CERIAS

The Center for Education and Research in Information Assurance and Security

# GAN-inspired Defense Against Backdoor Attack on Federated Learning Systems

*Agnideven Palanisamy Sundar† , Feng Li† , Xukai Zou†, and Tianchong Gao⋆*
*†Indiana University-Purdue University Indianapolis, Indianapolis, IN, USA.*
*⋆Southeast University, Nanjing, Jiangsu, China.*
*agpalan@iu.edu, tgao@seu.edu.cn, {fengli,xzou}@iupui.edu*

## OVERVIEW

**Federated Learning allows clients to privately train their local models, which is then combined in a central server to form the global model. A backdoor attack is a special type of attack where the malicious entities act as clients and implant a small trigger into the global model. Once implanted, the model performs the attacker desired task in the presence of the trigger but acts benignly otherwise. The unavailability of labeled benign and backdoored models has prevented researchers from building detection classifiers. In our work, we build a GAN-inspired defense mechanism that can detect and defend against the presence of such backdoor triggers. We tackle the data problem by utilizing the clients as Generators to construct the required dataset. Unlike traditional GAN, the goal here is to improve the performance of the Discriminator on the server-side, rather than boosting the Generator performance. We experimentally evaluated the proficiency of our approach with the image-based non-IID datasets, CIFAR10 and CelebA.**

| | |
|---|---|
| **GOAL** | **Build a binary classifier which can differentiate backdoored models from benign models.** |
| **PROBLEM** | **Lack of labeled dataset specific to the main task for both backdoored and benign models.** |
| **SOLUTION** | **Generate necessary datasets based on GAN principles during the FL process itself.** |

## METHODOLOGY

### Discriminator-focused GAN

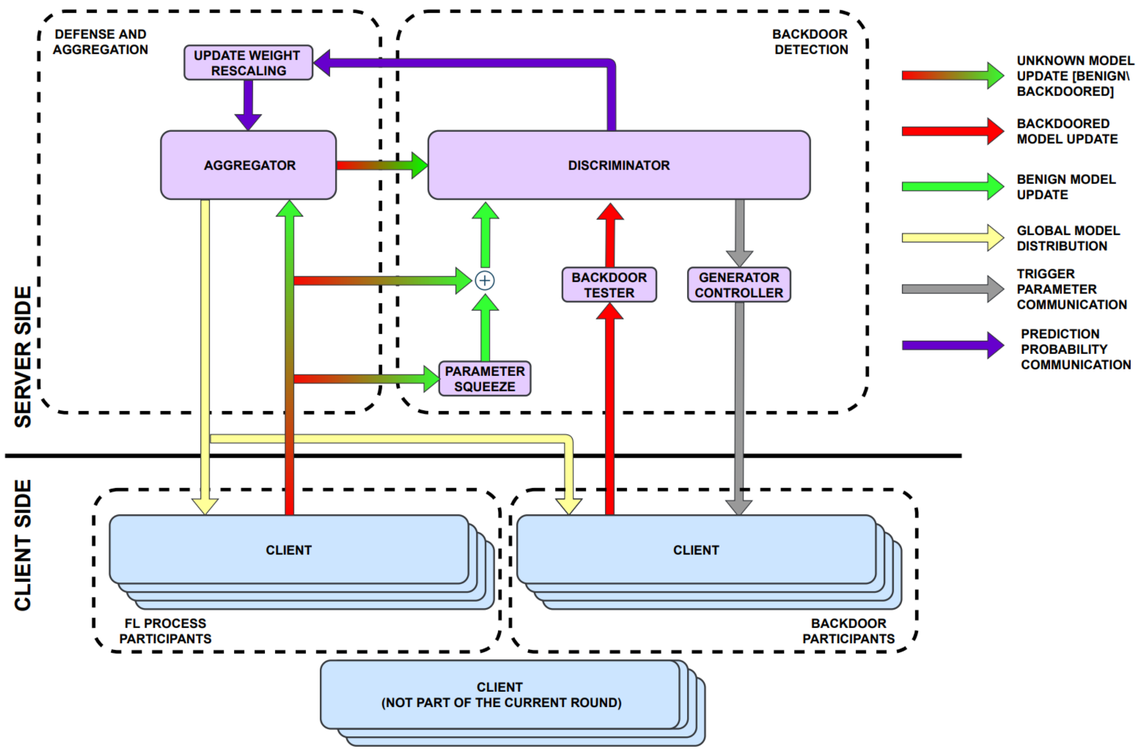
**Traditionally, the Discriminator is discarded once the generator has satisfactory performance. But, for our case, we focus on improving the performance of the Discriminator, which is the necessary binary classifier in our goal. Furthermore, we ensure that the performance of the Generator does not improve of learn additional information about the FL process.**

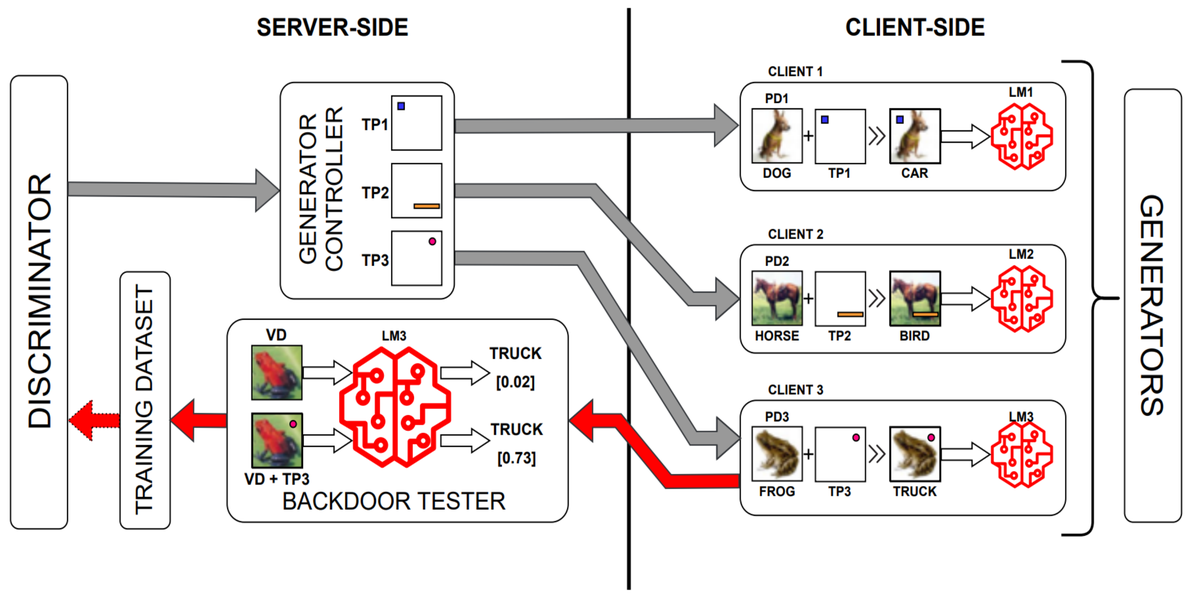
#### Generator [Clients]

- **The benign models are trained by following the regular Federated Learning Protocol.**
- **The malicious models are trained using backdoored images as per the parameter requirements of the server.**

#### Discriminator [Server]

- **The Discriminator is trained as a binary classifier, using the benign and malicious model datasets.**
- **The Generator Controller determines hyperparameter such as trigger size, position and shape for clients.**




### Circumventing Attacker Deception

- **Attacker will try to deceive the training process by sending the wrong model.**
- **The influence of the attack can be subdued by preprocessing before training.**

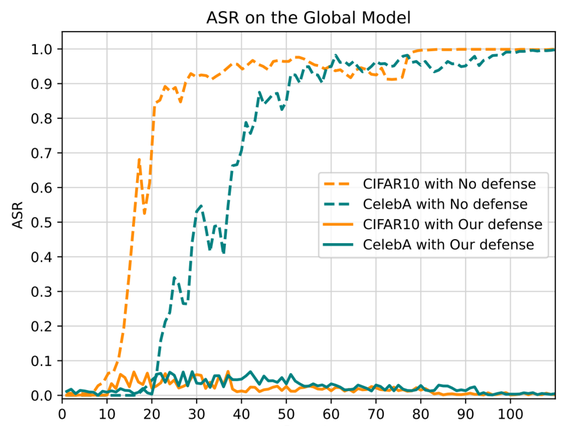
#### Benign Model Preprocessing

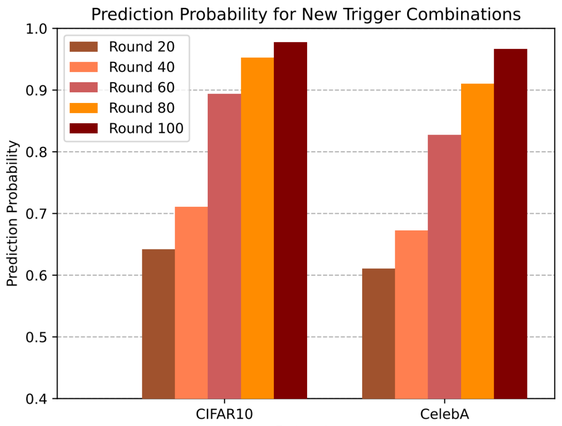
**The attacker can send backdoored model instead of benign models to corrupt the training dataset. To avoid this, we clip and squeeze 50% of the model updates before using it for training. Clipping has proven successful in mitigating the impact of the backdoor.**

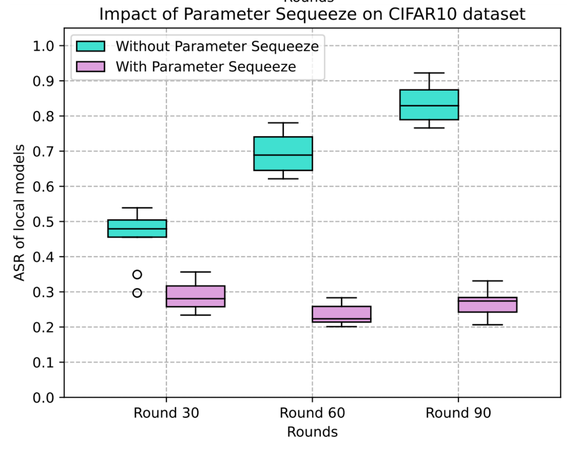
#### Backdoor Model Preprocessing

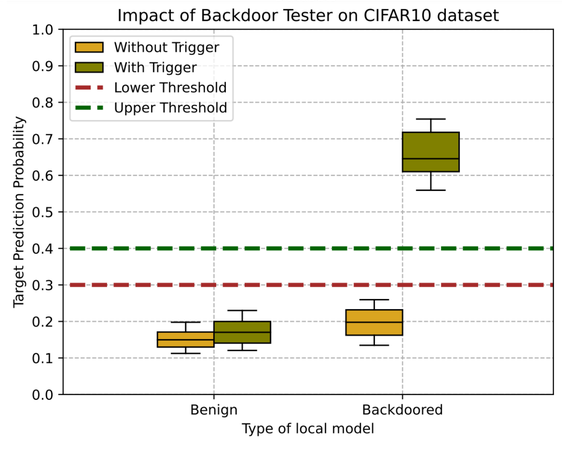
**Similarly, the attacker can send benign model updates instead of backdoored models. We use a backdoor tester to verify whether the model has actually been trained using the appropriate backdoor trigger with a validation dataset. If not, then it is discarded.**

## RESULTS




PURDUE UNIVERSITY Discovery Park

CERIAS