Targeted Clean-Label Poisoning Attacks on Federated Learning

Ayushi Patel(✉) and Priyanka Singh

Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar 382007, Gujarat, India
201801203@daiict.ac.in

Abstract. Federated Learning (FL) has become one of the most extensively utilized distributed training approaches since it allows users to access large datasets without really sharing them. Only the updated model parameters are exchanged with the central server after the model has been trained locally on the devices holding the data. Because of the distributed nature of the FL technique, there is a possibility of adversarial attacks that aim to manipulate the behavior of the model. This paper explores targeted clean-label attack in which adversaries inject poisoned images into compromised clients' dataset to alter the behaviour of the model on a specific target image at test time. The standard CIFAR10 dataset is used in this study to conduct various experiments and manipulate the image classifier. This study discovered that the behavior of a FL model can be altered maliciously towards a specific target image without significantly affecting the model's overall accuracy. In addition, the attack's impact grows in direct proportion to the number of injected poisonous images and malicious client (i.e. controlled by adversaries) participating in the FL process.

Keywords: Federated Learning · Targeted attacks · Data poisoning · Clean label

1 Introduction

The artificial intelligence industry is dominated by data-driven machine learning methods. For the model to work well in broader deployments, a large scale diversified dataset is needed, which is not always available due to a variety of factors such as competitive dynamics between different organisations, legal restrictions, user discomfort, privacy concerns, and so on. Because of the aforementioned challenges, there has been an increase in the number of proposals for various distributed training architectures. For example, under the Federated Learning (FL) paradigm, instead of collecting all essential data on a central server, the data remains on numerous edge devices e.g., computers, mobile phones, and IoT devices. Each data holding device is responsible for training the model using local data. Only the 'model parameters' are exchanged with the central server, where the global model is developed by aggregating the local parameters. So Federated

KC Santosh et al. (Eds.): RTIP2R 2022, CCIS 1704, pp. 231–243, 2023.
https://doi.org/10.1007/978-3-031-23599-3_17

Learning has evolved as a privacy-enhancing technology, allowing the model to be trained locally on data from millions of devices without actually sharing it.

Despite the fact that FL eliminates the need for a centralized database, it remains subject to adversarial attacks that might compromise the model's integrity and endanger data privacy in a situation where an adversary controls a portion of edge devices. By tempering local training data or model parameters, these devices may then be corrupted to achieve adversarial goals. Because there is no central authority to validate data, malicious clients can poison the trained global model. Despite the fact that FL restricts the malicious agent's access to a subset of the data available on a few devices, this can still significantly impair model performance, and they can be reverse-engineered to reveal clients' data.

This study explores vulnerability of FL systems to various adversarial attacks attempting to alter the behaviour of global model. Broadly, the attacks on FL can be classified into two main categories:

- Attacks on the model behaviour: The goal of these attacks is to alter the model's behaviour. There may be one or more malicious clients capable of causing model behaviour to degrade by delivering poisoned updates to the central server. These attacks are difficult to detect owing to the fact that the central server has no knowledge of the client's training data.
- Privacy attacks: The goal of these attacks is to infer sensitive information about the clients/participants in the FL process. They endanger not only the privacy of the data that the clients have, but also the local model parameters that the clients provide to the central server.

This paper concentrates on the first type of attacks: attacks on the model behaviour. Shafahi et al. [1] presented targeted data poisoning attacks on neural networks as part of attacks altering the behaviour of the model, as well as a method to generate 'clean-label poisoned instances' (i.e. poisoned instances are correctly labeled) to alter classifier's behaviour. The attack occurs during training time by carefully introducing poisoned instances into the training data with the purpose of manipulating classifier behaviour on one specific target instance at time. Since the attacks are targeted, the change in overall model accuracy is trivial and hence readily overlooked while accomplishing the intended misclassification.

FL, like machine learning, is vulnerable to adversarial attacks, particularly this type of poisoning attacks, and recent research [5,6,9] has demonstrated that the FL model's functionality can be significantly damaged. So, this research focuses on extending the attack described in [1] in the context of Federated Learning. To create clean-labeled poison instances, the study employs the optimization-based strategy described in [1]. The study looks into the model's behavior using two variables: the number of malicious clients involved in the FL process and the number of poisoned instances injected by them.

The rest of the paper is organized as follows: Sect. 2 gives a brief overview of the possible attacks on FL and discusses about an optimization method that can generate poisoned instances. Section 3 presents the related work and Sect. 4 provides experimental setup, threat model and framework for data poisoning attack

on FL. Section 5 contains four different experiments on images from CIFAR-10 dataset followed by a discussion about the results in Sect. 6 and finally, the conclusion.

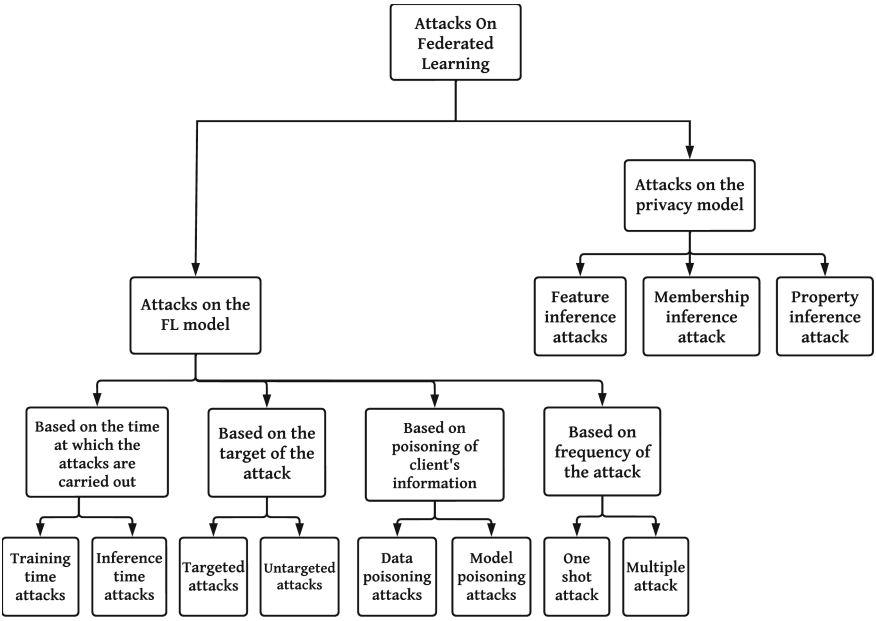

Fig. 1. Types of attacks on federated learning

2 Background Concepts

In this section, we will go through potential attacks that might occur in a federated learning situation. Thereafter, a brief overview of an optimization-based technique for generating poison instances is provided [1].

2.1 Overview of Attacks on FL

The distributed nature of FL can be exploited by the adversaries to influence model behaviour and infer sensitive information about FL participants. Broadly the attacks on FL can be categorized into: Attacks on the FL model and Privacy attacks.

Now, the attacks on the FL model can be classified further according to the four scenarios outlined below.

1. Time at which the attacks are carried out

- Training time attacks: These attacks take place during data gathering, data preparation, and model training processes. The primary purpose of these attacks is to either influence the behaviour of the model getting trained or infer information from the training data.
- Inference time attacks: These attacks occur after the model has been trained. The fundamental goal of these attacks is to obtain information about the model's characteristics.

2. Target of the attack

- Targeted attacks: These attacks are aimed at misclassifying a certain target instance or class as a different class chosen by the adversary. The key goal here is to question the system's integrity rather than to reduce overall model accuracy.
- Untargeted attacks: The primary purpose of these attacks is to change the model in a way that it ends up predicting any of the incorrect classes. These attacks cause a noticeable drop in model performance and are thus clearly detectable.

3. Poisoning of client's information

- Data poisoning attacks: The attacker tampers with the training data of one or more clients involved in the FL process. The primary goal of these attacks is to degrade the performance of the global model.
- Model poisoning attacks: The attacker directly modifies the updated weights sent to the central server by different clients.

4. Frequency of the attack

- One shot attack: These attacks are carried out at one precise moment during the training process.
- Multiple attacks: These attacks occur repeatedly during the training process, either throughout all of the training rounds or just a few of them.

Similarly, the privacy attacks can be classified further into the three categories outlined below.

1. Feature inference attacks

- The major goal of these attacks, as the name implies, is to retrieve the dataset of clients participating in the FL process. Clients either exchange their gradients or local model parameters with the central server during communication rounds, providing surface for the attackers.

2. Membership inference attacks

– Given the training data of clients participating in FL and the local model, these attacks attempt to determine whether or not this data was utilised to train the model.

3. Property inference attack

– These attacks are aimed at identifying whether or not certain properties are possessed by clients participating in FL process. For example, whether a client possesses certain qualities that are not directly related to the FL model's core goal.

2.2 Optimization Based Method to Generate Poisons

Shafahi et al. [1] proposed an optimization-based strategy for creating poisons which, when introduced to the training set, carries a change on the behaviour of the classifier. Let's start by deciphering a few terms.

– **Target instance** is an instance from the test dataset that we intend to misclassify at test time.
– **Base instance** is an instance from the test dataset. The model will misclassify the target instance with the label of base instance at test time.
– **Poison instance** is an instance generated by making imperceptible changes to base instance. It is injected to the training dataset to spur misclassification.

Let $f(x)$ denote feature representation of input x, p poison instance, b base instance and t target instance then Eq. (1) represents how poisons are generated via feature collision.

$$p = \underset{x}{argmin}\ ||f(x) - f(t)||_2^2 + \beta||x - b||_2^2 \tag{1}$$

The first term in Eq. (1) causes the poison instance to migrate in feature space toward the target instance. The second term in Eq. (1) transforms the poison instance into a base class instance. Shafahi et al. [1] proposes an algorithm to optimize Eq. (1). The forward step is a gradient descent update to reduce the L2 distance between base instance and target instance in feature space, whereas backward step is a proximal update to minimise the Frobenius distance from the base instance in input space [1].

3 Related Work

A variety of targeted attacks with the purpose of injecting a secondary or backdoor task into the model have been proposed in the literature. These are regarded effective as long as they are successful in preserving the overall accuracy of the model. Bagdasaryan et al. proposed a model-poisoning technique where the

attacker compromises one or more clients, connects basic patterns in training data with a specific target label and train the model locally using constraint-and-scale technique. The resulting model, thereafter replaces the joint model as a result of federated averaging [2]. They showcased that this type of model-poisoning attack is significantly powerful than data-poisoning attacks. Bhagoji et al. exploited the lack of transparency in the client updates and boosted the malicious client's update to overcome the effect of other clients [3]. Wang et al. trained the model using projected gradient descent (PGD) so that at every training round the attacker's model does not deviate much from the global model [4]. This was more robust than the simple model replacement strategies against a range of defense mechanisms provided in [2] and [3].

Sun et al. implemented data poisoning attacks exploiting the communication protocol and suggested bi-level data poisoning attacks - ATTacks on Federated Learning (AT^2FL) [5]. The extensive experiments carried out by the authors suggest that it can significantly damage performances of real world applications. Tolpegin et al. investigated targeted data poisoning attacks where malicious clients' 'mislabeled' data changes the behaviour of the global model [6]. They suggest a defensive approach for identifying malevolent clients who are engaged in the FL process. The study indicated that as the number of malevolent individuals rises, the detrimental influence on the global model also improves. In addition, increasing the number of malevolent players in later rounds of training can improve the efficacy of these attacks. For CIFAR-10 dataset, the findings show that if 40% of total clients are altered, the number of images correctly classified for the target class drops to 0% and overall model accuracy drops by 3.9%, from 78.3% to 74.4%. Cao et al. investigated the number of poisoned samples and attackers as variables influencing the performance of poisoning attacks [7]. The study found that the attack success rate grows linearly with the number of poisoned samples. When the number of poisoned samples is kept constant, the attack success rate increases with the increased number of attackers. The study also suggested a method called 'sniper' for removing poisoned local models from malevolent players during training.

Generative Adversarial Networks (GANs) have been widely employed in recent years to produce poisoned data since they improve the accuracy of backdoor tasks and secure the attack against potential defences. Zhang et al. developed a GAN-based approach for creating poisoned samples as well as a poisoning attack model for the FL framework [8]. It was further extended to an approach where the attacker acts as benign participant in the FL process, sending poisoned updates to the server [9]. Ligeng Zhu et al. showed that private training data can be obtained exploiting publicly shared gradients questioning the safety of gradient exchange [10].

4 Clean-Label Poisoning Attacks on Federated Learning

This paper takes into account attacks on FL models that aim to change the model behaviour during test time. The attacks are also targeted i.e., they intend to manipulate model behaviour on a single test instance. In addition, the attacker

is expected to be able to inject poisons into the training set of one or more clients, classifying the attack as targeted data poisoning attacks.

This section will cover the dataset utilized for the attack, as well as the experimental setting, the adversaries' goals, the framework used to produce poisons, and ultimately targeted data poisoning attacks in the case of FL.

4.1 Dataset

This study focuses on a range of experiments about the targeted clean-label attacks for the image classification task utilizing the CIFAR-10 [11] dataset. There are 50,000 training samples in the CIFAR-10 dataset, with 5000 samples in each of the 10 classes, i.e., aeroplane, car, bird, cat, deer, dog, frog, horse, ship, truck. There are also 10,000 test samples on the premises. This study utilizes RESNET18, which is an 18-layer deep convolutional neural network (CNN). Without the introduction of poisoned instances, the model achieved an accuracy of 78.07% in case of FL.

4.2 Experimental Setup

This research utilizes the PyTorch framework to construct a FL scenario in Python. The experiments are conducted by firstly loading the RESNET18 model that was trained on the IMAGENET dataset, but do not use *pretrained* weights and trains the model from scratch on CIFAR-10 dataset. In total, 50 clients were involved in the FL process. By default, 5 of the 50 clients were assumed to be under the control of adversaries. Also, it is assumed that the adversaries can only manipulate the training data on their local device. This study assumes a non-iid (independent and identically distributed) scenario, where each client is randomly assigned 1000 images from the CIFAR-10 dataset. This FL configuration has been trained for a total of 100 global epochs. Federated averaging (fedAvg) is utilized as central aggregator.

4.3 Threat Model

This study takes into consideration a scenario in which a portion of clients is controlled by adversaries who don't have access to training data but can access the model and its parameters. This can be exploited to generate poisoned instances. It is safe to make this assumption as a lot of traditional networks (i.e., RESNET, Inception, etc.) are utilized to extract features. It is also assumed that the federated aggregation operations are not compromised.

4.4 Adversarial Goal

The attack described in this work is targeted, focusing on misclassifying only a specific target instance during testing. It will result in a minor change in overall model accuracy, allowing the attack to go unnoticed while still accomplishing the desired purpose. Targeted attacks are harder to detect than the untargeted attacks as they are more covert.

4.5 Generating Poison Instances

To create the poison instance, the attacker will first select a base instance and a target instance from the test dataset. Thereafter he will make undetectable modifications to the base instance using the optimization-based technique described in [1]. In the experiments, we injected several poisoned instances generated with the watermarking approach mentioned in [1]. These poisoned instances were labeled as base class and seem remarkably similar to the original base instances, resulting in clean-label attacks.

4.6 Data Poisoning Attack on FL

To undertake targeted data poisoning attacks on a FL model with 50 clients, we first pick 5 clients at random to act as adversaries. Thereafter, these malicious clients train the model on poisoned dataset (clean dataset + 50 poison instances) to compromise the behaviour of the model. We consider, (1) 'bird' as our target class, (2) 'dog' as our base class (Figs. 2 and 3).

Fig. 2. Target instance from the class "bird"

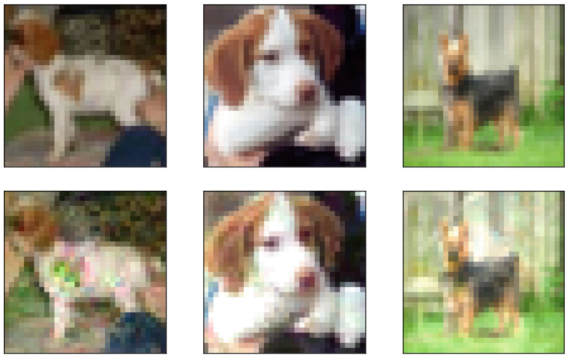

Fig. 3. The first row has three random base instances from the class "dog." The second row contains poison instances created from the respective base instances for the target instance depicted in Fig. 2.

5 Experiments and Results

We conducted the experiments on images from CIFAR-10 dataset. We considered an experimental scenario where we evaluate different *target instance* - *base instance* combinations with varying watermarking opacity. We are presenting results for three such pairs, which are described below.

1. dog vs bird (opacity: 30%)
2. airplane vs frog (opacity: 30%)
3. airplane vs frog (opacity: 20%)

Cao et al., investigated the number of poisonous images and malicious clients as variables influencing attack success rate [7]. The results obtained in this research are similar to the conclusions reached in [7].

5.1 Experiment 1

In the first experiment we have kept number of malicious clients as 5 i.e. constant and we vary number of poisonous images from 0 to 80. Figure 4 shows that the attack success rate grows almost linearly as the number of malicious images injected increases. As we increase the number of poisonous images, Model weights achieved after federated averaging of poisoned weights derives away from the original weights resulting in higher attack success rate.

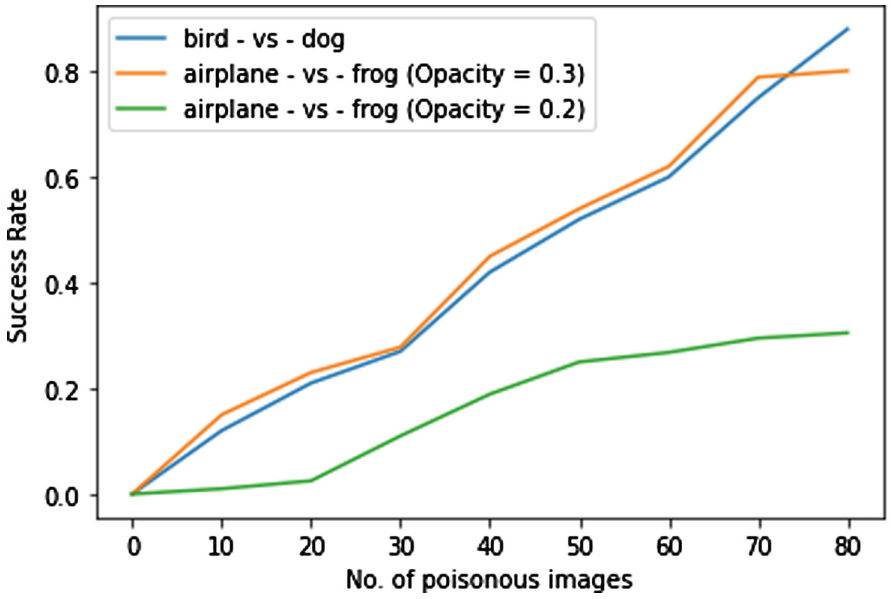

Fig. 4. Experiment 1

5.2 Experiment 2

In the second experiment, we maintained the number of poisonous images constant at 60 and varied the number of malicious clients from 0 to 6. Figure 5 shows that the attack success rate grows almost linearly as the number of malicious clients participating in the FL process increases. As we increase number of malicious clients, poisoned local models involved in the global model aggregation also increases resulting in higher attack success rate.

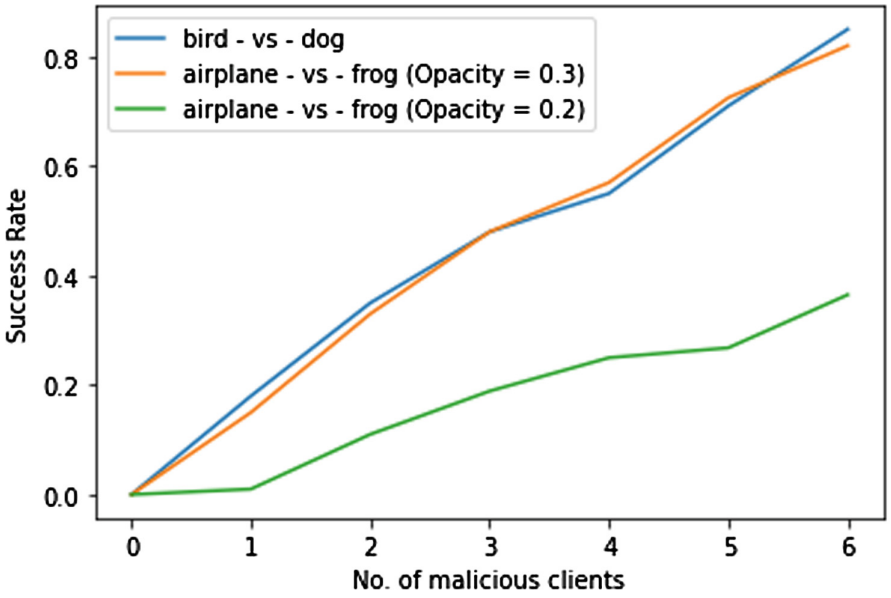

Fig. 5. Experiment 2

5.3 Experiment 3

In the third experiment, we examine overall accuracy of the model under conditions when poison instances are inserted into the dataset vs when they are not. For all cases, we maintained the total number of malicious clients participating in the FL process at 5. We will evaluate the change in overall model accuracy when the number of poisonous images increases from 50 to 70. When we insert 50 poisonous images, the initial 78.07% is reduced by an average of 0.1%. When we raise the number of poisonous images to 70, it reduces by 0.22% on average. This suggests that targeted data poisoning has little effect on the overall model accuracy while achieving the desired misclassification (Fig. 6).

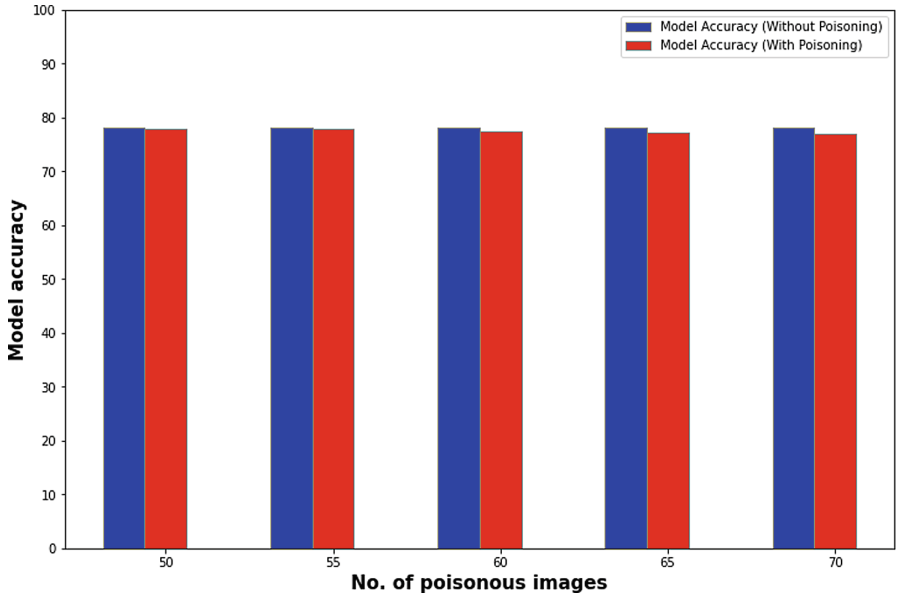

Fig. 6. Experiment 3

5.4 Experiment 4

In the fourth experiment, the study tries to determine minimum number of poisonous images required to incorrectly classify each image of a certain class. We used the CIFAR-10 dataset's 'bird' class as our target class and 'dog' as our base class. The experiment's findings show that if we create more than 80 poison images for each image in the 'bird' class, we can misclassify all of the images present in that class. In this experiment, the overall model accuracy drops significantly to 70% from the original 78.07%. This is due to the fact that in our previous experiments, only one specific image from the target class was incorrectly classified rather than all of the images from that class.

6 Discussion

If M_i^t represents local model parameter of i^th participant at t^th epoch then if we have N participants in the FL process then global parameter vector M_G^t

$$M_G^t = 1/N \sum_{i=1}^{N} M_i^t \tag{2}$$

Now, we assume that p out of these N participants are under the control of an adversary then poisoned global parameter vector $M^t_{G'}$

$$M^t_{G'} = 1/N(\sum_{i=1}^{p} M^t_i + \sum_{j=p+1}^{N} M^t_j) \tag{3}$$

The local model parameter vector of clients under an adversary's control is represented by the first term in Eq. (3). It is linearly proportional to the global parameter vector. As a result, when the number of participants under the control of an adversary increases, its contribution in calculating global parameter vector at the central server also increases. We were able to demonstrate the same in experiment 2 when we increased the number of malicious clients and saw an almost linear increase in the attack success rate.

7 Conclusion

In our research, we focused on targeted clean-label data poisoning attacks in FL environment. We were able to show how FL models are susceptible to data poisoning attacks. We demonstrated that the behaviour of a FL model can be altered to misclassify a specific test image, affecting the model's overall accuracy just slightly. We also demonstrated the impact of varied numbers of poisonous instances and malicious clients in the FL process. Based on the experiments, we concluded that the attack success rate increases as we increase the number of poisonous images and clients engaging in the learning process.

In future, we want to validate the scalability of the attack by testing the model on different datasets. In addition, we want to develop a defense mechanism against the targeted clean-label data poisoning attacks.

References

1. Shafahi, A., et al.: Poison frogs! Targeted clean-label poisoning attacks on neural networks. In: NIPS Conference (2018)
2. Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., Shmatikov, V.: How to backdoor federated learning. In: International Conference on Artificial Intelligence and Statistics, pp. 2938–2948 (2020)
3. Bhagoji, A., Chakraborty, S., Mittal, P., Calo, S.: Analyzing federated learning through an adversarial lens. In: Proceedings of the 36th International Conference on Machine Learning (ICML), vol. 97, pp. 634–643 (2019)
4. Wang, H., et al.: Attack of the tails: yes, you really can backdoor federated learning, In: Advances in Neural Information Processing Systems, vol. 33 (2020)
5. Sun, G., Cong, Y., Dong, J., Wang, Q., Lyu, L., Liu, J.: Data poisoning attacks on federated machine learning. IEEE Internet Things J. 1 (2021)
6. Tolpegin, V., Truex, S., Gursoy, M., Liu, L.: Data poisoning attacks against federated learning systems, pp. 480–501 (2020)
7. Cao, D., Chang, S., Lin, Z., Liu, G., Sun, D.: Understanding distributed poisoning attack in federated learning. In: 2019 IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS), pp. 233–239 (2019)

8. Zhang, J., Chen, B., Cheng, X., Binh, H.T.T., Yu, S.: PoisonGAN: generative poisoning attacks against federated learning in edge computing systems. IEEE Internet Things J. **8**, 3310–3322 (2021)
9. Zhang, J., Chen, J., Wu, D., Chen, B., Yu, S.: Poisoning attack in federated learning using generative adversarial nets. In: 2019 18th IEEE International Conference on Trust, Security and Privacy in Computing and Communications/13th IEEE International Conference on Big Data Science and Engineering (TrustCom/BigDataSE), pp. 374–380 (2019)
10. Zhu, L., Han, S.: Deep leakage from gradients. In: Yang, Q., Fan, L., Yu, H. (eds.) Federated Learning. LNCS (LNAI), vol. 12500, pp. 17–31. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-63076-8_2
11. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)