Poisoning Attacks in Federated Learning: An Evaluation on Traffic Sign Classification

Florian Nuding
florian.nuding@tuwien.ac.at
Vienna University of Technology, Austria

Rudolf Mayer
rmayer@sba-research.org
SBA Research gGmbH, Vienna, Austria

ABSTRACT

Federated Learning has recently gained attraction as a means to analyze data without having to centralize it from initially distributed data sources. Generally, this is achieved by only exchanging and aggregating the parameters of the locally learned models. This enables better handling of sensitive data, e.g. of individuals, or business related content. Applications can further benefit from the distributed nature of the learning by using multiple computer resources, and eliminating network communication overhead.

Adversarial Machine Learning in general deals with attacks on the learning process, and backdoor attacks are one specific attack that tries to break the integrity of a model by manipulating the behavior on certain inputs. Recent work has shown that despite the benefits of Federated Learning, the distributed setting also opens up new attack vectors for adversaries. In this paper, we thus specifically study this manipulation of the training process to embed a backdoor on the example of a dataset for traffic sign classification. Extending earlier work, we specifically include the setting of sequential learning, in additional to parallel averaging, and perform a broad analysis on a number of different settings.

CCS CONCEPTS

• **Security and privacy → Distributed systems security**; • **Computing methodologies → Supervised learning**.

KEYWORDS

Adversarial machine learning, Federated learning, Poisoning attacks

ACM Reference Format:
Florian Nuding and Rudolf Mayer. 2020. Poisoning Attacks in Federated Learning: An Evaluation on Traffic Sign Classification. In *Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy (CODASPY '20), March 16–18, 2020, New Orleans, LA, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3374664.3379534

1 INTRODUCTION AND RELATED WORK

Federated learning is the process of decentralizing the training of machine learning models. Rather than centralizing all data on a big endpoint, models are learned locally at endpoints. Federated learning is appealing specifically in settings when the data is already distributed in various computing nodes, when it was gathered or collected there. The reason for performing this type of learning can be manifold, but frequently, reasons of data confidentiality are a motivation – as data does not need to be exchanged, and the information contained in the exchanged models is already highly abstracted, it promises to solve several of the issues commonly associated with individual or otherwise sensitive data. The medical domain is e.g. a candidate for such settings, and studies have shown that e.g. the federated analysis of medical image data is comparable to centralized settings[7].

In federated settings, data can be partitioned either *vertically*, where each node gathers different features for the same observations, or *horizontally*, where each node has different observations, described by the same features[10]. In this work, we focus on the latter setting. The global model is often obtained by two different approaches. *Parallel averaging*[6] trains models at each node in parallel, before aggregation by a central coordinator. In *sequential learning*, sometimes called cyclic incremental learning, on the other hand, the model is passed from one node to the other, and the subsequent node continuous training from the state the previous node has ended training in. This is performed in a number of cycles.

Adversarial Machine Learning has recently gained attention, as a field concerned with the security of the machine learning process. Attacks can often be categorized to address the *confidentiality*, *integrity*, or *availability*, and be either performed during the training or prediction phase of the process[3]. Prominent attacks include e.g. adversarial examples, which are a prediction-time *evasion* attack, or *poisoning* attacks, which aim at creating a backdoor in that model during the training time, e.g. in deep neural networks[5]. Both attacks aim at influencing the *integrity*. Adversarial examples try to create a minute perturbation of an original, legitimate input. This is ideally not noticeable by a human observer, but tricks the classification model in assuming, with high confidence, a different class. Poisoning attacks inject a number of carefully modified samples to the training set. These samples contain a key that is a trigger for a specific (wrong) class to be predicted. The goal of the attacker is that the trained network memorizes these keys, and can thus be activated on demand by embedding the key in a target sample. The pattern may even be noticeable, but not suspicious – e.g. a sticker on a traffic sign, a person wearing a specific type of glasses, etc.

Recent work has studied the vulnerabilities of Federated Learning against a number of these attacks, e.g. against the model confidentiality, by trying to infer information about the samples used for training the models[8]. Poisoning attacks have been investigated, e.g. in [1]. As federated learning is a distributed system, infiltrating a single computation node might be easier for an attacker than a centralized system, especially if some of the participating nodes are not well protected.

CODASPY '20, March 16–18, 2020, New Orleans, LA, USA

ACM ISBN 978-1-4503-7107-0/20/03…$15.00
https://doi.org/10.1145/3374664.3379534

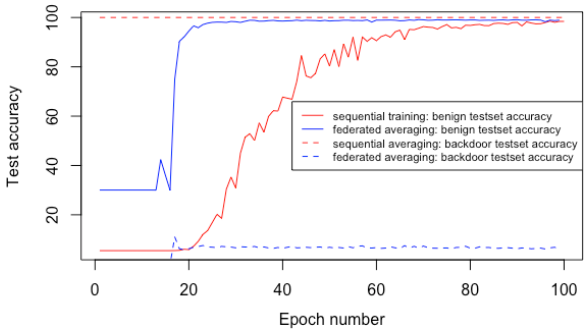


Figure 1: Accuracy comparison of sequential and parallel learning, with 20% malicious nodes

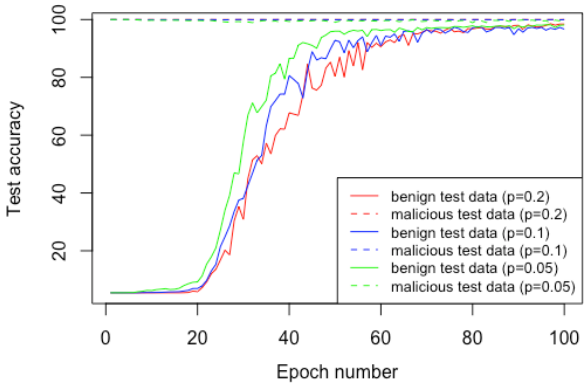


Figure 2: Effectiveness at different fractions of malicious nodes (p)

In this paper, we focus on poisoning attacks, and perform an evaluation of the possibilities to create a backdoor in federated learning settings. We utilize a dataset that has not yet been studied in adversarial, federated settings, specifically from the image categorization domain, with the task to classify traffic signs.

We contribute to current research by analyzing and comparing both parallel averaging as well as sequential federated learning. While previous work has focused on parallel averaging, sequential learning can be exploited in different ways by adversaries. Further, we extend earlier work by an in-depth evaluation on a variety of parameters that have not been considered previously.

2 DATASET AND EXPERIMENT SETUP

For our experiments, we utilize the European Traffic Signs dataset, [4], which combines some existing benchmark sets with new additions from other countries. The task is to correctly classify the traffic sign depicted in the image. As our model we chose to a convolutional neural network, based on recommendations from [4]. The hyper-parameters were chosen by running the classification task in a centralized setting, i.e. with all data in one node.

For the federated learning, we test several different settings: the *merge strategy* (parallel or sequential averaging), the *number of nodes* (from two to 20), and the *batch size* (in steps of 8, 32, 64) .

Regarding the attack, we vary the proportion of adversarial nodes, as well as, if applicable, the sequence when they are contributing to the learning (at the beginning or the end of one cycle). Further, two different strategies are utilized for influencing the final model – a basic strategy, which just returns the model learned by the malicious node, and a *model replacement* strategy[2], which aims at boosting the influence of the backdoor nodes by increasing their weights, to achieve a larger influence in the final model.

3 EVALUATION

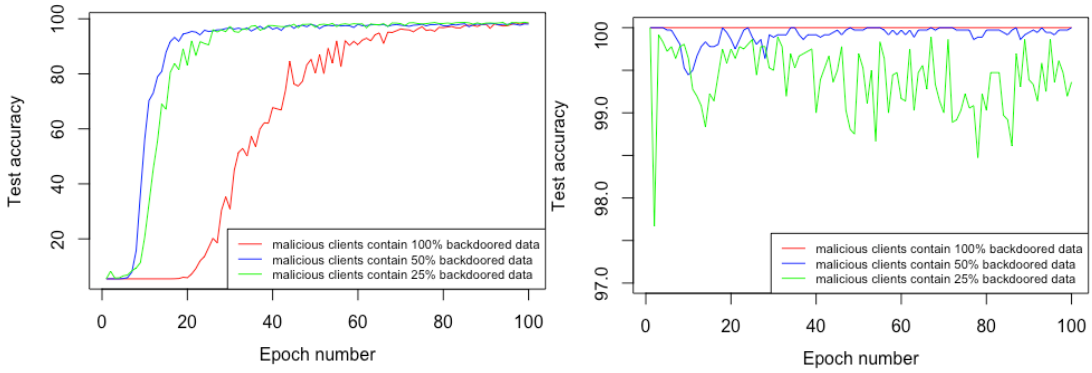
In Figure 1, the basic strategy for model learning is used both on sequential training and federated averaging. The x-axis corresponds to the number of epoch, and the y-axis corresponds to the accuracy on the test-set. The accuracy on the benign test-set reaches high results in both cases (the difference in how fast that is achieved might be a result of a bigger learning rate on federated averaging).

However, the more interesting observation is that when federated (parallel) averaging is used, the backdoor is not successfully implemented: the accuracy does not increase in this case, and is below 10% for all number of epochs. The loss on the benign set correctly decreases during training in both sequential training and federated averaging, while the loss of the malicious test-set using federated learning even increases. This inability to introduce a backdoor into the combined model on federated averaging using the basic attack strategy also corresponds to the observations made by [1]. When applying the model replacement strategy, the backdoor is more successfully embedded into the model.

For centralized learning, one parameter is the percentage of poisoned images used to install the backdoor (p%). With a still rather small percentage, a high effectiveness of the backdoor can already be achieved – up to 95% effectiveness of the backdoor with 10-20% poisoned samples[9]. We further vary the percentage of malicious nodes in the federation, with p between 5% and 20% (cf. Figure 2). All settings lead to a high accuracy of over 99% on test-data containing backdoor images, while also achieving an accuracy of over 95% on the benign test data, after 100 training epochs.

Extending the setting above, we keep the number of benign and malicious clients constant, but only change the proportion of benign and malicious samples in each malicious client. Bagdasaryan et. al. [1] do not test the influence of this relation, but keep this ratio constant (at 44 benign samples and 20 malicious samples in each batch of the malicious clients). The results are depicted in Figure 3. We can observe that while a larger setting of poisoned data delays the convergence on the benign data set, the backdoor can be very effective already with a lower number of backdoored samples.

While literature often follows an approach to first train a network with the benign samples, and then to install the backdoor by training with the poisoned samples (e.g.[5, 9]), we identify that for the *sequential* training, the order of the malicious nodes in the training sequence is of importance. The later in the training process, the fewer non-adversarial nodes have the ability to decrease the effectiveness of the backdoor by overwriting the model towards one focusing on benign images. We thus evaluate this setting, depicted in Figure 4, showing the extreme cases of being either first or last in the training process. While the adversary coming last results



(a) Performance on benign test data

(b) Backdoor performance on poisoned data

Figure 3: Effect of different ratio of benign and $p\%$ poisoned data (with a total of four benign and one malicious node)

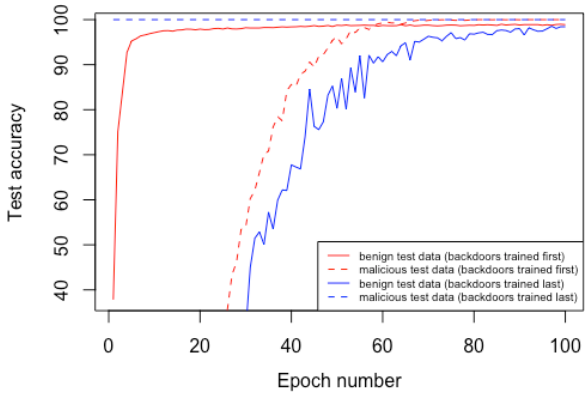


Figure 4: Performance of benign and malicious test set with different order of training on the poisoned data

in the backdoor being immediately effective, and the accuracy of the benign test set slowly increasing, until it reaches approx. 95% accuracy. For the other case, the adversary training first, the benign test set is very quickly reaching high accuracy level, but the backdoor takes a longer time to become at effective at all. Eventually, after a long enough training time, both scenarios converge to a similar state. The reason for this behavior is the training process, where especially in the first iterations of the training, when the network is far from a converged state, the most recent adaptions of the weights of the neurons have the highest influence, and lead to either the backdoor being very prominent, or mostly overwritten by the benign data.

4 CONCLUSIONS AND FUTURE WORK

In this paper, we evaluated the effectiveness of poisoning attacks in a federated learning setting, on a image categorization task. We distinguish two scenarios of federation: parallel averaging, and sequential (incremental) learning. We are able to confirm that federated settings are vulnerable to adversaries, and that it is possible to install backdoors in federated learning of a traffic sign classification. We observed differences in the effectiveness of federated averaging and sequential learning, and analyzed the effect of the order in the sequence an attacker is contributing to the learning.

It should be noted that in general, the required level of effectiveness of the backdoor depends on the attack scenario – to break a face authentication system, the attack should work with a high accuracy. In other settings, also a lower effectiveness might eventually lead to the attackers goal.

Future work will focus on confirming our results in different domains, detecting poisoning attacks in the federated setting, as well as designing mitigation and defense strategies.

ACKNOWLEDGMENTS

This work was partially funded by the Austrian Research Promotion Agency (FFG), contract No. 873979, and the EU Horizon 2020 programme under grant agreement No. 826078.

REFERENCES

[1] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. 2018. How To Backdoor Federated Learning. *CoRR*. abs/1807.00459.
[2] Arjun Nitin Bhagoji, Supriyo Chakraborty, Prateek Mittal, and Seraphin Calo. 2019. Analyzing Federated Learning through an Adversarial Lens. In *Proceedings of the 36th International Conference on Machine Learning*. Long Beach, CA, USA.
[3] Battista Biggio and Fabio Roli. 2018. Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning. *Pattern Recognition* 84 (December 2018).
[4] Citlalli Gamez Serna and Yassine Ruichek. 2018. Classification of Traffic Signs: The European Dataset. *IEEE Access* 6 (2018).
[5] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. 2017. BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain. In *Proceedings of the Machine Learning and Computer Security Workshop*. Long Beach, CA, USA.
[6] Jakub Konečný, H. Brendan McMahan, Felix X. Yu, Peter Richtarik, Ananda Theertha Suresh, and Dave Bacon. 2016. Federated Learning: Strategies for Improving Communication Efficiency. In *Workshop on Private Multi-Party Machine Learning, Conference on Neural Information Processing Systems (NIPS)*.
[7] Micah J. Sheller, G. Anthony Reina, Brandon Edwards, Jason Martin, and Spyridon Bakas. 2018. Multi-institutional Deep Learning Modeling Without Sharing Patient Data: A Feasibility Study on Brain Tumor Segmentation. In *International MICCAI Brainlesion Workshop BrainLes*. Springer, Granada, Spain.
[8] Stacey Truex, Ling Liu, Mehmet Emre Gursoy, Lei Yu, and Wenqi Wei. 2019. Demystifying Membership Inference Attacks in Machine Learning as a Service. *IEEE Transactions on Services Computing* (2019).
[9] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y. Zhao. 2019. Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks. In *IEEE Symposium on Security and Privacy (SP)*. San Francisco, CA, USA.
[10] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. Federated Machine Learning: Concept and Applications. *ACM Transactions on Intelligent Systems and Technology* 10, 2 (Jan. 2019).