Poster: A Good Representation Helps the Robustness of Federated Learning against Backdoor Attack

Kyung Ho Park[§]
SOCAR
kp@socar.kr

Sanghoon Jeon[§]
Samsung Electronics
sh47.jeon@samsung.com

Huy Kang Kim*
Korea Unversity
cenda@korea.ac.kr

***Abstract*—While federated learning accomplished competitive performance in distributed networks, they are vulnerable to security threats, such as backdoor attacks. As a countermeasure against the backdoor attack, we propose Representation-Guided FedAvg (RG-FedAvg), a novel framework that aims to elevate the robustness of the conventional federated learning approach. First, our approach pre-trains the server model with an external dataset that shares the label space with distributed networks. Then, we designed each client to perform *sample-wise filtering*, which aims to filter out samples suspicious of manipulation. Throughout the proof-of-concept level experiments, we discovered the proposed RG-FedAvg effectively enhances the robustness of federated learning. Upon the illustrated improvement avenues, we expect more robust federated learning to bring various benefits to society shortly.**

I. Introduction

Federated learning is a paradigm that trains machine learning models under the networks of distributed clients (devices) without sending the raw data to the server. While FedAvg [4] and its derived studies accomplished promising performance under the distributed circumstance, federated learning is known to be vulnerable to several threats, especially a backdoor attack. As shown in Fig. 1., malicious attacker who performs a backdoor attack manipulates the training sample with a triggering pattern to induce the machine learning model trained on the manipulated training set to perform a targeted incorrect prediction on the unseen validation sample, which bears the same trigger pattern. The nature of distributed learning (i.e., less number of training samples at each client, inherently homogeneous data distribution among clients) exposes it to the backdoor attack. Towards the robustness of federated learning, prior studies have proposed defensive approaches following two paradigms: *1) robust aggregation* [1] and *2) trustworthy client selection* [3]. The robust aggregation methods aim to robustly average local updates from each client at the central server, while the trustworthy client selection methods accept the local updates only if the client is not suspicious of adversarial manipulation.

While the aforementioned paradigms presumed the use of data within the distributed network, our study aims to cast a novel scenario on the robustness of federated learning where the central server can utilize a publicized dataset that shares the same label space with the clients' data. Many academia, government, and industry entities have started to freely distribute publicly-available datasets for machine learning practitioners due to the recent interest and consciousness in public datasets for society. Along with the trend of publicizing datasets, we cast a grand question for robustness in federated learning: 'Suppose the central server can access a publicized dataset that shares the same label space with clients' data. Can we leverage it to enhance the robustness of federated learning?'

[§]denotes equal contribution, and * denotes correspondence

(a) Label: *Sandal* (Benign) (b) Label: *Sneaker* (Backdoored)

Fig. 1. Examples of the benign sample (left) and backdoored sample (right). While the benign sample is correctly labeled without any manipulation, the backdoored sample has a cross-shape of trigger pattern and the mislabeled.

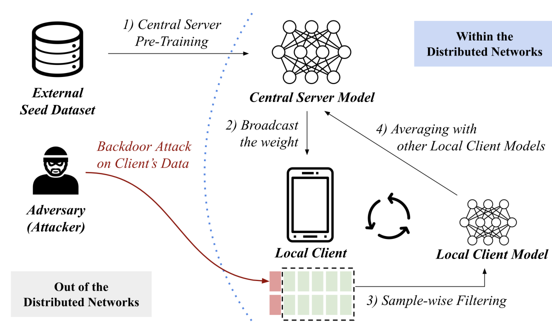


Fig. 2. Overview of the Representation-Guided FedAvg (RG-FedAvg)

II. Proposed Methodology

As a proof-of-concept level answer to the aforementioned question, we propose Representation-Guided FedAvg (RG-FedAvg), a simple but effective approach to escalating the robustness of the federated learning framework against the backdoor attack. Given the circumstance where the central server can utilize the publicly available external dataset, RG-FedAvg aims to eliminate local samples which are suspicious of backdoor attacks. As described in Fig. 2., the proposed RG-FedAvg consists of two stages: *1) Server pre-training* 2) *Sample-wise filtering*. First, the Sever pre-training stage acquires a seed dataset (which is the aforementioned external dataset), pre-trains the central server model, and broadcasts the pre-trained weight to the distributed clients. Note that we assume the seed dataset is fully sanitized without any adversarial samples, distributed in an *iid*-fashion, and shares the same label space with clients' data. The Server pre-training stage aims to establish a baseline representation power that understands the discriminative characteristics of each class. Second, during the Sample-wise filtering stage, each client drops samples suspicious of backdoor attack given the pre-trained model. As the first step of Sample-wise filtering, a client provides its samples to the pre-trained model and acquires the prediction results and their confidence. Suppose a particular sample's label differs from the predicted label with high confidence; we interpret the sample as the one suspicious of corruption. As the pre-trained model bears a particular amount of correct representation power regarding the label space, we expect the sample's local label to be inconsistent with the pre-trained model's confidently-predicted label to be reasonable evidence of the backdoor attack.

III. Experiments

A. Federated Learning Scenario

We set a federated learning scenario with Fashion-MNIST dataset [5], which has been widely utilized in prior studies regarding the backdoor attack. The Fashion-MNIST dataset includes 10 classes of 28x28 grayscale fashion-related images with a training set of 60,000 samples and a test set of 10,000 samples. We set a federated learning scenario with 10 clients, and the training samples are distributed *iid*-fashion among the client; thus, each client has 600 samples per class and the total number of training samples at a single client is 6,000. We employed FedAvg as a federated learning framework, utilized simple convolutional neural networks consisting of three convolution layers as a model, and set the objective of federated learning as a 10-class classification.

We presumed a backdoor attack scenario where an adversary manipulated sandal image with a cross pattern as a trigger (as shown in Fig. 1.) and changed the manipulated samples' label as *Sneaker*; thus, an adversary intends the model to misclassify *Sandal* images as *Sneaker*. Among 10 clients in the distributed networks, we presumed the backdoor attack manipulates 4 clients. Given no defensive method, we empirically examined an Attack Success Rate (ASR) of the aforementioned backdoor attack on FedAvg is 99.23%, which implies the backdoor attack successfully induced a faulty prediction of *Sneaker* samples as a *Sandal*.

B. Implementation of RG-FedAvg

Our study aims to examine how RG-FedAvg can effectively identify backdoored samples in the distributed networks. Note that we focused on examining whether the RG-FedAvg can identify backdoored samples in the distributed networks or not as our study intended to examine concept-level feasibility. To implement RG-FedAvg, we utilized a test set of the Fashion-MNIST (consists of 6,000 samples, nearly 10% of the number of samples at the distributed networks) as an external seed dataset for the 1) Server Pre-Training stage, and the pre-training accuracy was 98.74% on the seed dataset (pre-training performance). We broadcast the pre-trained weight to each client and let the client inspect every local sample. Following the 2) Sample-wise filtering stage, we eliminated a sample where its given label differs from the label predicted by the pre-trained model followed by high confidence. We measured a prediction's confidence with a Maximum Softmax Probability [2] and identified a prediction as a confident one of its confidence goes larger than 0.6, which is an empirically chosen threshold. Following the aforementioned setups, we could have justified that the RG-FedAvg effectively recognizes backdoored samples in the distributed networks with four evaluation metrics (Accuracy, Precision, Recall, and F1 Score) and the confusion matrix shown in Table I.

TABLE I. RG-FedAvg's backdoored sample detection performance

Backdoored Sample Detection Performance			
Accuracy	**Precision**	**Recall**	**F1 Score**
96.82%	55.69%	99.75%	71.47%

Confusion Matrix		Ground Truth	
		Backdoored	**Benign**
Prediction	**Backdoored**	2,394	1,905
	Benign	6	55,695

IV. Discussions and Conclusions

Compared to previously-proposed defense methods, our approach performs a sample-wise elimination in the distributed networks (while the trustworthy client selection methods do client-wise elimination) to minimize the number of both backdoored samples and falsely-eliminated benign samples in the distributed networks; however, this study still has avenues for future improvement. As a closing remark, we propose several improvement avenues to justify the effectiveness of RG-FedAvg for a robust federated learning framework.

- How can we perform more precise experiments along with various evaluation metrics such as ASR and the Central server model's accuracy?
- What's the effectiveness of RG-FedAvg under the various datasets, data distributions within the distributed networks (iid-fashion and *non-iid*-fashion), or the number of clients?
- How can be the RG-FedAvg more effective compared to the previous state-of-the-art approaches?
- What if the seed dataset does not share the label space with data in the distributed networks?

References

[1] S. Fu, C. Xie, B. Li, and Q. Chen, "Attack-resistant federated learning with residual-based reweighting," *arXiv preprint arXiv:1912.11464*, 2019.

[2] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," *arXiv preprint arXiv:1610.02136*, 2016.

[3] S. Li, E. Ngai, F. Ye, and T. Voigt, "Auto-weighted robust federated learning with corrupted data sources," *arXiv preprint arXiv:2101.05880*, 2021.

[4] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[5] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

Poster: A Good Representation Helps the Robustness of Federated Learning against Backdoor Attack

Kyung Ho Park*[1], Sanghoon Jeon*[2], and Huy Kang Kim [3]

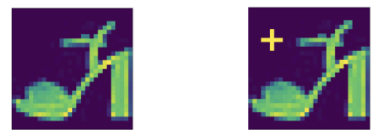
1) SOCAR 2) Samsung Electronics 3) Korea University

Introduction

- Although FedAvg and its derived studies accomplished promising performance under the distributed circumstance, federated learning is known to be vulnerable to several threats
- Prior studies have proposed defensive approaches following two paradigms below
  a) Robust Aggregation
  : It aims to robustly average local updates from each client at the central server
  b) Trustworthy Client Selection
  : It only accept the local updates only if the client is not suspicious of adversarial manipulation
- <u>While the aforementioned defensive paradigms presumed the use of data within the distributed network only, our study aims to cast a novel scenario on the robustness of federated learning where the central server can access a publicized dataset that shares the same label space with the clients' data</u>
- Recently, many academia, government, and industry entities have started to freely distribute publicly-available datasets for machine learning practitioners along with the increased consciousness in public datasets for society
- Along with the trend of publicizing datasets, our study casts a grand question for robustness in federated learning: "<u>Suppose the central server can access a publicized dataset that shares the same label space with clients's data. Can we leverage it to enhance the robustness of federated learning?</u>"

Threat: Backdoor Attack

- Malicious attacker who performs a backdoor attack manipulates the training sample with a triggering pattern to induce the machine learning model trained on the manipulated training set to perform a targeted incorrect prediction on the unseen validation sample, which bears the same triggering pattern
- Refer the examples of the benign sample (left) and backdoor sample (right) at the Fashion-MNIST dataset
  - A benign sample is correctly labeled as Sandal without any manipulations
  - A backdoor sample has a cross-shape of trigger pattern and is mislabeled as Sneaker (while the original label is Sandal)



(a) Label: *Sandal* (Benign) (b) Label: *Sneaker* (Backdoored)

Fig. 1. Examples of the benign sample (left) and backdoored sample (right).

Proposed Methodology: Representation-Guided FedAvg

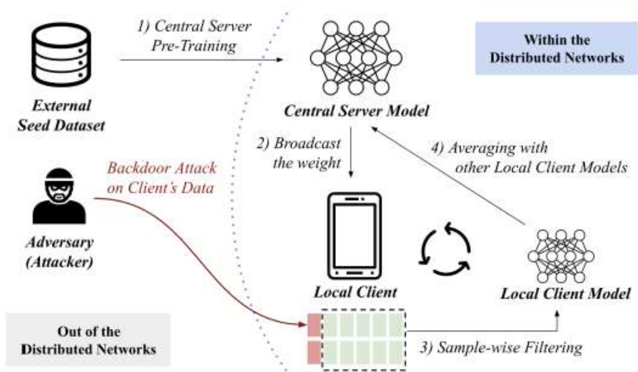


Fig. 2. Overview of the Representation-Guided FedAvg (RG-FedAvg)

- Our study proposes Representation-Guided FedAvg (RG-FedAvg), a simple but effective approach to escalating the robustness of federated learning framework against the backdoor attack
- <u>RG-FedAvg aims to eliminate local samples which are suspicious of backdoor attacks</u>
- Given the circumstance where the central server can utilize the publicly available external dataset, RG-FedAvg consists of two stages: a) Server pre-training and b) Sample-wise filtering
  a) Server pre-training 〈Refer the 1) of the Fig 1.〉
  - The server pre-training stage acquires a seed dataset, pre-trains the central server model, and broadcasts the pre-trained weight to the distributed clients

  b) Sample-wise filtering 〈Refer the 2) - 4) of the Fig. 1.〉
  - During the sample-wise filtering stage, each client drops samples suspicious of backdoor attack given the pre-trained model
  - If a particular sample's label differs from the predicted label with high confidence, we interpret the sample as the one suspicious of backdoor attack.
  - As the pre-trained model bears a particular amount of correct representation power regarding the label space, we expect the sample's local label to be inconsistent with the pre-trained model's confidently-predicted label to be reasonable evidence of the backdoor attack

Experiment Setups

a) <u>Federated Learning Scenario</u>
- Dataset: Fashion-MNIST (10-class, 60,000 Training samples, 10,000 Test samples)
- Distributed Networks Setting
  - Number of Clients and Data Distribution: 10 clients under iid setting
  - Model: Convolutional Neural Networks consisting of three convolution layers

b) <u>Backdoor Scenario</u>
- Among 10 clients in the distributed networks, we assumed 4 clients are manipulated by the attacker with backdoor attack
- Given no defensive method, we empirically examined an Attack Success Rate (ASR) of the aforementioned backdoor attack on FedAvg is 99.43%, which implies the backdoor attack successfully induced a faulty prediction of Sneaker samples as Sandal

c) <u>Implementation of RG-FedAvg</u>
- Our study aims to examine how RG-FedAvg can effectively identify backdoor samples in the distribution networks (a kind of backdoored sample detection)
- We sampled 6,000 samples from the test set of Fashion-MNIST as an external seed dataset to pre-train the central server
- During the a) Server pre-training, the pre-training performance Accuracy of 98.74%
- During the b) Sample-wise filtering, we measured the prediction's confidence with a Maximum Softmax Probability, and identified a prediction as a confident one if it goes larger than 0.6

Results and Improvement Avenues

a) Experiment Results (Backdoored Sample Detection Performance)

Backdoored Sample Detection Performance			
Accuracy	Precision	Recall	F1 Score
96.82%	55.69%	99.75%	71.47%

Confusion Matrix		Ground Truth	
		Backdoored	Benign
Prediction	Backdoored	2,394	1,905
	Benign	6	55,695

b) Improvement Avenues
- How can we perform more precise experiments along with various evaluation metrics such as ASR and the Central Server model's accuracy?
- How can we configure more diverse problem settings? What's the effectiveness of RG-FedAvg under the various datasets, data distributions within the distributed networks (iid, non-iid), or the number of clients?
- How can be the RG-FedAvg more effective compared to the previously-proposed state-of-the-art approaches?

Contributions

- Our study proposed a novel paradigm federated learning that leverages an external seed dataset to escalate the robustness against the backdoor attack
- We designed a RG-FedAvg, a simple but effective method that eliminates local samples suspicious of backdoor attacks utilizing the pre-trained model.
- We highly expect more in-depth analysis and improvement on our work can realize the benefits of federated learning into the real world shortly.