reducing accuracy by over 5%.

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.
In this paper the poisoning attack is already proposed in Zhu et al. [2019], Schwarzschild et al. [2021]

2 Threat Model

Attacker's Goal: In our experiments, the attacker's main objective is to misclassify instances from a particular class to a different class. We evaluated the CIFAR10 Krizhevsky et al. [2009] and MNIST Deng [2012] datasets to assess the effectiveness of these attacks. Specifically, 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: In our study, we consider a white-box attack scenario, where the attacker possesses knowledge of the model architecture and has access to the dataset. In the case of federated learning, the attacker may have access to a portion of the dataset.

Convex Polytope (CP) data poisoning attack: The Convex Polytope (CP) attack involves generating poisoned data so that the feature representation of the target instance can be expressed as a convex combination of the feature representations of the poisoned instances. This is achieved by solving an optimization problem to find the optimal variety of features for the poisons.

3 CP attack setting

Datasets and training settings: In this paper, we conducted a poison attack on the Resnet18 deep-learning model on CIFAR10 and MNIST datasets. CIFAR-10 Dataset has ten different categories of images in it. There are 60000 images of 10 classes: Airplane, Automobile, Bird, Cat, Deer, Dog, Frog, Horse, Ship, and Truck. Each class has $5,000$ images in the training set and 1000 in the testing set.
MNIST dataset comprises 28x28 pixel images depicting handwritten digits ranging from 0 to 9. It

includes a training set of 60,000 examples and a test set of 10,000 examples.
There are 25 poisoned data samples out of 2500 samples in the trainset. Subsequently, the number of poisoned data samples in the trainset is increased to 125. During the poisoning attack, the model is trained with three optimizers, SGD, Adam and RMSprop, on different trainset- sizes: 2500, 8000, and 14000 for both datasets.
Execution environment configuration: The CoLab Pro platform is utilized as the execution environment for this work. It is connected remotely to Visual Studio Code (VSC) through Ngrok, a secure tunnelling service that enables instant access to remote systems. This connection allows seamless collaboration and access to the CoLab environment from VSC.goo

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.

3.1 Poisoning attack on Centralized Environment

Firstly, 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 success rate is used as an evaluation metric that measures the number of times the attack successfully misclassifies the target class. In this case, the target class is 9, while the poisoned class is 1. Through 100 trials (using 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)

3.1.1 Summery of the results

Initially, The ResNet18 model achieves an overall accuracy of 88.7% on CIFAR10, while The 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.

3.2 Poisoning attack on Federated Learning Environment

3.2.1 Flower Federated Learning framework

Flower is an open-source framework that simplifies the development and deployment of federated learning systems. It offers developers a high-level interface and tools to define tasks, specify network architectures, and implement communication protocols while providing a user-friendly environment for seamless integration. The framework supports various learning algorithms, including federated averaging and SGD. It gives flexibility for customizing the training process and enables efficient and scalable implementation of federated learning. Beutel et al. [2020].

3.2.2 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.

3.2.3 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.

4 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. This suggests that factors beyond the quantity of poisoned data influence the attack's effectiveness.
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.
Overall, these results highlight the importance of considering dataset characteristics, optimizer selection, and the presence of poisoned samples when evaluating the vulnerability of machine learning models to data poisoning attacks.

5 Conclusion

This study investigates data poisoning attacks on machine learning models in two distinct environments: the traditional centralized and federated learning environments. The evaluation is conducted on CIFAR10 and MNIST datasets, considering different train-set sizes and optimizer choices to analyze the variations between the two environments. The findings emphasize the importance of factors such as dataset characteristics, optimizer selection, and the existence of poisoned samples when assessing the vulnerability of machine learning models to data poisoning attacks.

6 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

Colaboratory. `https://research.google.com/colaboratory/faq.html`. Verified: 2023-07-5.

D. J. Beutel, T. Topal, A. Mathur, X. Qiu, J. Fernandez-Marques, Y. Gao, L. Sani, K. H. Li, T. Parcollet, P. P. B. de Gusmão, et al. Flower: A friendly federated learning research framework. *arXiv preprint arXiv:2007.14390*, 2020.

L. Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

M. A. Ramirez, S. Yoon, E. Damiani, H. A. Hamadi, C. A. Ardagna, N. Bena, Y.-J. Byon, T.-Y. Kim, C.-S. Cho, and C. Y. Yeun. New data poison attacks on machine learning classifiers for mobile exfiltration. *arXiv preprint arXiv:2210.11592*, 2022.

P. P. Ray. A review on tinyml: State-of-the-art and prospects. *Journal of King Saud University-Computer and Information Sciences*, 34(4):1595–1623, 2022.

A. Schwarzschild, M. Goldblum, A. Gupta, J. P. Dickerson, and T. 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.

L. Shi, Z. Chen, Y. Shi, G. Zhao, L. Wei, Y. Tao, and Y. Gao. Data poisoning attacks on federated learning by using adversarial samples. In *2022 International Conference on Computer Engineering and Artificial Intelligence (ICCEAI)*, pages 158–162. IEEE, 2022.

Z. Tian, L. Cui, J. Liang, and S. Yu. A comprehensive survey on poisoning attacks and countermeasures in machine learning. *ACM Computing Surveys*, 55(8):1–35, 2022.

V. Tolpegin, S. Truex, M. E. Gursoy, and L. Liu. Data poisoning attacks against federated learning systems. In *Computer Security–ESORICS 2020: 25th European Symposium on Research in Computer Security, ESORICS 2020, Guildford, UK, September 14–18, 2020, Proceedings, Part I 25*, pages 480–501. Springer, 2020.

C. Zhu, W. R. Huang, H. Li, G. Taylor, C. Studer, and T. Goldstein. Transferable clean-label poisoning attacks on deep neural nets. In *International Conference on Machine Learning*, pages 7614–7623. PMLR, 2019.

A Appendix

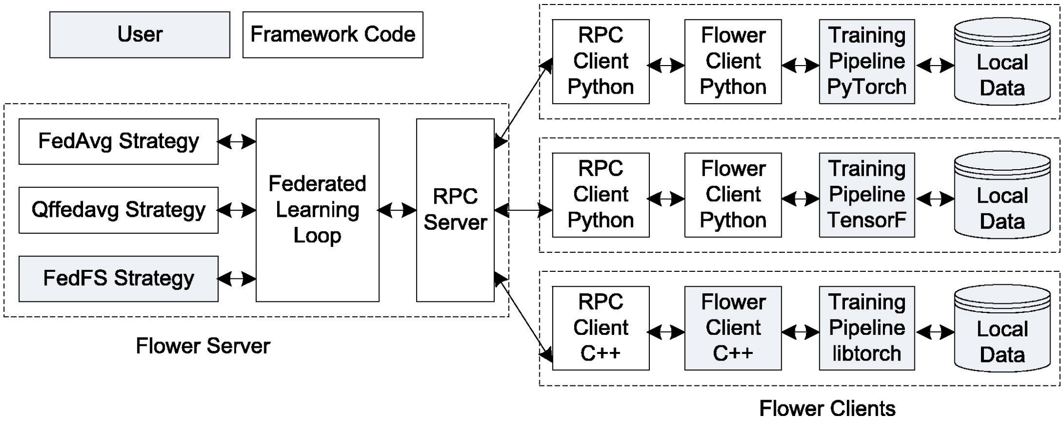

Figure1: Flower federated learning architecture Ray [2022].

Number of poisoned data on trainset	Trainset Size	Attacksuccess rate
25	2500	10%
	8000	14%
	14000	12%
125	2500	9%
	8000	12%
	14000	11%

Table 1: Attack Success Rate on CIFAR10 dataset.

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%	
FederatedLearning (Flower)	SGD	2500	14%	97.7%
		8000		97.6%
		14000		96.3%
	Adam	2500		97.9%
		8000		96.6%
		14000		96.6%
	RMSprop	2500		97.99%
		8000		97.99%
		14000		96.7%

Table 2: Attack Success Rate on MNIST dataset.