Data Poisoning attacks against Federated  
Learning Systems: A Survey

ABSTRACT

First introduced by Google in 2016, Federated Learning (FL) is an emerging machine learning system designed to maintain higher levels of data privacy through distributed model training. However just like other Machine Learning (ML) systems trained on user-provided data, these systems are vulnerable todata poisoningattacks. Data poisoning attacks manipulate the training data in order to corrupt the learned model. It still remains somewhat unclear as to how ‘toxic’ data poisoning attacks are outside of an academic setting. The aim of this paper is to firstly review the current data poisoning attacks which affect FL but also to assess the impact in which these malicious attacks pose within a realistic setting.

1. INTRODUCTION

The arrival of the big data era has shifted attention from the amount of data needed for ML applications, to the privacy and security of the data being used. Federated Learning (McMahan et al., 2017) has emerged within the last few years as a popular alternative to traditional large scale deep neural network (DNN) training. Traditional ML will ordinarily use the centralized method of data processing – centralized collection, unified processing, cleaning and modelling (C. Zhang, Y. Xie, H. Bai et al., 2021, p.2). Federated Learning utilizes distributed training among the remote devices, sharing only the model parameter updates which are then aggregated by a centralized parameter server. The server is designed without any visibility to a remote devices’ local data or training process (Bhagoji et al., 2019, p.1). This guarantees FL does not leak user privacy at the data level and therefore not in violation of GDPR and other data privacy legislation.

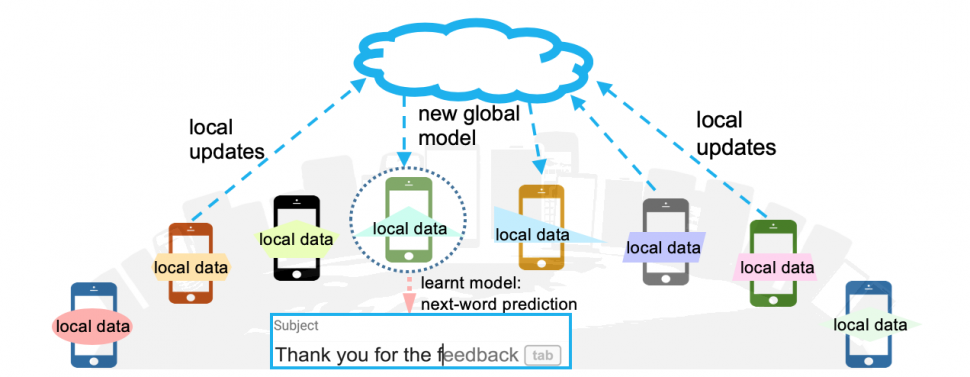
Although federated learning provides a privacy-aware paradigm for model training, works within the last few years have demonstrated that FL can still be prone to privacy leakage. It has been shown that even just communicating model updates throughout the training process can still reveal private information (Bhowmick et al., 2018). As shown by (Boenisch, 2021), even a passive attacker observing gradients can reconstruct an individual’s user data. While federated learnings privacy-preserving protection may not be but as strong as initially hoped is also prone to the threat of Data poisoning and in particular model poisoning attacks.

To explore this issue the paper is split into four sections and is structured as follows. The first section provides an overview of federated learning, both is applications and the mechanics behind it. The second section provides the same technical detailing in relation to data poisoning. The third section then surveys current data poisoning attacks that have been specifically levied against FL. In this section, the degree of potency data poisoning has on FL systems within a realistic setting is analyzed.

1. OVERVIEW OF FEDERATED LEARNING

2.1 How does FL work?

Federated Learning can be split into for main steps. The first step is client selecting/sampling. The server either randomly selects or uses an algorithm to select devices to train the model. The second step is parameter broadcasting. The server transmits the global model to the selected clients. The third step is local model training. Here the selected clients retrain the model sent to them using their nonidentically distributed local data. The final step is model aggregation. The selected clients then



send their updated model parameters back to the server where they are then aggregated towards the global model. This new global model is then sent back to another subset of selected devices. This training process continues until convergence is reached or a stopping criterion is met.

So the standard FL objective is to learn a single global statistical model from data stored on potentially millions of remote devices. The constraints in this case are that the data is not only stored and processed on the local devices but that the global model is only partially updated on a periodic basis. Although the modelling approach can differ depending on the objective or application, the general problem formulation is to minimize the following function:

In this case, $m$ = total number of devices, $p\_k$$\geq 0$ and $\displaystyle\sum$$\_k p\_k=1$, and $F\_k$ is the local objective function for the $k$th device. The local objective function is often defined as the empirical risk over local data (Li, 2020, p.50). $P\_k$ refers to each devices relative impact, with two settings being $p\_k=1/n)$ or $p\_k=(n\_k/n)$ where $n=$$\displaystyle\sum$$\_k n\_k$ refers to the total number of samples

TYPES OF FEDERATED LEARNING

Federated learning has quite a few different architectures. Based on the distributive characteristics of that data, FL can be categorized into three distinct groups: horizontal federated learning, vertical federated learning and federated transfer learning.

**Horizontal FL**

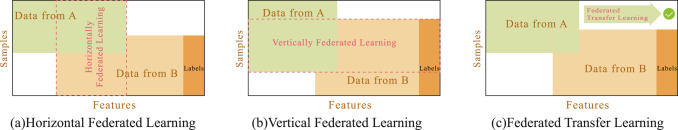
Horizontal federated learning, also referred to as sample based federated learning, is used in situations where the local data sets share the same features but are different in samples. That is to say the user features overlap a lot but the users overlap little. For example, two banks in different locations may have very different users and so there is little overlap. However, with both businesses being banks, the features spaces are very similar. The benefit of this architecture is that a global model can be trained that not only increases the total number of training samples but can also improve the accuracy of the model (Zhang, Xie, Bai et al, 2021, p.3).

**Vertical FL**

Vertical federated learning, or feature based FL, is used in instances where the data sets share the same sample ID but differ in feature space (Q. Yang et al, 2019, p.7). For example, you may have two different companies that reside in the same city. In this instance, their users are likely to overlap however the user features will be different. Under such an architecture, vertical FL could aggregate these different features to create prediction models about the users. Such research already undertaken in this field include Cooperative Statistical Analysis (Du & Atallah, 2001), classification (Du, han, chen, 2004) and association rule mining (Vaidya & Clifton, 2004).

**Federated Transfer Learning**

Federated Transfer learning applied to situations where there is neither, or very little overlap in not only samples but also in feature space. In this case a common representation is learned using the limited sample overlap and later applied to make predictions.



* 1. Applications of Federated Learning

**Healthcare**

Smart healthcare is an exciting domain which can benefit from federate learning systems. Electronic Health Records (EHR) is considered the main source of healthcare data for ML applications (Ghassemi, Naumann, Schulam, Beam et al., 2020). These EHR’s which contain disease symptoms, genome sequences, medical reports are particularly confidential and stored in individual hospitals and GP surgery’s. Training models using only limited data hinders the performance of machine learning models in healthcare settings. In this instance, federated learning could help build a collaborative model for health care data while maintaining the privacy of sensitive patient information private.

**Transportation**

With the increase in sensors in vehicular networks, federated learning methods could capitalize on the abundance of data. At the moment, ML based models have been applied to vehicle and traffic management (Tan, Bremner, Kernec, Imran, 2020). In relation to traffic management, specifically traffic flow predictions, large amounts of data is required and due to the data being stored amount various organizations, it cannot be exchanged to protect privacy (Liu, Kang, Niyato, Zhang, 2020). Federated learning could be applied to address such roadblocks..

**Finance**

Using federated learning to build a risk assessment ML model for loan risk assessment may be one of its best uses (Cheng, Liu, Chen, Yang, 2020). Banks could potentially combine not only their customer credit card reports but other factors like “taxation and reputation by collaborating with other financial institutions and e-commerce companies” (Mammen, 2021, p.3). Again, the privacy issue of sharing consumers private information could be alleviated by using federated learning techniques.

**OVERVIEW OF DATA POISONING**

Having covered the mechanics and application of federated learning, this essay now turns to its main focus of analysing the risk data and model poisoning poses to federated learning. Data poisoning is a security risk for machine learning systems as a whole. As alluded to at the beginning of the essay, data poisoning attacks attempt to “control the behaviour of a system by manipulating its training data” (Schwarzchild, p.1). In particular these deep learning systems are prone to such attacks to due the sheer volume of data needed to train the models. At the scale of data needed, it is often unfeasible to adequately scrutinize it. Understandably this has caused concern at industry level where datasets for AI systems which are not manually reviewed could potentially produce corrupted models \Jiang2017corr. A recent survey of 28 industry organizations further cements this point, with data poisoning being the primary fear out of all the adversarial machine learning threats \kumar2020adversial. General Poisoning attacks can be either a) random attacks or b) targeted attacks (Huang et al., 2011). Random attacks aims to reduce the accuracy of the federated learning model whereas targeted attacks attempt to modify the FL model to output the target label specified by the attacker (lyu, 2020). Poisoning attacks are highly relevant in a number of domains such as spam filtering (Burzstein, 2018; Nelson 2008), computer vision (Papernot, McDaniel, 2018), disease diagnosis (Mozaffari-Kermani, M., Sur-Kolay, S., Raghunathan) and recommender systems (Fang, M., Yang, G., Gong, N.Z., Liu, 2018).

However the works cited above have been conducted using traditional ML models where the training data is centrally stored and processed. Two examples can be provided to illustrate why these attacks and defences for centralized ML models are not suitable for federated learning. Poisoning attacks that rely on observing the training data distribution to create poison instances are not applicable as the malicious user can only access and modify the single device training data that they hold. In the same vein, server side defences that use anomaly detection to filter poison instances (Paudice, A., Muñoz-González, L., Gyorgy, A, 2018) are not applicable since the server can only see the parameter updates rather than the individual instances.

**HOW TOXIC ARE DATA POISONING ATTACKS ON FEDERATED LEARNING?**

Although traditional poisoning attacks may not be effective against FL systems, they can be modified to be made effective. As mentioned previously, there are two types of poisoning attacks in federated learning, data and model poisoning. In order to carry out an data poisoning attack on an FL model, the adversary will have to inject multiple poison instance or change existing instances. In this case the local learning process is not modified (Tolpegin p.496).

In model poisoning, a slightly different process occurs. The malicious participant “modifies its learning process in order to create adversarial gradients and parameter updates” (Tolpegin, p.496). Work by Bhagoji(cite) were able to use targeted and untargeted model poisoning attacks to create high model error. Amazingly in their paper, simulating an adversary “controlling a single malicious agent were able to achieve targeted misclassification at the global level with 100% confidence while ensuring convergence of the global model for deep neural networks trained on both datasets” (Bhagoji, p.635). Another example of the return strong result can be seen in (Tolpegin). In their paper they were able to successfully carry out label flipping poisoning attacks. The adversarial goal was to manipulate the learned parameters so that the final global model had high errors for particular classes. Using label flipping attack to implement targeted data poisoning they were able to oberserve substantial change in target class recall (>6% drop) tolpegin, p.488

516

**CONCLUSION**

The aim of this essay has been firstly to account for what federated learning is, the different iterations it may come in and the applications it can be utilized for. While federated learning faces many challenges relating to performance optimization, security and systems and data heterogeneity, this essay chose to focus on adversarial attacks levied against FL, specifically data and model poisoning. Federated learning systems were found to be vulnerable to label flipping poisoning attacks. It was also seen that as the proportion of adversarial participants increase, the higher the negative impact on the global model. It was also observed that targeted poisoning attacks are possible, that is to say, with enough malicious participants, adversaries are able to induce a specific desired output from the global model. While federated learning is certainly an exciting prospect, there are undeniably kinks still needing to be worked out. In order for the potential of Federated Learning to be fully realised, future research will have to be continually focused on not only performance optimization but on adversarial defences.