Towards Mitigation of Edge-Case Backdoor Attacks in Federated Learning

Fatima Elhattab

Towards Mitigation of Edge-Case Backdoor Attacks in Federated Learning

Fatima ELHATTAB
INSA Lyon, France
Ph.D. from Nov. 2020 to Oct. 2023
fatima.elhattab@insa-lyon.fr

Abstract

Federated Learning (FL) allows many data owners to train a joint model without sharing their training data. However, FL is vulnerable to poisoning attacks where malicious workers attempt to inject a backdoor task in the model at training time, along with the main task that the model was initially trained for. Recent works show that FL is particularly sensitive to edge-case backdoors that are introduced by data points having unusual out-of-distribution features. Such attacks are among the most difficult to counter in today's FL robust systems.

In this paper, we first implement two poisoning attacks and show that state-of-the-art robust FL systems, that are meant to counter malicious behavior, are actually vulnerable to this type of attacks. Then, we propose a defense mechanism called *ARMOR* that uses Generative Adversarial Networks to uncover edge-case backdoor attacks. Instead of monitoring the statistical shapes of users' model updates as in most of existing defense mechanisms, *ARMOR* extracts data features from the model updates in order to identify the backdoor patterns. In addition, *ARMOR* is the first FL defense mechanism against targeted poisoning attacks that is compatible with secure aggregation, thus providing better privacy than its competitors. Our extensive experimental evaluations with different datasets and neural network models show that *ARMOR* is able to counter edge-case backdoors, and outperforms existing robust FL systems by +48% to +100% in terms of resilience to attacks, while providing equivalent model quality.

ACM Reference Format:
Fatima ELHATTAB. 2022. Towards Mitigation of Edge-Case Backdoor Attacks in Federated Learning. In *Proceedings of ACM Conference (Conference'22).* Rennes, France, 3 pages.

1 Introduction

Federated Learning (FL) is a promising paradigm that is gaining grip in the context of privacy-preserving Machine Learning (ML). Thanks to FL, several data owners called *clients* (*e.g.*, mobiles devices in cross-device FL, or organizations in cross-silo FL) can collaboratively train a model on their private decentralized data, without having to send their raw data to external service providers. To this end, clients iteratively update a global model using their local training data, and send only their model updates to a central party called the *server* that orchestrates the training process. The FL server aggregates the received model updates to produce a new version of the global model, which is, in turn, distributed to the clients. FL was rapidly adopted in several thriving application domains such as next-word prediction [16], healthcare [11], banking [8], and many more.

Threat of Edge-Case Backdoors. Although FL has improved the privacy of machine learning by decentralizing the data and the learning process, many recent works have demonstrated that FL systems are highly vulnerable to various kinds of poisoning attacks, where *malicious clients* contribute poisoned model updates to poison the global model [1, 3, 7, 12, 15]. We specifically put our focus on *edge-case backdoor attacks*, as we argue that they are among the most difficult poisoning attacks to tackle [15].

Edge-case backdoors are introduced by altering label data points that, while usually correctly classified by the model, are under-represented, or unlikely to be part of the regular training or test data. Moreover, these attacks often use techniques such as the Projected Gradient Descent (PGD) to project the malicious model updates' vector so that it looks similar to a benign model updates' vector [15].

Our Contributions. In this work, we propose *ARMOR*, a FL defense mechanism that handles edge-case bakdoors. Specifically, we make the following key contributions:

- We address one of the most powerful poisoning attacks in FL, namely edge-case backdoors followed by model updates projection of the global FL model using PGD. This attack evades the strongest defense mechanisms.
- We design and implement *ARgan*, a new Generative Adversarial Network (GAN) architecture that can generate class representatives of the FL global model without having access to private real training data samples that are hidden from the FL server.
- We design and implement *MORpheus*, a poisoning mitigation mechanism which relies on the synthetically

Conference'22, April 2022, Rennes, France

generated data samples of class representatives as a test set, to uncover edge-case backdoors that are potentially introduced in the FL global model.

- By assembling *ARgan* and *MORpheus*, we propose *ARMOR* a framework for robust FL. As far as we know, *ARMOR* is the first FL defense mechanism to protect against edge-case poisoning attacks. It is also compatible with secure aggregation, thus, providing better privacy than its competitors since no client reveals its plaintext local model updates to the FL server.
- We compare *ARMOR* against five existing FL defenses, using two widely used datasets and neural networks. Our evaluation shows that, contrary to existing FL defenses, *ARMOR* can counter edge-case backdoors, with 95% resilience to attacks, and without hurting the model quality.

2 Related Work

Existing FL defense mechanisms, such as Multi-Krum [4], Trimmed Mean [17], NDC [14], FLTrust [6] and DLMP[7], fail to fully counter edge-case backdoors. Most of these mechanisms aim to detect malicious model updates as client updates that are statistically different from the majority [4, 10, 14, 17]. Or they assume the existence of a server-side test set to measure fluctuations in the model predictions' quality, to detect potential attacks [6, 7]. Although these defenses are robust to other types of attacks such as untargeted poisoning attacks [13], as pointed out in a recent study, they are not resilient to edge-case backdoors [15].

Furthermore, in order to protect against honest-but-curious FL servers, the basic FL protocol relies on secure aggregation which, roughly speaking, allows the server to access the sum of clients' model updates to be aggregated, without allowing the server to inspect each individual client update [5]. However, many existing FL defense mechanisms require analyzing individual model updates [2, 4, 6, 17], and are therefore incompatible with secure aggregation, which makes them more vulnerable to privacy leakage [9].

3 Overview of the Proposed Approach

We propose *ARMOR*, a novel FL defense mechanism that counters powerful edge-case backdoor attacks without breaking secure aggregation guarantees, nor having access to private real data samples to carry model inspection.

The overall architecture of *ARMOR* is described in Figure 1, with two main components: *ARgan* and *MORpheus*. *ARgan* is used to generate a synthetic dataset based on model updates, which is further leveraged by *MORpheus* to provide proper mitigation against poisoning attacks.

ARMOR defense mechanism does not make any assumptions neither on the proportion of attackers in the system nor on their data distribution. The insight behind *ARMOR* is as follows. Let B be a backdoor task that aims to misclassify the data samples holding a particular data pattern P^* from a

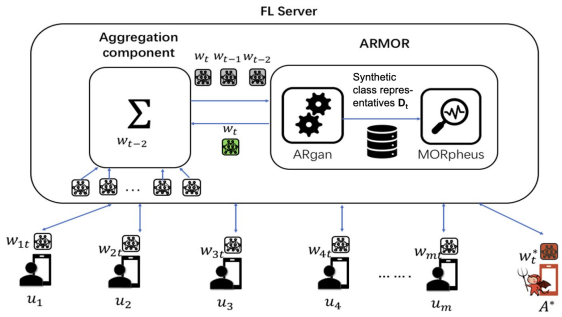


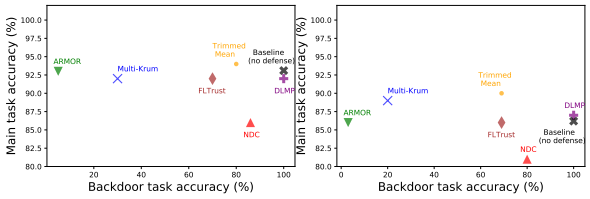
Figure 1. *ARMOR* architecture

source class C_{source} to a target class C_{target}. Let us consider that the the model w_t (at round t) is poisoned with such a backdoor B. The poisoned class representatives of C_{target} generated from w_t would be misclassified by previous non-poisoned model (*e.g.*, w_{t-1}). Here, when auditing a model w_t, *ARMOR* monitors the difference between the loss obtained when feeding class representatives to this model w_t and the loss when feeding the representatives to the models of the s previous rounds $\{w_{t-1}, ..., w_{t-s}\}$. If the difference is higher than a given threshold, the current model is considered to be corrupted, and *ARMOR* applies a mitigation technique to reduce the impact of the new model updates. Roughly spealing, the malicious updates are multiplied by a scaling factor θ, in order to reduce the impact of these updates on the global model. Here, θ is inversely proportional to the loss, *i.e.*, the higher the loss is, the lower the scaling factor is and, thus, the lower the impact on the global model is.

4 Preliminary Evaluation Results

We evaluate the effectiveness of *ARMOR* to counter edge-case backdoors, and compare it to various existing FL defense mechanisms. The resilience of a FL system to backdoors may come at the expense of a lower utility, *i.e.*, a lower model quality. In Figure 2, we present the trade-off between utility (*i.e.*, the model's main task accuracy) and resilience to edge-case backdoor attacks. We evaluate the FL systems with an attack occurring every round on FashionMNIST and CIFAR datasets. DLMP and FLTrust, which fall into the category of FL defense mechanisms that assume the existence of a validation set, have a main task accuracy which is close to the one of the FL baseline system where no defense mechanism is used. Indeed, such systems do not modify the aggregation applied by the FL server and, thus, provide a good model utility compared to the baseline. However, these defense mechanisms are not resilient to edge-case backdoors. In contrast, defense mechanisms that are based on specific aggregation approaches, such as Multi-Krum, Trimmed Mean and NDC, induce a much higher difference in model utility compared to the FL baseline system, ranging from 7.1% to +2.8%. Indeed, aggregation-based approaches such as NDC, Multi-Krum

and Trimmed Mean, may impact model accuracy, although they do not sufficiently mitigate attacks. In comparison, with *ARMOR* the backdoor task accuracy does not exceed 5% without hurting the main task accuracy, thus, providing the best trade-off between resilience and utility.



(a) FashionMNIST (b) CIFAR

Figure 2. Trade-off between resilience and utility

5 Related Work

Existing FL defense mechanisms, such as Multi-Krum [4], Trimmed Mean [17], NDC [14], FLTrust [6] and DLMP[7], fail to fully counter edge-case backdoors. Most of these mechanisms aim to detect malicious model updates as client updates that are statistically different from the majority [4, 10, 14, 17]. Or they assume the existence of a server-side test set to measure fluctuations in the model predictions' quality, to detect potential attacks [6, 7]. Although these defenses are robust to other types of attacks such as untargeted poisoning attacks [13], as pointed out in a recent study, they are not resilient to edge-case backdoors [15].

Furthermore, in order to protect against honest-but-curious FL servers, the basic FL protocol relies on secure aggregation which, roughly speaking, allows the server to access the sum of clients' model updates to be aggregated, without allowing the server to inspect each individual client update [5]. However, many existing FL defense mechanisms require analyzing individual model updates [2, 4, 6, 17], and are therefore incompatible with secure aggregation, which makes them more vulnerable to privacy leakage [9].

6 Conclusion and Ongoing Work

ARMOR is a novel Federated Learning defense method against edge-case backdoor attacks. The key difference between *ARMOR* and existing FL defenses is twofold: (i) *ARMOR* is the first FL poisoning defense method that counter edge-case backdoor attacks; (ii) contrary to most of existing FL defenses, *ARMOR* is compatible with secure aggregation, thus, providing better privacy protection. This work opens an interesting research perspective to further study the trade-off between robustness and privacy in federated learning, and how to handle these two antagonistic aspects in a consistent way.

Acknowledgments

This work is done under the supervision of Sara Bouchenak.

References

[1] Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., and Shmatikov, V. How to Backdoor Federated Learning. In *International Conference on Artificial Intelligence and Statistics* (Virtual Conference, Aug. 2020).

[2] Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., and Shmatikov, V. How to Backdoor Federated Learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS 2020)* (Virt. Conference, July 2020).

[3] Bhagoji, A. N., Chakraborty, S., Mittal, P., and Calo, S. Analyzing Federated Learning Through an Adversarial Lens. In *Int. Conference on Machine Learning (ICML 2019)* (Long Beach, CA, USA, June 2019).

[4] Blanchard, P., El Mhamdi, E. M., Guerraoui, R., and Stainer, J. Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent. In *Advances in Neural Information Processing Systems* (Mar. 2017).

[5] Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., and Seth, K. Practical Secure Aggregation for Privacy-Preserving Machine Learning. In *ACM SIGSAC Conference on Computer and Communications Security* (Dallas, TX, USA, 2017).

[6] Cao, X., Fang, M., Liu, J., and Gong, N. FLTrust: Byzantine-robust Federated Learning via Trust Bootstrapping. In *Network and Distributed Systems Security Symposium (NDSS 2021)* (Virt. Conference, Feb. 2021).

[7] Fang, M., Cao, X., Jia, J., and Gong, N. Local Model Poisoning Attacks to Byzantine-Robust Federated Learning. In *29th USENIX Security Symposium (USENIX Security 2020)* (Virtual Conference, Aug. 2020).

[8] Long, G., Tan, Y., Jiang, J., and Zhang, C. Federated Learning for Open Banking. *arXiv:2108.10749* (Aug. 2021).

[9] Melis, L., Song, C., De Cristofaro, E., and Shmatikov, V. Exploiting Unintended Feature Leakage in Collaborative Learning. *2019 IEEE Symposium on Security and Privacy (SP 2019)* (May 2019).

[10] Pillutla, K., Kakade, S. M., and Harchaoui, Z. Robust Aggregation for Federated Learning. arXiv:1912.13445.

[11] Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., et al. The Future of Digital Health with Federated Learning. *NPJ Digital Medicine 3*, 1 (Mar. 2020).

[12] Shejwalkar, V., and Houmansadr, A. Manipulating the Byzantine: Optimizing Model Poisoning Attacks and Defenses for Federated Learning. In *Network and Distributed Systems Security Symposium (NDSS 2021)* (Virtual Conference, Feb. 2021).

[13] Shejwalkar, V., Houmansadr, A., Kairouz, P., and Ramage, D. Back to the Drawing Board: A Critical Evaluation of Poisoning Attacks on Federated Learning. In *2022 IEEE Symposium on Security and Privacy (SP)* (San Francisco, CA, USA, May 2022).

[14] Sun, Z., Kairouz, P., Suresh, A. T., and McMahan, H. B. Can You Really Backdoor Federated Learning? *arXiv:1911.07963* (Dec. 2019).

[15] Wang, H., Sreenivasan, K., Rajput, S., Vishwakarma, H., Agarwal, S., yong Sohn, J., Lee, K., and Papailiopoulos, D. S. Attack of the Tails: Yes, You Really Can Backdoor Federated Learning. In *Conference on Neural Information Processing Systems 2020,NeurIPS 2020* (Virtual Conference, Dec. 2020).

[16] Yang, T., Andrew, G., Eichner, H., Sun, H., Li, W., Kong, N., Ramage, D., and Beaufays, F. Applied Federated Learning: Improving Google Keyboard Query Suggestions. *arXiv:1812.02903* (Dec. 2018).

[17] Yin, D., Chen, Y., Kannan, R., and Bartlett, P. Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates. In *35th International Conference on Machine Learning, ICML 2018* (Stockholm, Sweden, July 2018).