Data Poisoning Attacks in Federated Learning

Rudolf Mayer

COMET Competence Centers for Excellent Technologies
www.ffg.at/comet
SBA Research

Problem & Motivation

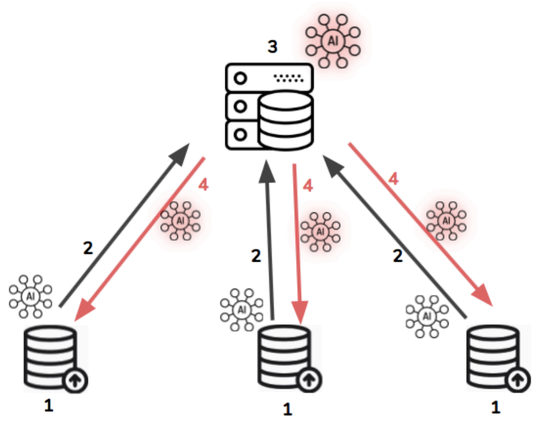
Federated Learning promises advances over centralized learning:

- No need to exchange distributed data for collaborative learning
  - Alleviating many risks and obstacles related to data privacy
- Computing resources at the data holders can be utilized, thus distributing the computation

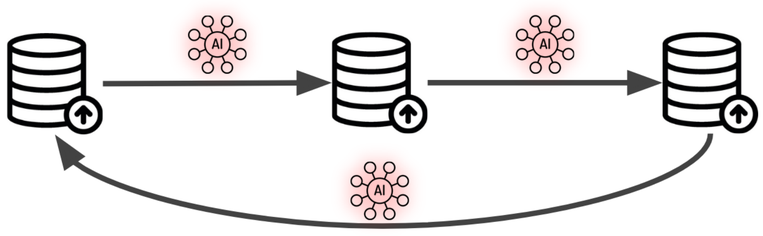
Federated Learning is however still (or even more) exposed to adversarial attacks. We evaluate to what extend an attacker can disturb the training process to successfully embed a backdoor in the common model.

Federated Learning

Rather than moving the data to the model, Federated Learning is based on the principle of creating a model where the data is generated. Two different architectures of federated machine learning can be distinguished.
Parallel Federated Learning: Each training round consists of several steps: The clients train models based on their local data (1), which are send to the aggregation server (2). There, the local models are combined (e.g. by averaging the models' parameters) (3). Finally, they are distributed back to the clients (4).



Sequential Federated Learning (aka cyclic incremental learning): A client trains its model locally, and sends it to the next client for further training. This does not require a central aggregation process.



Poisoning Attacks

Despite obvious benefits, the distributed nature of Federated Learning enables new attack vectors for adversaries. Backdoor attacks are an attack targeting the model's integrity during the training phase. According to this strategy, an adversary poisons the training data by adding samples containing a certain pattern (the so-called "backdoor"). The goal is to trigger malicious behavior on data containing this pattern during the deployment phase. *Note*: The appearance of the backdoor patterns as such is not a primary concern. While the created backdoors are noticeable, they are chosen to be unsuspicious as they naturally occur in the selected data.

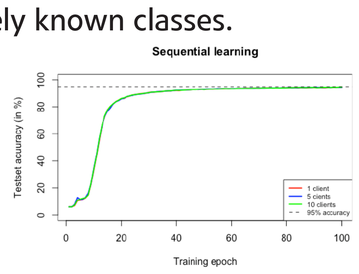
(a) Original image (Yale Face dataset) *(b)* Backdoor pattern „glasses" *(c)* Original image (Traffic Sign dataset) *(d)* Backdoor pattern „black square"

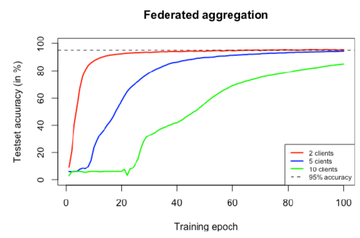
Goals

We compare federated machine learning to central machine learning on aspects such as the number of participants in the network, or the ability of handling non-independent and non-identically distributed (non-i.i.d.) data. We measure the effectiveness of different ML models with common metrics such as test set accuracy. Furthermore, we perform backdoor attacks in Federated Learning settings to gain insight into the impact on effectiveness by varying properties such as the pattern's appearance (size, shape or color), attack strategies, or varying numbers of attackers.

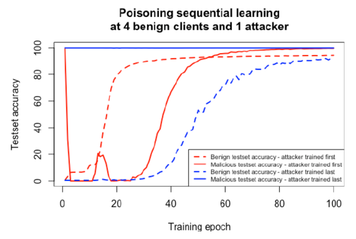
Results

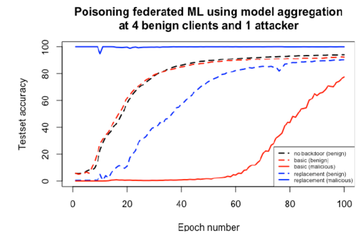
Tested on i.i.d. data, the number of clients has no influence on the effectiveness of the machine learning model in the case of sequential learning. However, in the case of federated (parallel) aggregation, a higher number of clients leads to a slower convergence of the global model. Sequential learning on non-i.i.d. data suffers from catastrophic forgetting, meaning that data trained in early stages is underrepresented in the resulting model. Federated averaging also suffers from reduced performance on sparsely known classes.



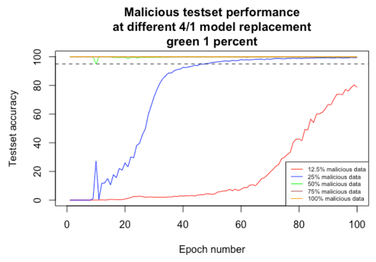


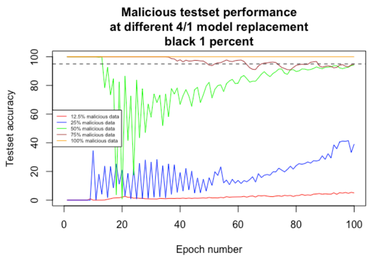
Backdoor attacks can be successfully introduced in both federated settings, shown below in networks consisting of 4 benign clients and 1 attacker. In sequential learning, the point of time the attacker participates in the learning cycle has a big impact on the performance. In a federated aggregation setting, especially the model replacement strategy [1] leads to a high effectiveness on benign and malicious test data.





We can further see that black color is less effective for the backdoor: the only backdoor attack considered as successful is when malicious client uses 50% poisoned data. If we use only 25%, the performance of the malicious test set is low, and the attack is not successful





Conclusions

We evaluated different types of Federated Machine Learning techniques regarding effectiveness on benchmark datasets for image classification. Federated ML offers advantages regarding privacy and utilisation of resources, but opens up new attack vectors for adversaries. We designed and implemented strategies for backdoor attacks and were able to confirm that Federated ML is highly susceptible to these attacks. Future work will put an emphasis especially on defence strategies.

E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov. How To Backdoor Federated Learning. In *23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, Palermo, Italy, June 2020.
F. Nuding and R. Mayer. Poisoning Attacks in Federated Learning – an Evaluation on Traffic Sign Classification. In *10th ACM Conference on Data and Application Security and Privacy (CODASPY)*, New Orleans, U.S., March 2020.

Bundesministerium Klimaschutz, Umwelt, Energie, Mobilität, Innovation und Technologie
Bundesministerium Digitalisierung und Wirtschaftsstandort
FFG Forschung wirkt.
wirtschaftsagentur wien Ein Fonds der Stadt Wien
Europäische Kommission
FWF Der Wissenschaftsfonds.
netidee OPEN INNOVATIONS

SBA Research (SBA-K1) is a COMET Centre within the framework of COMET – Competence Centers for Excellent Technologies Programme and funded by BMK, BMDW, and the federal state of Vienna. The COMET Programme is managed by FFG.