Liu *et al. Cybersecurity* (2022) 5:4
https://doi.org/10.1186/s42400-021-00105-6

Cybersecurity

SURVEY

Open Access

Threats, attacks and defenses to federated learning: issues, taxonomy and perspectives

Pengrui Liu[†], Xiangrui Xu[†] and Wei Wang[*]

Abstract

Empirical attacks on Federated Learning (FL) systems indicate that FL is fraught with numerous attack surfaces throughout the FL execution. These attacks can not only cause models to fail in specific tasks, but also infer private information. While previous surveys have identified the risks, listed the attack methods available in the literature or provided a basic taxonomy to classify them, they mainly focused on the risks in the training phase of FL. In this work, we survey the threats, attacks and defenses to FL throughout the whole process of FL in three phases, including *Data and Behavior Auditing Phase*, *Training Phase* and *Predicting Phase*. We further provide a comprehensive analysis of these threats, attacks and defenses, and summarize their issues and taxonomy. Our work considers security and privacy of FL based on the viewpoint of the execution process of FL. We highlight that establishing a trusted FL requires adequate measures to mitigate security and privacy threats at each phase. Finally, we discuss the limitations of current attacks and defense approaches and provide an outlook on promising future research directions in FL.

Keywords: Federated learning, Security and privacy threats, Multi-phases, Inference attacks, Poisoning attacks, Evasion attacks, Defenses, Trusted

Introduction

As smart cities grow in popularity, the amounts of multi-source heterogeneous data generated by various organizations and individuals have become increasingly diverse. However, businesses and people are hesitant to exchange data due to the concern about data privacy, leading to the emergence of *data silos*. Several attempts have been made to solve the data privacy threats, where FL has shown its superiority as it allows multiple local workers to train together without revealing sensitive information about local data (Lyu et al. 2020). In December 2018, the IEEE Standards Committee approved the standard project of *architectural framework and application of Federated machine learning*. Subsequently, more and more scholars and technical experts joined the standards working group and participated in drafting IEEE Standards about FL.

At present, FL combined with Multi-task Learning (Smith et al. 2017), Reinforcement Learning (Qi et al. 2021), Graph Neural Network (Wu et al. 2021) or other artificial intelligence algorithms have been proposed and applied in many fields. In addition, similar to FL, some collaborative learning methods like Assisted Learning (Xian et al. 2020), Split Learning (Vepakomma et al. 2018) have also been proposed. In Natural Language Processing (NLP), the application of FL is also being widely studied. Lin et al. (2021) opened up a research-oriented FedNLP framework, which aims to study privacy-preserving methods in NLP with FL. Many aggregation algorithms and open-source frameworks for FL have also been proposed (Mothukuri et al. 2021), such as FedAvg (McMahan et al. 2017), SMC-AVG (Bonawitz et al. 2016), FedProx (Li et al. 2018), and FATE, Tensorflow-Federated, PySyft etc.

*Correspondence: wangwei1@bjtu.edu.cn
[†]Pengrui Liu and Xiangrui Xu contributed equally to this work and should be considered co-first authors.
Beijing Key Laboratory of Security and Privacy in Intelligent Transportation, Beijing Jiaotong University, Beijing 100044, China

Although FL can effectively break data silos, there are many inborn security and privacy threats. Before the model is trained, malicious local workers may destroy the integrity, confidentiality, and availability of data, and thus contaminate the model. In general, the key roles of FL include two parts: central server and local workers (or local clients). The adversary can compromise the central server or a part of local workers. When the model is being trained, the adversary can manipulate the global model by controlling the samples or model updates. This will result in degraded performance of the global model, or leave a backdoor. In addition, in the model training and predicting phases, the adversary can also infer the private information of other honest local workers, including membership inference and attribute inference. Even though differential privacy and other privacy-preserving algorithms have been implemented within FL, attacks against FL can still succeed (Cheu et al. 2021).

Many existing surveys mainly focused on listing and describing various attack methods and defense strategies (Lyu et al. 2020; Mothukuri et al. 2021; Enthoven and Al-Ars 2020). However, these surveys only analyze security and privacy threats in the training phase. In this work, we analyze the security and privacy threats according to the multi-phase framework of the FL execution, including *Data and Behavior Auditing*, *Training* and *Predicting*. We identify the issues and provide a taxonomy of threats, attacks and defenses on FL. We also provide perspectives on how to build a trusted FL.

FL concepts and challenges

Definition

FL is defined as a machine learning paradigm in which multiple clients work together to train a model under the coordination of a central server, while the training data remains stored locally (Kairouz et al. 2019). According to the type of local workers, FL can be divided into cross-device and cross-silo. Cross-device workers are primarily mobile phones, tablets, speakers, and other terminal IoT devices. These local workers may disconnect at any time in the process of model training. The workers of cross-silo are mainly large institutions that have high data storage and computing capabilities. In the fully decentralized setting, FL can be combined with blockchain (Warnat-Herresthal et al. 2021) or secure multi-party computing technology (Song et al. 2020). In this work, we focus on security and privacy threats against centralized FL.

A Categorization of Federated Learning

In FL, models are trained locally and aggregated at a central server. A global model is obtained after several parameter/gradient aggregation updates. Unlike distributed machine learning, the central server of FL does not have access to the local worker's data. The data distribution among local workers can be independent and identically distributed (i.i.d) or non-independent and identically distributed (non-i.i.d). The types of FL mainly include Horizontal FL (HFL), Vertical FL (VFL) and Federated Transfer Learning (FTL) (Yang et al. 2019). The specific description of each type is as follows:

HFL is suitable for local workers with less sample repetition and more overlapping features. Most existing work mainly focused on the security and privacy towards HFL. VFL is suitable for the scenarios where local workers have the same sample ID and less overlapping features. VFL consists of encrypted entity alignment and encrypted model training. As the number of workers increases, the amount of calculations increases accordingly. SecurBoost (Cheng et al. 2019) is the most representative model of vertical FL, which supports multiple workers to participate in VFL in the FATE framework. FTL (Liu et al. 2018) is suitable for scenes with few sample ID and feature overlap.

Fully decentralized learning

To avoid malicious or semi-honest third parties (central servers), fully decentralized learning emerged (Kim et al. 2018). The fully decentralized learning is usually combined with blockchain, which has proven to be effective in protecting data privacy (Wang et al. 2021; Li et al. 2018). Warnat-Herresthal et al. (2021) proposed a decentralized collaborative computing method called Swarm Learning (SL), which combines privacy-preserving, edge computing and blockchain based peer-to-peer network. Weng et al. (2021) proposed DeepChain, realizing data confidentiality and calculating auditability based on blockchain incentive mechanism and privacy-preserving methods. Based on the combination of blockchain technique and privacy-preserving algorithms, it can be seen that fully decentralized learning enhances the trust guarantee of collaborative computing.

Learning mechanisms

The idea of FL is to jointly train a global model by optimizing the parameters θ with multiple local workers' updates. Basically, there are two aggregating mechanisms named synchronized SGD (Shokri and Shmatikov 2015) and FedAvg (McMahan et al. 2017). In synchronized SGD, each local worker computes the gradient at one batch from its own data and uploads it to the server. In FedAvg, each local worker performs several epochs of gradient descent and provides the updated parameters to the server. Then, the central server will aggregate those gradients or parameters.

The relationship between FL and privacy computing

Privacy computing refers to a range of information technologies that analyze data while ensuring that the data providers do not reveal the private information. In other words, privacy computing is a collection of "data available but not visible" technologies, including FL, secure multi-party computing (MPC), trusted execution environment (TEE), differential privacy (DP), etc. Among them, FL is a derivative technique that integrates distribution machine learning with privacy techniques; secure multi-party computing is a cryptography-based privacy computing technique; trusted execution environment is a trusted hardware-based privacy computing technique; differential privacy is a rigorous mathematical definition of privacy. These techniques are often used in combination to accomplish computing and analyzing data while ensuring the security and privacy of the original data.

The challenge of heterogeneity

With the diversification and complexity of the local workers, the concerns of mutual trust, efficiency, and convergence quality become increasingly obvious. In practical applications, FL needs to break through the heterogeneity of devices in storage, computing and communication capabilities, non-i.i.d data, and model requirements in different local application environments. One effective method to addressing these heterogeneous challenges is to implement personalized FL in three aspects: device (Liu et al. 2020), data (Li et al. 2020) and model (Smith et al. 2017).

The challenge of communication

Reducing communication costs is a major bottleneck for federated computing, as local workers need to multiple interact with a central server and the connections are often unstable. Therefore, how to improve the transmission efficiency while ensuring the accuracy of the joint calculation is an important issue. Existing work indicated that sparse matrix (Konečný et al. 2016) and model compression (Chen et al. 2018) can significantly reduce the communication overhead with little impact on the model accuracy.

The challenge of security and privacy threats

The attack surfaces of FL have expanded due to the characteristics of distribution. For example, malicious local workers may try to steal the privacy information of honest local workers, or malicious local workers can launch collusive attacks to impairing the performance of the final global model.

Multi-phases framework of trusted FL

As shown in Fig. 1, the multi-phases framework of the FL execution can be divided into three phases, including *Data and Behavior Auditing*, *Training* and *Predicting*. The model faces different security and privacy threats at each phase of FL execution. We argue that establishing a trusted FL requires taking effective measures at each phase to fully mitigate security and privacy threats.

- *Data and behavior auditing phase*

  In general, contaminated data and malicious behavior are the main factors affecting model performance. On the one hand, the data of local workers may be contaminated by label noise or feature noise. On the other hand, the historical behavior of local workers may be malicious. The local workers' systems may have some vulnerabilities. These vulnerabilities may have been exploited by adversaries. These threats will impact the subsequent training and prediction of FL. If the risk of data and behavior auditing phase is minimized, the probability of poisoning attacks and privacy inference attacks may decrease.
- *Training phase*

  FL requires multiple local workers working collaboratively to train a global model. In the model training phase, malicious local worker can manipulate their data, model gradients and parameters. Therefore, if adversaries compromise the local workers, they can disturb the integrity of the training dataset or model to impair the performance of the global model. Besides, the central server can also launch passive or active inference attacks. In addition, during the upload and download of model updates, the models may be eavesdropped by intermediaries in the communication channel, resulting in model updates being tampered or stolen. Therefore, it is necessary to protect the transfer of model updates between the local workers and the central server.
- *Predicting phase*

  Once the model is trained, the global model is deployed onto the local worker devices, regardless of whether they participated in the training or not. In this phase, the evasion attacks and privacy inference attacks occur frequently. Evasion attacks usually do not change the target model, but cheat the model to produce false prediction. Privacy inference attacks can reconstruct the characteristics of the model and raw data. The effectiveness of these attacks depend on the knowledge available to the adversaries.

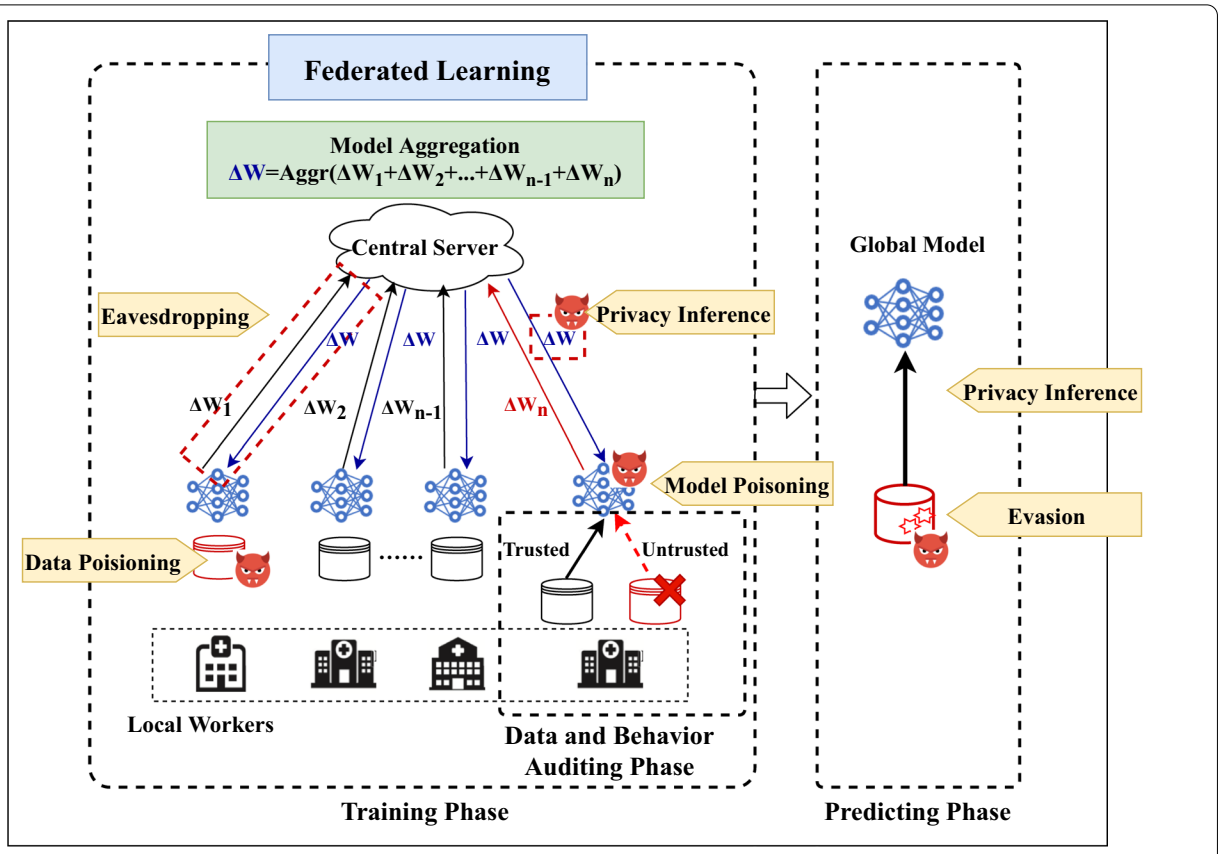

Fig. 1 The multi-phases framework of FL including data and behavior auditing, model training and model predicting

Data and behavior auditing phase

The performance of FL depends on the high quality data of the local workers and the bengin historical behavior of the local workers and central server. Once the data quality is low or there exists malicious behavior, the trained model may become ineffective or even harmful. This section analyzes the threat model, attacks and defenses during the data and behavior auditing phase.

Threat model

In FL, the data of each local worker is available and invisible. Local workers have absolute right of control over their data. This rule makes it difficult to audit the data quality and historical behavior of all local workers. Therefore, a malicious local worker can silently modify the training data to influence the final global model. In addition, the data quality issues, such as unlabeled, noisy or incomplete, may occur during data collection, transmission and processing. These may cause a significant impact on data-based decision-making (Jiang et al. 2021).

Attacks

The data and behavior auditing phase is the first line of defense to ensure the credibility of FL. If this line is breached, a malicious local worker can use low-quality or poisoned data to decrease the performance of the global model or even corrupt the model.

In this phase, the local workers and central server are exposed to existing system, software, and network security threats. Adversaries can cause damage to data and systems by social engineering, penetration attacks, backdoor attacks, and advanced persistent threat (APT) attacks etc. For example, in the cross-device scenario, if the devices have some vulnerabilities, the adversaries can exploit these vulnerabilities to compromise the data and the model (Wang et al. 2014). In addition, insiders can also directly undermine core data and systems by abusing their authorities. Inadvertent errors, and environmental factors in the phases of data collection and transmission will also have a certain impact on the subsequent data analysis.

Defenses

Before the model training phase, one method to ensure the credibility of the FL model is auditing the data quality of the local workers. High-confidence data can effectively reduce the occurrence of poisoning attacks and improve the effectiveness of the model. However, there are few works on *the data quality assessment* in FL. The fact that the data of local workers cannot be aggregated poses some challenges to the overall data quality assessment in FL. Other method is evaluating the historical behavior of local workers and central server. Credibility measurement and credibility verification methods should be proposed based on the system logs.

In addition, the trustworthiness of the local workers should also be dynamically evaluated in the training process (Akujuobi et al. 2019). In general, malicious local workers usually behave differently than most trusted local workers. Therefore, by auditing the model behavior uploaded to the central server, the untrusted local workers can be eliminated.

Training phase

As mentioned earlier, the training phase of FL mainly involves poisoning attacks and privacy inference attacks (white-box). An adversary may launch privacy inference attacks to obtain the victim's privacy information, or launch poisoning attacks locally to affect the performance of the global model. We explain these two attacks in detail in the following subsections.

Privacy inference attacks

FL (McMahan et al. 2016) has recently emerged as a solution to protect data privacy. However, existing work suggested that adversaries can infer different levels of sensitive information from the updated gradients in FL (Hitaj et al. 2017; Nasr et al. 2019; Zhu and Han 2020). In this section, we analyze the reasons for privacy leakage, threat models, attack methods and defense strategies for privacy inference attacks in the training phase.

The reasons of privacy leakage

Several common forms of privacy leakage are listed below.

- *Leakage from embedding layer*

  When a deep learning model learns non-digital data with sparse and discrete input space, it will first convert the input into a low-dimensional vector representation through the embedding layer. For example, in the natural language processing scenario, each word in its vocabulary V signifies a discrete token, and is mapped to a vector after learning. The parameters of the embedding layer can be represented by the matrix $W_{emb} \in R^{|V| \times d}$, where $|V|$ donates the size of the vocabulary and d donates the dimensionality of the embedding. For a specific text, the gradient update of the embedding layer is only based on the words that appear in the text, and the gradient update corresponding to other words are 0. Based on this observation, an adversary can infer which words the local workers used during the FL training period directly (Melis et al. 2019).
- *Leakage from FC layer*

  The fully connected (FC) layer is usually an indispensable component in a deep learning model. The main function of the FC layer is to map the distributed features to the sample label space. Recent studies demonstrated that both the ground-truth labels (Zhao et al. 2020)and the inputs to any FC layer (Geiping et al. 2020; Pan et al. 2020) can be restored from the gradients.

  In regular classification tasks, the deep learning model generally ends with the FC layer, and the loss is calculated by cross-entropy after softmax activation. After the activation function, the output values are between 0-1. Therefore, the sign of gradient according to the correct label is negative and positive otherwise. Hence, the ground-truth labels can definitely be reconstructed from the shared gradients(Zhao et al. 2020). In addition, the input of the fully connected layer can always be calculated from the gradients, regardless of the position in the neural network(Geiping et al. 2020).
- *Leakage from the model gradients*

  The model training is usually regarded as a high-level representation of the data (Lyu 2018), which makes the gradient-based privacy inference attack possible (Aono et al. 2017). Recent work demonstrated that gradients can determine whether an exact sample was used to training (Melis et al. 2019; Shokri et al. 2017), reveal the properties or the representatives of the training samples (Melis et al. 2019), and even completely restore the original training data (Stella et al. 2021; Zhang et al. 2020; Hitaj et al. 2017).

Threat model

In FL, the local workers, the central server, and the communication between the central server and the local workers are considered viable points for the implementation of attack methods. Since FL requires the central server and local workers to exchange gradients/parameters information, white-box attacks can be implemented in FL setting. A comparison of the threat models is summarized in Table 1.

Table 1 Threat model of privacy inference attacks in the training phase, Y (Yes), N (No)

	Knowledge			Ability			Auxiliary data
	Model structure	Weights	Gradients	Train model	Design model	Modify update	
Server	Y	Y	Y	N	Y	Y	N
Eavesdropping	N	Y	Y	N	N	N	N
Workers							
k = 2	Y	Y	Y	Y	N	Y	Y
k > 2	Y	Y	N	Y	N	Y	Y

- *Sever-side attacks*

  A server can be assumed as an honest-but-curious server or a malicious server. The server's knowledge includes the model's structure, weights, and gradients for each epoch of the local workers. Basically, honest-but-curious adversaries may not modify the network structure or send malicious global parameters, while malicious servers vice. Meanwhile, it is usually assumed that the adversaries have unlimited computing resources.
- *Eavesdropping attacks*
  The adversaries located in the communication channel between central server and local workers can launch eavesdropping attacks. The adversaries can steal or tamper some meaningful information, such as model weights or gradients, in each communication.
- *Worker-side attacks* It can be assumed that K workers (of which $K \geq 2$) collaboratively train a joint model using local datasets with negotiating a common FL algorithm.

  When $K = 2$, one of the workers is the adversary, whose goal is to steal information about the training data of another targeted local worker. In this case, an adversary can access the model structure, weights, and gradients of the target worker, just like a server-side adversary. In addition, the adversary takes the responsibility of training the model but cannot modify the model structure.
  When $K > 2$, there are workers who are neither the adversary nor the victim. In this case, the adversary cannot accurately obtain the gradient of the target victim, which increases the difficulty of the attack.

Attacks

According to different inference targets, privacy inference attacks can be summarized as membership inference attacks, class representative inference attacks, property inference attacks, and data reconstruction attacks. Table 2 lists the representative privacy inference attacks against FL in the training phase.

- *Membership inference attacks*

Table 2 Privacy inference attacks against FL in the training phase

	Assumption			Goal	Limitation
	Adversary	Active/Passive	Auxiliary data		
GAN attack (Hitaj et al. 2017)	Worker	Active	No	Classrepresentative inference	Allclassmembers similar
CPA (Nasr et al. 2019)	Worker	Active/Passive	No	Membershipinference	Lackstheoretical proofofthe-bounds
UFL (Melis et al. 2019)	Worker	Active/Passive	Yes	Propertiesinference	Auxiliarycondition maynotmeet
DLG (Zhu and Han 2020)	Server	Passive	No	Inferringtraining dataandlabel	Shallowand smoothnetworks
iDLG (Zhao et al. 2020)	Server	Passive	No	Inferringtrainingdata withimage-labelrecovery	Asingleinput point
Invert gradient (Geiping et al. 2020)	Server	Passive	No	Inferringtraining dataandlabel	Lowperformance atgeneralcase
GradInversion (Yin et al. 2021)	Server	Passive	No	Largebatchimagerecovery forcomplexdatasets	Gradientsonlyupdate onceatlocal ineachiteration
GRNN (Ren et al. 2021)	Server	Passive	No	Generatingtraining dataandlabel	

Membership inference attacks target on determining whether an exact sample was used to train the network (Shokri et al. 2017). An adversary can conduct both passive and active membership inference attacks (Nasr et al. 2019; Melis et al. 2019) to infer whether an exact data was used to train. Passive attacks generally do not modify the learning process, and only make inferences by observing the updated model parameters. Active adversaries can tamper with the training protocol of the FL model and trick other participants into exposing their privacy. A straightforward way is that the adversary shares malicious updates and induces the FL global model to reveal more information about the local data of other local workers. In Nasr et al. (2019), the author presented a comprehensive privacy analysis (CPA) of deep learning by exploiting the privacy vulnerabilities of the SGD algorithm. Experimental results concluded that the gradients are closer to the output layer leak more information, i.e., members and non-members produce different distributions during training. However, their work lacks theoretical proof of the boundaries of privacy breaches.

- *Class representative inference attacks*
  Class Representatives inference attacks aim to obtain the prototypical samples of a target label that the adversary does not own. Hitaj et al. (2017) proposed an active inference attack at inside, called Generative Adversarial Networks Attack, on collaborative deep learning models. Experimental results demonstrated that any malicious local workers using this method could infer privacy information from other participants. However, the experiments require that all class members are similar, and the adversary has prior knowledge of the victim's data labels.
- *Property inference attacks*
  The goal of property inference attacks is to infer meta characteristics of other participants' training data (Melis et al. 2019). Adversaries can obtain specific properties of victim's training data through active or passive inference based on auxiliary label information about the target properties. Passive adversaries can only observe model updates and train a binary attribute classifier of target property to perform inferences. Active adversaries can deceive the FL model to better separate data with and without target attributes, thereby stealing more information. However, the attack condition of auxiliary training data may limit its applicability.
- *Data reconstruction attacks*

Data reconstruction attacks aim to reconstruct training samples and/or associated labels accurately that were used during training.

1. DLG/iDLG

   Previous work has made some contributions in inferring training data features from gradients, but these methods are generally considered "shallow" leakage. Deep Leakage from Gradient (DLG) (Zhu and Han 2020) was the first exploration to fully reveal the private training data from gradients, which can obtain the training inputs as well as the labels in only a few iterations. The core idea of DLG is to synthesis pairs of "dummy" inputs and labels by matching their "dummy" gradients close to the real ones, which can be described as a euclidean matching term (1).

$$\arg\min_{x} ||\nabla_\theta L_\theta(x,y) - \nabla_\theta L_\theta(x^*,y^*)||^2 \quad (1)$$

   Where (x, y) denotes the "dummy" input and the corresponding "dummy" label, and (x^*, y^*) denotes the ground-truth training data and label. Experimental results demonstrated that the training image and label can be jointly reconstructed with a batch size up to 8 and image resolution up to 6464 in shallow and smooth architectures.

   Although DLG has superior performance than the previous "shallow" leakage methods, it suffers from obtaining the ground-truth labels consistently and often fails due to a lousy initialization. In the following, the Improved Gradient Depth Leakage (iDLG) (Zhao et al. 2020) presented theoretically as well as empirically that the ground-truth labels can be recovered with 100% accuracy from the signs of corresponding gradients, such that it improves the fidelity of the extracted data. However, such an algorithm only works for sharing gradients of a single input data.
2. Inverting gradients
   The effectiveness of DLG/iDLG is based on a strong assumption of shallow network and low-resolution recovery, but it is far from realistic scenarios. (Geiping et al. 2020) noted that these assumptions are not necessary if in a right attack. As such, it proposed to use cosine similarity (Chinram et al. 2021) with Total Variation (TV) restriction (Rudin et al. 1992) as the cost function.

$$\arg\min_{x\in[0,1]^n} 1 - \frac{< \nabla_\theta L_\theta(x,y), \nabla_\theta L_\theta(x^*,y) >}{||\nabla_\theta L_\theta(x,y)||||\nabla_\theta L_\theta(x^*,y)||} + \alpha TV(x) \tag{2}$$

Experimental results demonstrated that it is possible to restore the image even in realistic deep and non-smooth architectures

3. GradInversion
The recovery of a single image's label in iDLG has yield great benefits for image restoration (Geiping et al. 2020). In GradInversion (Yin et al. 2021), it implemented a batch-wise labels reconstruction from the final FC layer gradients, enabling a larger batch images restoration in complex training settings. To recover more specific details, GradInversion also introduced a set of regularization, such as image fidelity regularization and group consistency regularization. The optimization function can be formulated as (3):

$$\hat{\mathbf{x}}^* = \underset{\hat{x}}{\operatorname{argmin}} \mathcal{L}(\hat{\mathbf{x}}; W, \Delta W) + \mathcal{R}_{\text{fidelity}}(\hat{\mathbf{x}}) + \mathcal{R}_{\text{group}}(\hat{\mathbf{x}}) \tag{3}$$

Where $\hat{\mathbf{x}}$ is a dummy input batch, and W denotes a network weights, ΔW denotes a batch-averaged gradient of images x^* and labels y^*.
Experimental results indicated that even for complex datasets and deep networks, batch-wise images can be reconstructed with high fidelity through GradInversion. However, this paper only discussed the gradient in one descent step at local.

4. GRNN
Generative Adversarial Networks (GAN) have been shown to be effective in recovering data information (Liu et al. 2021). However, GAN based techniques require additional information, such as class labels which are generally unavailable for privacy persevered learning (Hitaj et al. 2017). Recently, Ren et al. (2021) proposed Generative Regression Neural Network (GRNN), which is capable of restoring training data and their corresponding labels without auxiliary data. Experimental results indicted that GRNN outperforms the DLG/IDLG method with stronger robustness, better stability and higher accuracy. However, same as GradInversion, it only discussed the gradient in one descent step at local.

Defenses

Existing strategies to resisting private inference are usually based on the processing of shared gradient information, including: (1) Compression Gradients; (2) Cryptology Gradients; and (3) Perturbation Gradients, as shown in Table 3.

- *Compression gradients*

The compressibility and sparsity of the gradients are mainly considered as tricks to reduce communication and computational overhead (Haddadpour et al. 2021). Abdelmoniem (2021) illustrated a statistical-based gradient compression technique for distributed training systems, which effectively improves model communication and calculation efficiency. Intuitively, these methods can be directly transferred to FL privacy protection because they reduce the information sources for privacy inferences. In DLG (Zhao et al. 2020), the experimental results suggested that compressing the gradients can successfully prevent deep leakage.

Table 3 Defense methods against privacy inference attacks in the training phase

	Actor	Guarantee			Weakness
		Model	Aggregated value	Local releasedvalue	
Compression gradients					
Pruning	Worker	Y	N	Y	Failintext inferringtask
Dropout	Worker	Y	N	Y	Slightlydecrease modelaccuracy
Cryptology gradients					
SMC	Worker	N	Y	Y	Computationand communicationconsuming
HE	Worker	N	Y	Y	
Perturbation gradients					
CDP	Server	N	Y	N	Requirea trustaggregator
LDP	Worker	N	N	Y	Needenough calibrationnoise
DDP	Worker	N	N	Y	Computation consuming

Another straightforward measure to increase the randomness of gradient is dropout (Zeng et al. 2021). However, dropout produces more generalized features while increasing uncertainty (Srivastava et al. 2014; Chang et al. 2017), which may facilitate inference on generalized data. Experimental results in UFL (Melis et al. 2019) demonstrated that dropout can have a positive impact on their attacks, albeit slightly degrading the performance of the joint model.

- *Cryptology gradients*
  The encryption algorithms often used in FL can be broadly classified as Homomorphic Encryption (HE) (Fang and Qian 2021; Reagen et al. 2021) and secure multi-party computing (SMC) (Li et al. 2020; Liu et al. 2020). HE allows the data to be encrypted and processed, and the decrypted result is equivalent to the operation performed on the original data. Since homomorphic encryption does not change the original data information, it can theoretically ensure that there is no performance loss in model convergence (Yousuf et al. 2021; Wu et al. 2021; Park and Tibouchi 2020). However, the effectiveness of HE comes at the expense of computation and memory (Rahman et al. 2020; Gaid and Salloum 2021), which limits its application (Lyu et al. 2020; Aono et al. 2017). SMC (Yao 1982) enables individual participants to perform joint calculations on their inputs without revealing their own information. This process ensures a high degree of privacy and accuracy. However, it is also computation and communication consuming (Chen et al. 2019). In addition, SMC in FL scenario requires each worker to coordinate with each other during the training process, which is usually impractical.
- *Perturbation gradients*
  The core idea of differential privacy(DP) (Abadi et al. 2016; Triastcyn and Faltings 2019) is to protect data privacy by adding random noise to sensitive information. Basically, DP can be divided into three categories: centralized DP (CDP), local DP (LDP) and distributed DP (DDP) (Lyu et al. 2020; Wei 2020). In FL, CDPs add noise to the aggregated local model gradient through a trusted aggregator to ensure the privacy of the entire data (Lyu et al. 2020). The effectiveness of CDPs requires numerous workers in the FL, which is not apply to H2B scenarios with small-scale workers (Zheng et al. 2021). For LDPs and DDPs, the workers control noise disturbances, which can provide stronger privacy protection. However, LDPs usually need to add sufficient calibration noise to guarantee the data privacy, which may impair the performance of the model (Seif et al. 2020). DDPs guarantee the privacy of each worker by incorporating encryption protocols, which can lead to higher training costs.

Poisoning attacks

Poisoning attacks on machine learning models have been widely studied. These attacks occur in the training phase against FL. On the one hand, adversaries can impair the performance of the final global model on untargeted tasks. On the other hand, adversaries can inject a backdoor into the final global model. In general, poisoning attacks can be categorized as data poisoning and model poisoning.

Threat model

The adversaries can manipulate some local workers to participate in the training process of FL and modify the model updates. The modification methods include changing data features, labels, model parameters, or gradients. The proportion of local workers being manipulated and the amount of modification of training data are the key factors affecting the final training effect. Due to the distributed setting and practical application of FL, the data distribution can be i.i.d., and non-i.i.d. These attacks may be carried out under different data distribution conditions.

Attacks

In general, poisoning attacks can be divided into data poisoning attacks and model poisoning attacks, as well as targeted attacks (backdoor attacks) and untargeted attacks (byzantine attacks) (Lyu et al. 2020; Mothukuri et al. 2021; Enthoven and Al-Ars 2020).

- *Data poisoning and model poisoning attacks*

  Data poisoning attacks Data poisoning attacks are mainly changing the training dataset. The data can be changed by adding noise or flipping the labels.

  Model poisoning attacks The purpose of model poisoning attacks is to arbitrarily manipulate the model updates. These attacks can cause the global model to deviate from the normal model, resulting in degraded model performance or leaving backdoors in the final global model.

  Moreover, local workers sometimes just get the global model, but do not contribute data and computing resources. Such local workers can upload virtual updates, e.g. random parameters, to the central server. These attacks are called free riding

attacks (Lin et al. 2019; Zong et al. 2018). Free riding attacks can also be classified as the model poisoning attacks.

What are the differences and similarities between data poisoning and model poisoning attacks?

For data poisoning attacks, adversaries can only add specific noise to the data or change the labels to affect the performance of the global model. For model poisoning attacks, adversaries usually actively influences the update of the model, e.g., changing objective function. Data poisoning attacks may not as effective as model poisoning attacks (Bhagoji et al. 2019).

The amount of existing data poisoning and model poisoning attacks to construct poisoned samples is to add a specific trigger to the data or to flip the labels. There are not many methods to implement poisoning attacks by adding triggers and unchanging labels.

- *Byzantine and backdoor attacks*
  Byzantine attacks Byzantine attacks are the untargeted attacks and their goal is to cause the failure of the global model.
  Backdoor attacks The goal of a backdoor attack is to make the model fail in a particular task, while the normal task cannot be affected. To some degree, backdoor attack is one type of targeted poisoning attacks.
  Backdoor attacks insert hidden triggers in the global model after training, generally by changing specific features. In the predicting phase, only when there are samples that can trigger backdoor task, the attack will succeed. Therefore, only the adversaries who know how to trigger the backdoor task can successfully launch the attack (this idea can also be applied to model identity authentication (Xiangrui et al. 2020)). However, the current work mainly focus on image datasets, and how to inject backdoor attacks on text datasets needs to be further explored.
- *Perspectives of poisoning attacks*
  We summarize the perspectives of poisoning attacks with the following five questions.

Q1. How to improve the effectiveness of poisoning attacks?

Bagdasaryan et al. (2020) indicated that any local worker can upload a malicious model to the central server during the training phase. They presented a general method called "restrict-and-scale", which enabled adversaries to generate a model with high accuracy in both main task and backdoor task. In addition, they used an objective function to avoid being detected. The objective function includes rewarding the accuracy of the model and punishing the model that deviates from the "normal" of the aggregator. By adding the penalty item L_{ano}, the objective (loss) function is modified as follows:

$$L_{model} = \alpha L_{class} + (1 - \alpha) L_{ano} \tag{4}$$

The adversary's training data include both normal inputs and backdoor inputs, so that L_{class} can balance the accuracy of main task and backdoor task. L_{ano} can be any type of regularization, such as p-norm distance between weight/gradient matrices. In fact, the model poisoning attacks are mainly realized by modifying the objective function. Bhagoji et al. (2019) mainly studied a targeted attack on FL initiated by a few malicious local workers. They proposed the idea of simple boosting. In this processing, malicious local workers try to overcome the impact of the normal local workers and central server on model updates.

In order to improve the stealth of this attack, Bhagoji et al. (2019) proposed the idea of steady model pooling and alternating minimization, making the adversaries avoid being detected by central server.

Q2. What are the conditions for a successful poisoning attack?

Sun et al. (2019) compared the "random sampling attack" with "fixed frequency attack". "Random sampling attack" randomly selects malicious local workers in each round. And "fixed frequency attack" ensures one malicious local worker per f round. They proved that the performance of attacks depends on the proportion of malicious local workers. Baruch et al. (2019) indicted that the model changed within a certain small range is enough to lead to a non-omniscient attack, and some existing defenses (Krum, Trimmed Mean, Bulyan) can be bypassed when the data of each participant satisfy i.i.d.

Q3. How to make a backdoor task more secret?

Xie et al. (2020) proposed a distributed backdoor attack. The original trigger added to samples in one local worker is disassembled into many sub-triggers added to samples in different local workers. Hence, each compromised local worker trains the local model using partial triggers. In the predicting phase, all sub-triggers can be clustered on a single sample to launch a backdoor attack. In this way, the detection difficulty will increase after the triggers are distributed.

Q4. What are the triggers that can launch a successful backdoor attack ?

Wang et al. (2020) indicted that using tail edge samples as triggers can effectively launch backdoor attacks. These samples are unlikely to be part of the training or test data. This provided an idea for finding backdoor triggers.

Q5. Can poisoning attacks bypass the defense strategies?

Existing work presented that the answer to this question is **"YES"**.

Fang et al. (2020) studied model poisoning attacks against byzantine robust FL. It demonstrated that poisoning attacks can succeed even using robust aggregation algorithms such as Krum, Bulyan, Trimmed Mean and Median. Their work can greatly improve the error rate of the global model learned by the above four robust aggregation algorithms.

Defenses

There are two types of defense methods for poisoning attacks, namely robustness aggregation and differential privacy.

- *Robustness aggregation*

  The central server can independently verify the performance of the global model with the validation dataset. The central server can also check whether the malicious local workers' updates are statistically different from other local workers' updates (Bhagoji et al. 2019).

  Various byzantine-robust aggregation methods have been proposed to defend against malicious local workers. Sun et al. (2019) proved that norm threshold of updates can mitigate the attack without affecting the model performance. Fang et al. (2020) generalized RONI and TRIM which were designed to defend against data poisoning attacks to defend against their model poisoning attacks. RFA (Pillutla et al. 2019) aggregated the local models by computing a weighted geometric median using the smoothed Weiszfeld's algorithm. FoolsGold (Fung et al. 2018) is a defense method against sybil attacks on FL. FoolsGold adapts the learning rate (aggregate weight) of local models based on the model similarity in each round. In the Median method (Yin et al. 2018), the central server sorts the parameters of local models, and takes the median as the next round global model. Same as Median, in the Trimmed Mean method (Yin et al. 2018), the server will also sort the parameter of local models. Then, the central server removes the largest and smallest β parameters, and computes the mean of the remaining $m - 2\beta$ parameters as the next round global model. Blanchard et al. (2017) selects one of the local models which is similar to other local models as the global model. Even if the selected local model comes from the compromised local workers, its influence will be limited. Mhamdi et al. (2018) combined Krum and a variant of trimmed mean. Specifically, Bulyan first iteratively applies Krum to select θ local models. It then uses a variant of Trimmed Mean to aggregate the θ local models.
- *Differential privacy*

  Sun et al. (2019) added Gaussian noise with small standard deviations to the aggregated global model to mitigate threats. Naseri et al. (2020) demonstrated that both LDP and CDP can defend against backdoor attacks.

Predicting phase

In the model predicting phase, there are still security and privacy threats, as shown in Table 4. The global model are visible to the local workers and central server, which may increase the possibility of launching attacks in the predicting phase. Malicious local workers or central server may infer honest local workers' sensitive information from the global model.

Evasion attacks

Evasion attacks aim to cheat the target model by constructing specific samples called adversarial examples. Usually, some subtle noise added to the input samples cannot be detected by human beings, and cause the

Table 4 Evasion and privacy inference attacks in the model predicting phase

Attack Types	Goal	Attack Methods	Defense Strategies
Evasion	Making the model misclassification on adversary examples	Based on optimization; Based on gradient; Based on decision-making and so on	Empirical defense; Certified defense
Model Inversion	Obtaining privacy information of the original data	Attribute inference; Property inference	Model structure defense; Information obfuscation; Query control; Differential privacy
Membership Inference	Testing whether a specific point was part of the training dataset	Shadow model; Boundary attack	
Model Extraction	Obtaining relevant information about the target model	Model parameter; Hyperparameter	

model to give incorrect classification results. A classic example is that a panda image with a small amount of noise is identified as a gibbon (Szegedy et al. 2014). The adversarial examples can be attributed to the linear characteristics in high-dimensional space (Goodfellow et al. 2015) and the non-robust characteristics (Gilmer et al. 2019).

According to the optimization objective, evasion attacks can be divided into targeted attacks with class-specific errors, and untargeted attacks that do not consider class-specific errors. The evasion attacks have attracted wide attentions and been applied to many scenes, such as attacking autonomous driving (Lu et al. 2017), internet of things (Yulei 2021), face recognition (Sharif et al. 2016), and speech recognition (Carlini et al. 2016).

Threat model

From the perspective of the adversary's knowledge, the attack can be divided into white-box and black-box attacks. Under the white-box attacks, the adversary has complete knowledge about the target model, including neural network structure, model parameters and output. In contrast, under the black-box attacks, the adversary does not know the neural network architecture, parameters, and other target model information. The attack can be implemented according to the query results of the target model.

Attacks

The main research direction of the evasion attacks (adversarial examples attacks) is to design adversarial examples and to break through the robustness of the model.

- *In computer vision (CV)*

  White-box evasion attacks are mainly based on optimization, gradient, classification hyperplane and so on. For the optimization-based methods, how to find the minimum possible attack disturbance is defined as an optimization problem. The most representative method is C&W (Carlini and Wagner 2017) and L-BFGS (Szegedy et al. 2014). For the gradient-base methods, their core idea is to modify the input sample in the gradient direction. The main methods include one attack, such as FGSM (Goodfellow et al. 2015) and iterative attack, such as i-FGSM (Kurakin et al. 2017). For the classification hyperplane-based methods, their purpose is to find the minimum disturbance that fool deep networks, such as Deepfool (Moosavi-Dezfooli et al. 2016). Black-box evasion attacks are mainly based on transferability, gradient estimation and decision-making (Ji et al. 2021).
- *In natural language processing (NLP)*

  Evasion attacks in CV domain have made significant breakthroughs in attack methods. However, there are still many challenges in NLP tasks. Due to the inherent differences between image and text data, the evasion attacks for the CV tasks cannot be directly applied to the NLP tasks. First, image data (such as pixel value) is continuous, but *text data is discrete*, so that it is a challenge to disturb along the gradient direction. Second, a tiny change in the pixel values can cause image data disturbance, and this disturbance is challenging to be detected by human beings. However, minor disturbances can be easily detected for text data.

  The adversarial examples for text data can be char, word and sentence levels (Zeng et al. 2020). There are three representative methods of generating adversarial examples in text classification: genetic attack (Ren et al. 2020), HoTFLip (Ebrahimi et al. 2018) and MHA (Zhang et al. 2019).

Defenses

- *Empirical defense*

  Many researchers suggest that image preprocessing and feature transformation can defend against evasion attacks. However, these methods are almost ineffective in the scenario where the adversary knows the defense methods (Ji et al. 2021). Security-by-obscurity mechanism improves the model security by hiding information, mainly including model fusion, gradient mask and randomization (Ji et al. 2021). The main methods affecting decision boundary are adversarial training (Madry et al. 2018). In order to improve the robustness of the model, the defender generates the adversary examples and mixes them with the original samples to train the model. However, in CV, adversarial training tends to overfit the model to the specific constraint region, which leads to the degradation of generalization performance of the model.
- *Certified defense*

  Certified defense (Lécuyer et al. 2019; Li et al. 2018) has been studied in recent years, and it is provably robust to certain kinds of adversarial perturbations. Cohen et al. (2019) prove a tight robustness guarantee in l_2 norm for smoothing with Gaussian noise. Strong empirical results suggest that randomized smoothing is a promising direction for future research into robust adversarial classification.

Privacy inference attacks

Privacy inference attacks also happened in predicting phase. These attacks include model inversion, membership inference, and model extraction.

Threat model

In the model predicting phase, the adversaries may have no knowledge of the parameters of the model, and only have access to query the model. In particular, different assumptions about adversary's knowledge, such as with or without auxiliary data, and knowing the confidence vector or label-only, make the attack and defense methods difficult to be generally applicable.

Attacks

- *Model inversion*

  Model inversion attacks mainly use some APIs provided by a machine learning system to obtain the preliminary information of the model. With this preliminary information, the adversaries can analyze the model to obtain some relevant information about the original data (Jayaraman and Evans 2019). We argue that model inversion attacks are categorized as attribute inference attacks and property inference attacks.

  Attribute inference attacks(Fredrikson et al. 2014; Yeom et al. 2018) aim to learn hidden sensitive attributes of a sample. The prediction results of machine learning models often contain a lot of reasoning information about the sample. Fredrikson et al. (2014) proposed that the input information contained in the confidence output can be used as a measure of the input inversion attacks. Property inference attacks (Song and Shmatikov 2020) try to infer whether the training dataset has a specific property. We argue that the difference between attribute and property inference attacks is that attribute inference attacks obtain the features involved in the main task, while the property inference attacks obtain the features independent of the main task.
- *Membership inference*

  Membership inference attacks aim to test whether a specific point is part of the training dataset. Shokri et al. (2017) first proposed this attack catting it as a supervised learning problem. Specifically, the adversary trains multiple shadow models to mimic the behavior of the target model, and trains an attack model from data derived from the shadow models' outputs. Salem et al. (2019) pointed that the above method has many assumptions on the adversary, such as the use of several shadow models, knowledge of the target model structure, and a dataset from the same distribution as the target model's training dataset. They relax these assumptions and study three different types of attacks. Choquette-Choo et al. (2021) and Li and Zhang (2021) focus on how to implement the attack in the case of label-only. These methods based on an intuition that it is more difficult to perturb the member inputs to mislead the target model than to perturb the non-member inputs. The fundamental reason for the success of the membership inference attacks is the overfitting of the target model.

  Yeom et al. (2018) assumed that the adversary has full white-box access to the target model, along with some auxiliary information. Under the same settings, Nasr et al. (2019) obtained the activation function output and gradients of the model as the features to train the attack model. Leino and Fredrikson (2020) presented a white-box membership inference attack based on the intimate understanding of information leakage through the target model's idiosyncratic use of features. Chen et al. (2020) studied membership inference attacks against generative models under various threat models, and the attack calibration technique proposed significantly boosts the attack performance.
- *Model extraction*

  The adversaries obtain relevant information about the target model through a circular query to simulate the decision boundary of the target model. Model extraction attacks can be divided into model parameter extraction and hyperparameter extraction attacks. Model parameter extraction attacks aim to recover the model parameters via black-box access to the target model. The main methods include adversarial learning, based on meta-model, alternative-model or equation-solving attacks (Ren et al. 2021; Tramèr et al. 2016). Hyperparameter extraction attacks try to recover the underlying hyperparameters, such as regularization coefficient (Wang and Gong 2018).

Defenses

Grosso et al. (2021) analysed fundamental bounds on information leakage, which can help us to construct privacy-preserving ML models. Ren et al. (2021) concluded that the following types of data privacy-preserving measures could be adopted: model structure defense (e.g. reducing the sensitivity of the model to training samples and overfitting of the model), information obfuscation defense (e.g. confusing the output of the model), and query control defense (e.g. controlling query times). The reasons of successful attacks are very important for

studying defense methods. Facts have proved that the existing defense methods still have some defects. For example, overfitting is the main reason why membership inference attacks can succeed, and the data enhancement mechanism can effectively prevent overfitting. However, Kaya and Dumitras (2021) evaluated the implementation of two membership inference attacks on seven data enhancement mechanisms and differential privacy. They found that "applying augmentation does not limit the risk", so that we should to study more robust defense methods.

In particular, differential privacy is used to protect data privacy (Papernot et al. 2018). At training, random noise may add to the data, objective function, gradients, parameters, or output. At Inferring, due to the noise added in the training process, the model's generalization performance will be reduced, so that there is a trade-off between privacy and utility. In order to achieve the utility-loss guarantees, Jia et al. (2019) added crafted noise to each confidence score vector to turn it into an adversarial example against black-box membership inference attacks. This method can mislead the adversary's attack model, and it belongs to information obfuscation defense.

Perspectives

- *Security and privacy threats on VFL and FTL*

  Most previous work has focused on security and privacy threats in HFL, while work on security and privacy threats in VFL/FTL is limited. In VFL, usually only one local worker has the label of training data. Hence, whether the threats in HFL still exit in VFL/FTL and whether there are new threats in VFL/FTL deserve further study (Lyu et al. 2020). Some attacks against VFL have been proposed. For example, Luo et al. (2020) proposed a feature inference attack against VFL in the predicting phase. Weng et al. (2020) implemented two practical attacks against VFL based on logistic regression and XGBoost.
- *Limitations of attack scenarios*
  For property and membership inference attacks in the training phase, if the adversaries are local workers, they can only obtain the sum of information from other local workers. Therefore, they can only infer that there is a specific sample or property in the overall dataset of other local workers. How to confirm the specific information belonging to which honest local worker is an open problem.
  For data reconstruction attacks, the existing work assumed that adversaries are located in the central server. They can collect the parameters or gradients about all local workers and launch a white-box data reconstruction attack. However, these attacks can only recover a single sample or a batch of samples when $iteration = 1$, where iteration means stochastic gradient update steps per epoch. How to implement data reconstruction attacks under $epoch > 1$ and $iteration > 1$ is a big challenge.
  For evasion attacks and poisoning attacks, the key to the success depends on finding or generating the appropriate samples as triggers. For the discrete datasets, further work on evasion and poisoning attacks is needed (Wang et al. 2020). Except for the most obvious difference, namely that evasion attacks occur in the predicting phase and poisoning attacks occur in the training phase, it is valuable to analyze the connections and differences between them in theory (Pang et al. 2020; Suciu et al. 2018; Demontis et al. 2019).
- *Weakness of the defense strategies*
  Recent evidence suggests that the defense methods of FL have some shortcomings. For example, robust aggregation algorithms can be circumvented by poisoning attacks; DP affects the usability of the model; SMC and HE can cause model inefficiency to some extent (Kanagavelu et al. 2020). With the continuous improvement of attack methods, targeted defense strategies need to be put forward as soon as possible to ensure the security and privacy of FL.
  Besides, previous work emphasized that detecting whether the local workers are trusted. The local workers should confirm whether the central server is trusted (Guowen et al. 2020; Guo et al. 2021) in the training phase. Previous work also established that adversaries can extract memorized information from the model (Song et al. 2017). Therefore, how to make the trained model remember less information about data is also a research direction (He et al. 2021).
- *Building a trustworthy FL*
  There are many threats against FL in every phase from data and behavior auditing, model training to predicting. In particular, the data and behavior auditing for FL should be paid more attention, as it is the first line of defense for FL security and privacy. In addition, more trustworthiness measurement and assessment methods can be investigated to evaluate the trustworthiness of local staff and central servers before the model training phase. In the model training phase, the centralized FL needs to employ privacy-preserving and security technologies, and advances machine learning algorithms. Warnat-Herrestha et al. (2021) construct a decentralized collaborative learning platform based on blockchain. This platform fully considers the trusted access of institutions, and employs Trusted Execution Environment

(TEE), DP and HE to protect private information. This platform can provide experience for centralized FL. Building a FL systems on Blockchain may be more reliable due to its nature of immutability and decentralization.

Conclusion

Federated Learning (FL) has recently emerged as a solution to the issues of data silos. However, FL itself is still riddled with attack surfaces that arouse the risk of data privacy and model robustness. In this work, we identify the issues and provide the taxonomy of FL based on the multi-phases it works with, including data and behavior auditing phase, training phase and predicting phase. Finally, we present the perspectives of FL. Our work indicate that FL is promising in privacy enhancement technology. However, building a trusted FL system is confronted with security and privacy issues inherited by its distributed nature. One should consider the threats existing in all the phases on which the execution of FL follows, including the data and behavior auditing phase, training phase and predicting phase.

Acknowledgements
We are very grateful to Chao Li, Hao Zhen and Xiaoting Lyu for their useful suggestions.

Authors' contributions
PL is responsible for writing the contents except "Privacy inference attacks" in Section "Training phase" and revising the expression of the full text. XX is responsible for writing "Privacy inference attacks" in Section "Training phase" and revising the expression of the full text. WW is responsible for the proposal of innovation points and the overall grasp of the structure and content of the full text. All authors read and approved the final manuscript.

Authors' information
Pengrui Liu received the BA degree in 2017 at Shanxi University and the MA degree in 2020 at North University of China. He is currently pursuing a Ph.D. degree in Beijing Jiaotong University, China. His research interests include privacy enhancement technology and security of deep learning.
Xiangrui Xu received the BA and MA degrees in 2018 and 2021, respectively, at Wuhan Polytechnic University. Since 2018, she has been a Research Assistant with the Artificial Intelligence Laboratory of Mathematics and Computer College. She is currently pursuing a Ph.D. degree at Beijing Jiaotong University, China. Her research interests include privacy enhancement technology and security of deep learning.
Wei Wang is currently a full professor and chairs the Department of Information Security, Beijing Jiaotong University, China. He earned his Ph.D. degree from Xi'an Jiaotong University, in 2006. He was a postdoctoral researcher in University of Trento, Italy, during 2005–2006. He was a postdoctoral researcher in TELECOM Bretagne and in INRIA, France, during 2007–2008. He was a European ERCIM Fellow in Norwegian University of Science and Technology (NTNU), Norway, and in Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg, during 2009-2011. He visited INRIA, ETH, NTNU, CNR, and New York University Polytechnic. He has authored or co-authored over 100 peer-reviewed papers in various journals and international conferences. His main research interests include privacy enhancement technology and blockchain.

Funding
This work was supported in part by National Key R&D Program of China, under Grant 2020YFB2103802, in part by the National Natural Science Foundation of China, uder grant U21A20463, and in part by the Fundamental Research Funds for the Central Universities of China under Grant KKJB320001536.

Availability of data and materials
Not applicable.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 2 August 2021 Accepted: 1 December 2021
Published online: 02 February 2022

References

Abadi M, Chu A, Goodfellow I, McMahan HB, Mironov I, Talwar K, Zhang L (2016) Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security, pp 308–318. https://doi.org/10.1145/2976749.2978318
Abdelmoniem AM, Elzanaty A, Alouini M-S, Canini M (2021) An efficient statistical-based gradient compression technique for distributed training systems. In: Proceedings of Machine Learning and Systems, 3
Akujuobi U, Han Y, Zhang Q, Zhang X (2019) Collaborative graph walk for semi-supervised multi-label node classification. In: Wang J, Shim K, Wu X (eds) 2019 IEEE international conference on data mining, ICDM 2019, Beijing, China, November 8–11, 2019, pp 1–10. IEEE. https://doi.org/10.1109/ICDM.2019.00010
Aono Y, Hayashi T, Wang L, Moriai S et al (2017) Privacy-preserving deep learning via additively homomorphic encryption. IEEE Trans Inf Forensics Secur 13(5):1333–1345. https://doi.org/10.1109/TIFS.2017.2787987
Bagdasaryan E, Veit A, Hua Y, Estrin D, Shmatikov V (2020) How to backdoor federated learning. In: Chiappa S, Calandra R (eds) The 23rd international conference on artificial intelligence and statistics, AISTATS 2020, 26–28 August 2020, Online [Palermo, Sicily, Italy], volume 108 of proceedings of Machine Learning Research. PMLR, pp 2938–2948
Baruch G, Baruch M, Goldberg Y (2019) A little is enough: circumventing defenses for distributed learning. In: Wallach HM, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox EB, Garnett R (eds) Advances in neural information processing systems 32: annual conference on neural information processing systems 2019, NeurIPS 2019, December 8–14, 2019, Vancouver, BC, Canada, pp 8632–8642
Bhagoji AN, Chakraborty S, Mittal P, Calo SB (2019) Analyzing federated learning through an adversarial lens. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th international conference on machine learning, ICML 2019, 9–15 June 2019, Long Beach, California, USA, volume 97 of proceedings of machine learning research. PMLR, pp 634–643
Blanchard P, Mhamdi EEM, Guerraoui R, Stainer J (2017) Machine learning with adversaries: byzantine tolerant gradient descent. In: Guyon I, von Luxburg U, Bengio S, Wallach HM, Fergus R, Vishwanathan SVN, Garnett R (eds) Advances in neural information processing systems 30: annual conference on neural information processing systems 2017, December 4–9, 2017, Long Beach, CA, USA, pp 119–129
Bonawitz K, Ivanov V, Kreuter B, Marcedone A, McMahan HB, Patel S, Ramage D, Segal A, Seth K(2016) Practical secure aggregation for federated learning on user-held data. arXiv preprint arXiv:1611.04482
Carlini N, Mishra P, Vaidya T, Zhang Y, Sherr M, Shields C, Wagner DA, Zhou W (2016) Hidden voice commands. In: Holz T, Savage S (eds) 25th USENIX security symposium, USENIX security 16, Austin, TX, USA, August 10–12, 2016. USENIX Association, pp 513–530
Carlini N, Wagner DA (2017) Towards evaluating the robustness of neural networks. In: 2017 IEEE symposium on security and privacy, SP 2017, San Jose, CA, USA, May 22–26, 2017, pp 39–57. IEEE Computer Society. https://doi.org/10.1109/SP.2017.49
Chang C-H, Rampasek L, Goldenberg A (2017) Dropout feature ranking for deep learning models. arXiv preprint arXiv:1712.08645
Chen C-Y, Choi J, Brand D, Agrawal A, Zhang W, Gopalakrishnan K (2018) Adacomp : adaptive residual gradient compression for data-parallel distributed training. In: McIlraith SA, Weinberger KQ (eds) Proceedings

of the thirty-second AAAI conference on artificial intelligence, (AAAI-18), the 30th innovative applications of artificial intelligence (IAAI-18), and the 8th AAAI symposium on educational advances in artificial intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2–7, 2018. AAAI Press, pp 2827–2835

Cheng K, Fan T, Jin Y, Liu Y, Chen T, Yang Q (2019) Secureboost: a lossless federated learning framework. CoRR arXiv:1901.08755

Chen V, Pastro V, Raykova M(2019) Secure computation for machine learning with spdz. arXiv preprint arXiv:1901.00329

Chen D, Yu N, Zhang Y, Fritz M (2020) Gan-leaks: a taxonomy of membership inference attacks against generative models. In: Ligatti J, Ou X, Katz J, Vigna G (eds) CCS '20: 2020 ACM SIGSAC conference on computer and communications security, virtual event, USA, November 9–13, 2020. ACM, pp 343–362. https://doi.org/10.1145/3372297.3417238

Cheu A, Smith AD, Ullman JR (2021) Manipulation attacks in local differential privacy. In: 42nd IEEE symposium on security and privacy, SP 2021, San Francisco, CA, USA, 24–27 May 2021. IEEE, pp 883–900

Chinram R, Mahmood T, Ur Rehman U, Ali Z, Iampan A (2021) Some novel cosine similarity measures based on complex hesitant fuzzy sets and their applications. J Math. https://doi.org/10.1155/2021/6690728

Choquette-Choo CA, Tramèr F, Carlini N, Papernot N (2021) Label-only membership inference attacks. In: Meila M, Zhang T (eds) Proceedings of the 38th international conference on machine learning, ICML 2021, 18–24 July 2021, virtual event, volume 139 of proceedings of machine learning research. PMLR, pp 1964–1974

Cohen JM, Rosenfeld E, Kolter JZ (2019) Certified adversarial robustness via randomized smoothing. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th international conference on machine learning, ICML 2019, 9–15 June 2019, Long Beach, California, USA, volume 97 of proceedings of machine learning research. PMLR, pp 1310–1320

Demontis A, Melis M, Pintor M, Jagielski M, Biggio B, Oprea A, Nita-Rotaru C, Roli F (2019) Why do adversarial attacks transfer? Explaining transferability of evasion and poisoning attacks. In: Heninger N, Traynor P (eds) 28th USENIX security symposium, USENIX security 2019, Santa Clara, CA, USA, August 14–16, 2019, pp 321–338. USENIX Association

Ebrahimi J, Rao A, Lowd D, Dou D (2018) Hotflip: white-box adversarial examples for text classification. In: Gurevych I, Yusuke M (eds) Proceedings of the 56th annual meeting of the association for computational linguistics, ACL 2018, Melbourne, Australia, July 15–20, 2018, volume 2: short papers. Association for Computational Linguistics, pp 31–36. https://doi.org/10.18653/v1/P18-2006

Enthoven D, Al-Ars Z (2020) An overview of federated deep learning privacy attacks and defensive strategies. CoRR arXiv:2004.04676

Fang H, Qian Q (2021) Privacy preserving machine learning with homomorphic encryption and federated learning. Future Internet 13(4):94

Fang M, Cao X, Jia J, Gong NZ (2020) Local model poisoning attacks to byzantine-robust federated learning. In: Capkun S, Roesner F (eds) 29th USENIX security symposium, USENIX security 2020, August 12–14, 2020. USENIX Association, pp 1605–1622

Fredrikson M, Lantz E, Jha S, Lin SM, Page D, Ristenpart T (2014) Privacy in pharmacogenetics: an end-to-end case study of personalized warfarin dosing. In: Fu K, Jung J (eds) Proceedings of the 23rd USENIX Security Symposium, San Diego, CA, USA, August 20–22, 2014. USENIX Association, pp 17–32

Fung C, Yoon CJM, Beschastnikh I (2018) Mitigating sybils in federated learning poisoning. CoRR arXiv:1808.04866

Gaid ML, Salloum SA (2021) Homomorphic encryption. In: The international conference on artificial intelligence and computer vision. Springer, pp 634–642

Geiping J, Bauermeister H, Dröge H, Moeller M (2020) Inverting gradients—how easy is it to break privacy in federated learning? arXiv preprint arXiv:2003.14053

Gilmer J, Ford N, Carlini N, Cubuk ED (2019) Adversarial examples are a natural consequence of test error in noise. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th international conference on machine learning, ICML 2019, 9–15 June 2019, Long Beach, California, USA, volume 97 of proceedings of Machine Learning Research. PMLR, pp 2280–2289

Goodfellow IJ, Shlens J, Szegedy C (2015) Explaining and harnessing adversarial examples. In: Bengio Y, LeCun Y (eds) 3rd international conference on learning representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, conference track proceedings

Grosso GD, Pichler G, Palamidessi C, Piantanida P (2021) Bounding information leakage in machine learning. CoRR arXiv:2105.03875

Guo X, Liu Z, Li J, Gao J, Hou B, Dong C, Baker T (2021) Verifl: communication-efficient and fast verifiable aggregation for federated learning. IEEE Trans Inf Forensics Secur 16:1736–1751. https://doi.org/10.1109/TIFS.2020.3043139

Guowen X, Li H, Liu S, Yang K, Lin X (2020) Verifynet: secure and verifiable federated learning. IEEE Trans Inf Forensics Secur 15:911–926. https://doi.org/10.1109/TIFS.2019.2929409

Haddadpour F, Kamani MM, Mokhtari A, Mahdavi M (2021) Federated learning with compression: unified analysis and sharp guarantees. In: International conference on artificial intelligence and statistics. PMLR, pp 2350–2358

He Y, Meng G, Chen K, He Jn, Hu X (2021) Deepobliviate: a powerful charm for erasing data residual memory in deep neural networks. CoRR arXiv:2105.06209

Hitaj B, Ateniese G, Perez-Cruz F (2017) Deep models under the gan: information leakage from collaborative deep learning. In: Proceedings of the 2017 ACM SIGSAC conference on computer and communications security, pp 603–618. https://doi.org/10.1145/3133956.3134012

Jayaraman B, Evans D (2019) Evaluating differentially private machine learning in practice. In: Heninger N, Traynor P (eds) 28th USENIX security symposium, USENIX security 2019, Santa Clara, CA, USA, August 14–16, 2019. USENIX Association, pp 1895–1912

Ji SL, Du TY, Li JF et al (2021) Security and privacy of machine learning models: a survey. Ruan Jian Xue Bao/J Softw 32(1):41–67 (in Chinese)

Jiang G, Wang W, Qian Y, Liang J (2021) A unified sample selection framework for output noise filtering: an error-bound perspective. J Mach Learn Res 22:18:1-18:66

Jia J, Salem A, Backes M, Zhang Y, Gong NZ (2019) Memguard: defending against black-box membership inference attacks via adversarial examples. In: Lorenzo C, Johannes K, XiaoFeng W, Jonathan K (eds) Proceedings of the 2019 ACM SIGSAC conference on computer and communications security, CCS 2019, London, UK, November 11–15, 2019, pp 259–274. ACM. https://doi.org/10.1145/3319535.3363201

Kairouz P, McMahan HB, Avent B et al (2019) Advances and open problems in federated learning. CoRR arXiv:1912.04977

Kanagavelu R, Li Z, Samsudin J, Yang Y, Yang F, Goh RSM, Cheah M, Wiwatphonthana P, Akkarajitsakul K, Wang S (2020) Two-phase multi-party computation enabled privacy-preserving federated learning. In: 20th IEEE/ACM international symposium on cluster, cloud and internet computing, CCGRID 2020, Melbourne, Australia, May 11-14, 2020. IEEE, pp 410–419. https://doi.org/10.1109/CCGrid49817.2020.00-52

Kaya Y, Dumitras T (2021) When does data augmentation help with membership inference attacks? In: Meila M, Zhang T (eds) Proceedings of the 38th international conference on machine learning, ICML 2021, 18–24 July 2021, virtual event, volume 139 of proceedings of Machine Learning Research. PMLR, pp 5345–5355

Kim H, Park J, Bennis M, Kim S-L (2018) On-device federated learning via blockchain and its latency analysis. CoRR arXiv:1808.03949

Konečný J, McMahan HB, Yu FX, Richtárik P, Suresh AT, Bacon D (2016) Federated learning: strategies for improving communication efficiency. CoRR arXiv:1610.05492

Kurakin A, Goodfellow IJ, Bengio S (2017) Adversarial examples in the physical world. In: 5th international conference on learning representations, ICLR 2017, Toulon, France, April 24–26, 2017, workshop track proceedings. OpenReview.net

Lécuyer M, Atlidakis V, Geambasu R, Hsu D, Jana S (2019) Certified robustness to adversarial examples with differential privacy. In: 2019 IEEE symposium on security and privacy, SP 2019, San Francisco, CA, USA, May 19–23, 2019. IEEE, pp 656–672. https://doi.org/10.1109/SP.2019.00044

Leino K, Fredrikson M (2020) Stolen memories: leveraging model memorization for calibrated white-box membership inference. In: Capkun S, Roesner F (eds) 29th USENIX security symposium, USENIX security 2020, August 12–14, 2020. USENIX Association, pp 1605–1622

Li L, Liu J, Cheng L, Qiu S, Wang W, Zhang X, Zhang Z (2018) Creditcoin: a privacy-preserving blockchain-based incentive announcement

network for communications of smart vehicles. IEEE Trans Intell Transp Syst 19(7):2204–2220
Li Y, Zhou Y, Jolfaei A, Dongjin Y, Gaochao X, Zheng X (2020) Privacy-preserving federated learning framework based on chained secure multi-party computing. IEEE Internet Things J. https://doi.org/10.1109/JIOT.2020.3022911
Li B, Chen C, Wang W, Carin L (2018) Second-order adversarial attack and certifiable robustness. CoRR arXiv:1809.03113
Li X, Huang K, Yang W, Wang S, Zhang Z(2020) On the convergence of fedavg on non-iid data. In: 8th international conference on learning representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020. OpenReview.net
Lin BY, He C, Zeng Z, Wang H, Huang Y, Soltanolkotabi M, Ren X, Avestimehr S (2021) Fednlp: a research platform for federated learning in natural language processing. CoRR arXiv:2104.08815
Lin J, Min D, Liu J (2019) Free-riders in federated learning: attacks and defenses. CoRR arXiv:1911.12560
Li T, Sahu AK, Zaheer M, Sanjabi M, Talwalkar A, Smith V (2018) Federated optimization in heterogeneous networks. arXiv preprint arXiv:1812.06127
Liu J, Yuan Tian Yu, Zhou YX, Ansari N (2020) Privacy preserving distributed data mining based on secure multi-party computation. Comput Commun 153:208–216. https://doi.org/10.1016/j.comcom.2020.02.014
Liu M-Y, Huang X, Jiahui Yu, Wang T-C, Mallya A (2021) Generative adversarial networks for image and video synthesis: algorithms and applications. Proceed IEEE 109(5):839–862. https://doi.org/10.1109/JPROC.2021.3049196
Liu Y, Chen T, Yang Q (2018) Secure federated transfer learning. CoRR arXiv:1812.03337
Liu L, Zhang J, Song S, Letaief KB (2020) Client-edge-cloud hierarchical federated learning. In: 2020 IEEE international conference on communications, ICC 2020, Dublin, Ireland, June 7–11, 2020. IEEE, pp 1–6. https://doi.org/10.1109/ICC40277.2020.9148862
Li Z, Zhang Y (2021) Membership leakage in label-only exposures. CoRR arXiv:2007.15528
Luo X, Wu Y, Xiao X, Ooi BC (2020) Feature inference attack on model predictions in vertical federated learning. CoRR arXiv:2010.10152
Lu J, Sibai H, Fabry E (2017) Adversarial examples that fool detectors. CoRR arXiv:1712.02494
Lyu L (2018) Privacy-preserving machine learning and data aggregation for Internet of Things. PhD thesis
Lyu L, Yu H, Ma X, Sun L, Zhao J, Yang Q, Yu PS (2020) Privacy and robustness in federated learning: attacks and defenses. arXiv preprint arXiv:2012.06337
Lyu L, Yu H, Ma X, Sun L, Zhao J, Yang Q, Yu PS (2020) Privacy and robustness in federated learning: attacks and defenses. CoRR arXiv:2012.06337
Lyu L, Yu H, Yang Q (2020) Threats to federated learning: a survey. arXiv preprint arXiv:2003.02133
Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A (2018) Towards deep learning models resistant to adversarial attacks. In: 6th international conference on learning representations, ICLR 2018, Vancouver, BC, Canada, April 30–May 3, 2018, conference track proceedings. OpenReview.net
McMahan HB, Moore E, Ramage D, Arcas BA (2016) Federated learning of deep networks using model averaging. arXiv preprint arXiv:1602.05629
McMahan B, Moore E, Ramage D, Hampson S, Arcas BA (2017) Communication-efficient learning of deep networks from decentralized data. In: Singh A, Zhu X (eds) Proceedings of the 20th international conference on artificial intelligence and statistics, AISTATS 2017, 20–22 April 2017, Fort Lauderdale, FL, USA, volume 54 of proceedings of machine learning research. PMLR, pp 1273–1282
Melis L, Song C, De Cristofaro E, Shmatikov V (2019) Exploiting unintended feature leakage in collaborative learning. In: 2019 IEEE symposium on security and privacy (SP). IEEE, pp 691–706. https://doi.org/10.1109/SP.2019.00029
Mhamdi EEM, Guerraoui R, Rouault S (2018) The hidden vulnerability of distributed learning in byzantium. In: Dy JG, Krause A (eds) Proceedings of the 35th international conference on machine learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10–15, 2018, volume 80 of proceedings of machine learning research. PMLR, pp 3518–3527
Moosavi-Dezfooli S-M, Fawzi A, Frossard P (2016) Deepfool: a simple and accurate method to fool deep neural networks. In: 2016 IEEE conference on computer vision and pattern recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016. IEEE Computer Society, pp 2574–2582. https://doi.org/10.1109/CVPR.2016.282
Mothukuri V, Parizi RM, Pouriyeh S, Huang Y, Dehghantanha A, Srivastava G (2021) A survey on security and privacy of federated learning. Future Gener Comput Syst 115:619–640. https://doi.org/10.1016/j.future.2020.10.007
Naseri M, Hayes J, De Cristofaro E (2020) Toward robustness and privacy in federated learning: experimenting with local and central differential privacy. CoRR arXiv:2009.03561
Nasr M, Shokri R, Houmansadr A (2019) Comprehensive privacy analysis of deep learning: passive and active white-box inference attacks against centralized and federated learning. In: 2019 IEEE symposium on security and privacy (SP). IEEE, pp 739–753. https://doi.org/10.1109/SP.2019.00065
Pang R, Shen H, Zhang X, Ji S, Vorobeychik Y, Luo X, Liu AX, Wang T (2020) A tale of evil twins: adversarial inputs versus poisoned models. In: Ligatti J, Ou X, Katz J, Vigna G (eds) CCS '20: 2020 ACM SIGSAC conference on computer and communications security, virtual event, USA, November 9–13, 2020. ACM, pp 85–99. https://doi.org/10.1145/3372297.3417253
Pan X, Zhang M, Ji S, Yang M (2020) Privacy risks of general-purpose language models. In: 2020 IEEE symposium on security and privacy (SP). IEEE, pp 1314–1331. https://doi.org/10.1109/SP40000.2020.00095
Papernot N, McDaniel PD, Sinha A, Wellman MP (2018) Sok: security and privacy in machine learning. In: 2018 IEEE European symposium on security and privacy, EuroS&P 2018, London, United Kingdom, April 24–26, 2018. IEEE, pp 399–414. https://doi.org/10.1109/EuroSP.2018.00035
Park J, Tibouchi M (2020) Shecs-pir: somewhat homomorphic encryption-based compact and scalable private information retrieval. In: European symposium on research in computer security. Springer, pp 86–106. https://doi.org/10.1007/978-3-030-59013-0_5
Pillutla VK, Kakade SM, Harchaoui Z (2019) Robust aggregation for federated learning. CoRR arXiv:1912.13445
Qi J, Zhou Q, Lei L, Zheng K (2021) Federated reinforcement learning: techniques, applications, and open challenges. CoRR arXiv:2108.11887
Rahman MS, Khalil I, Atiquzzaman M, Yi X (2020) Towards privacy preserving AI based composition framework in edge networks using fully homomorphic encryption. Eng Appl Artif Intell 94:103737. https://doi.org/10.1016/j.engappai.2020.103737
Reagen B, Choi W-S, Ko Y, Lee VT, Lee H-HS, Wei G-Y, Brooks D (2021) Cheetah: optimizing and accelerating homomorphic encryption for private inference. In 2021 IEEE international symposium on high-performance computer architecture (HPCA). IEEE, pp 26–39. https://doi.org/10.3390/fi13040094
Ren K, Meng QR, Yan SK et al (2021) Survey of artificial intelligence data security and privacy protection. Chin J Netw Inf Secur 7(1):1–10
Ren H, Deng J, Xie X (2021) Grnn: generative regression neural network—a data leakage attack for federated learning. arXiv preprint arXiv:2105.00529
Ren Y, Lin J, Tang S, Zhou J, Yang S, Qi Y, Ren X (2020) Generating natural language adversarial examples on a large scale with generative models. In: De Giacomo G, Catalá A, Dilkina B, Milano M, Barro S, Bugarín S, Lang J (eds) ECAI 2020—24th European conference on artificial intelligence, 29 August–8 September 2020, Santiago de Compostela, Spain, August 29–September 8, 2020—including 10th conference on prestigious applications of artificial intelligence (PAIS 2020), volume 325 of frontiers in artificial intelligence and applications. IOS Press, pp 2156–2163. https://doi.org/10.3233/FAIA200340
Rudin LI, Osher S, Fatemi E (1992) Nonlinear total variation based noise removal algorithms. Physica D Nonlinear phenom 60(1–4):259–268. https://doi.org/10.1016/0167-2789(92)90242-F
Salem A, Zhang Y, Humbert M, Berrang P, Fritz M, Backes M (2019) Ml-leaks: model and data independent membership inference attacks and defenses on machine learning models. In: 26th annual network and distributed system security symposium, NDSS 2019, San Diego, California, USA, February 24–27, 2019. The Internet Society
Seif M, Tandon R, Li M (2020) Wireless federated learning with local differential privacy. In: 2020 IEEE international symposium on information theory (ISIT). IEEE, pp 2604–2609. https://doi.org/10.1109/ISIT44484.2020.9174426
Sharif M, Bhagavatula S, Bauer L, Reiter MK (2016) Accessorize to a crime: real and stealthy attacks on state-of-the-art face recognition. In: Weippl

ER, Katzenbeisser S, Kruegel C, Myers AC, Halevi S (eds) Proceedings of the 2016 ACM SIGSAC conference on computer and communications security, Vienna, Austria, October 24–28, 2016, pp 1528–1540. ACM. https://doi.org/10.1145/2976749.2978392

Shokri R, Shmatikov V (2015) Privacy-preserving deep learning. In: Ray I, Li N, Kruegel C (eds) Proceedings of the 22nd ACM SIGSAC conference on computer and communications security, Denver, CO, USA, October 12–16, 2015. ACM, pp 1310–1321. https://doi.org/10.1145/2810103.2813687

Shokri R, Stronati M, Song C, Shmatikov V (2017) Membership inference attacks against machine learning models. In: 2017 IEEE symposium on security and privacy (SP). IEEE, pp 3–18. https://doi.org/10.1109/SP.2017.41

Smith V, Chiang C-K, Sanjabi M, Talwalkar AS (2017) Federated multi-task learning. In: Guyon I, von Luxburg U, Bengio S, Wallach HM, Fergus R, Vishwanathan SVN, Garnett R (eds) Advances in neural information processing systems 30: annual conference on neural information processing systems 2017, December 4–9, 2017, Long Beach, CA, USA, pp 4424–4434

Song L, Haoqi W, Ruan W, Han W (2020) Sok: training machine learning models over multiple sources with privacy preservation. CoRR arXiv:2012.03386

Song C, Ristenpart T, Shmatikov V(2017) Machine learning models that remember too much. In: Thuraisingham BM, Evans D, Malkin T, Xu D (eds) Proceedings of the 2017 ACM SIGSAC conference on computer and communications security, CCS 2017, Dallas, TX, USA, October 30–November 03, 2017, pp 587–601. ACM. https://doi.org/10.1145/3133956.3134077

Song C, Shmatikov V (2020) Overlearning reveals sensitive attributes. In: 8th international conference on learning representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020. OpenReview.net

Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15(1):1929–1958

Stella H, Youyang Q, Bruce G, Longxiang G, Jianxin L, Yong X (2021) Dp-gan: differentially private consecutive data publishing using generative adversarial nets. J Netw Comput Appl 185:103066. https://doi.org/10.1016/j.jnca.2021.103066

Suciu O, Marginean R, Kaya Y, Daumé III H, Tudor D (2018) When does machine learning fail? Generalized transferability for evasion and poisoning attacks. In: Enck W, Felt AP (eds) 27th USENIX security symposium, USENIX security 2018, Baltimore, MD, USA, August 15–17, 2018, pp 1299–1316. USENIX Association

Sun Z, Kairouz P, Suresh AT, McMahan HB (2019) Can you really backdoor federated learning? CoRR arXiv:1911.07963

Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow IJ, Fergus R (2014) Intriguing properties of neural networks. In: Bengio Y, LeCun Y (eds) 2nd international conference on learning representations, ICLR 2014, Banff, AB, Canada, April 14–16, 2014, conference track proceedings

Tramèr F, Zhang F, Juels A, Reiter MK, Ristenpart T (2016) Stealing machine learning models via prediction apis. In: Holz T, Savage S (eds) 25th USENIX security symposium, USENIX security 16, Austin, TX, USA, August 10–12, 2016. USENIX Association, pp 601–618

Triastcyn A, Faltings B (2019) Federated learning with Bayesian differential privacy. In: 2019 IEEE international conference on Big Data (Big Data). IEEE, pp 2587–2596. https://doi.org/10.1109/BigData47090.2019.9005465

Vepakomma P, Gupta O, Swedish T, Raskar R (2018) Split learning for health: distributed deep learning without sharing raw patient data. CoRR arXiv:1812.00564

Wang W, Wang X, Feng D, Liu J, Han Z, Zhang X (2014) Exploring permission-induced risk in android applications for malicious application detection. IEEE Trans Inf Forensics Secur 9(11):1869–1882

Wang W, Song J, Guangquan X, Li Y, Wang H, Chunhua S (2021) Contractward: automated vulnerability detection models for ethereum smart contracts. IEEE Trans Netw Sci Eng 8(2):1133–1144

Wang B, Gong NZ (2018) Stealing hyperparameters in machine learning. In: 2018 IEEE symposium on security and privacy, SP 2018, proceedings, 21–23 May 2018, San Francisco, California, USA, pp 36–52. IEEE Computer Society. https://doi.org/10.1109/SP.2018.00038

Wang Y, Han Y, Bao H, Shen Y, Ma F, Li J, Zhang X (2020) Attackability characterization of adversarial evasion attack on discrete data. In: Gupta R, Liu Y, Tang J, Aditya Prakash B (eds) KDD '20: the 26th ACM SIGKDD conference on knowledge discovery and data mining, virtual event, CA, USA, August 23–27, 2020. ACM, pp 1415–1425. https://doi.org/10.1145/3394486.3403194

Wang H, Sreenivasan K, Rajput S, Vishwakarma H, Agarwal S, Sohn J, Lee K, Papailiopoulos DS (2020) Attack of the tails: yes, you really can backdoor federated learning. In: Larochelle H, Ranzato M, Hadsell R, Balcan M-F, Lin H-T (eds) Advances in neural information processing systems 33: annual conference on neural information processing systems 2020, NeurIPS 2020, December 6–12, 2020, virtual

Warnat-Herresthal S, Schultze H, Shastry KL et al (2021) Swarm learning for decentralized and confidential clinical machine learning. Nature 594:265–270. https://doi.org/10.1038/s41586-021-03583-3

Wei K, Li J, Ding M, Ma C, Yang HH, Farokhi F, Jin S, Shi TQS, Poor HV (2020) Federated learning with differential privacy: algorithms and performance analysis. IEEE Trans Inf Forensics Secur 15:3454–3469. https://doi.org/10.1109/TIFS.2020.2988575

Weng J, Weng J, Zhang J, Li M, Zhang Y, Luo W (2021) Deepchain: auditable and privacy-preserving deep learning with blockchain-based incentive. IEEE Trans. Dependable Secur Comput 18(5):2438–2455

Weng H, Zhang J, Xue F, Wei T, Ji S, Zong Z (2020) Privacy leakage of real-world vertical federated learning. CoRR arXiv:2011.09290

Wu T, Zhao C, Zhang Y-JA (2021) Privacy-preserving distributed optimal power flow with partially homomorphic encryption. IEEE Trans Smart Grid. https://doi.org/10.1109/TIFS.2017.2787987

Wu C, Wu F, Cao Y, Huang Y, Xie X (2021) Fedgnn: federated graph neural network for privacy-preserving recommendation. CoRR arXiv:2102.04925

Xiangrui X, Li Y, Yuan C (2020) "identity bracelets" for deep neural networks. IEEE Access 8:102065–102074

Xian X, Wang X, Ding J, Ghanadan R (2020) Assisted learning: a framework for multi-organization learning. In: Larochelle H, Ranzato M, Hadsell R, Balcan M-F, Lin H-T(eds) Advances in neural information processing systems 33: annual conference on neural information processing systems 2020, NeurIPS 2020, December 6–12, 2020, virtual

Xie C, Huang K, Chen P-Y, Li B (2020) DBA: distributed backdoor attacks against federated learning. In: 8th international conference on learning representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020. OpenReview.net

Yang Q, Liu Y, Chen T, Tong Y (2019) Federated machine learning: concept and applications. ACM Trans Intell Syst Technol 10(2):12:1-12:19. https://doi.org/10.1145/3298981

Yao AC (1982) Protocols for secure computations. In: 23rd annual symposium on foundations of computer science (sfcs 1982). IEEE, pp 160–164. https://doi.org/10.1109/SFCS.1982.38

Yeom S, Giacomelli I, Fredrikson M, Jha S (2018) Privacy risk in machine learning: analyzing the connection to overfitting. In: 31st IEEE computer security foundations symposium, CSF 2018, Oxford, United Kingdom, July 9–12, 2018, pp 268–282. IEEE Computer Society. https://doi.org/10.1109/CSF.2018.00027

Yin D, Chen Y, Ramchandran K, Bartlett PL (2018) Byzantine-robust distributed learning: towards optimal statistical rates. In: Dy JG, Krause A (eds) Proceedings of the 35th international conference on machine learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10–15, 2018, volume 80 of proceedings of machine learning research. PMLR, pp 5636–5645

Yin H, Mallya A, Vahdat A, Alvarez JM, Kautz J, Molchanov P (2021) See through gradients: image batch recovery via gradinversion. arXiv preprint arXiv:2104.07586

Yousuf H, Lahzi M, Salloum SA, Shaalan K (2021) Systematic review on fully homomorphic encryption scheme and its application. Recent Adv Intell Syst Smart Appl. https://doi.org/10.1007/978-3-030-47411-9_29

Yulei W (2021) Robust learning-enabled intelligence for the internet of things: a survey from the perspectives of noisy data and adversarial examples. IEEE Internet Things J 8(12):9568–9579. https://doi.org/10.1109/JIOT.2020.3018691

Zeng Y, Dai T, Chen B, Xia S-T, Lu J (2021) Correlation-based structural dropout for convolutional neural networks. Pattern Recognit. https://doi.org/10.1016/j.patcog.2021.108117

Zeng G, Qi F, Zhou Q, Zhang T, Hou B, Zang Y, Liu Z, Sun M (2020) Openattack: an open-source textual adversarial attack toolkit. CoRR arXiv:2009.09191

Zhang Y, Jia R, Pei H, Wang W, Li B, Song D (2020) The secret revealer: generative model-inversion attacks against deep neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 253–261

Zhang H, Zhou H, Miao N, Li L(2019) Generating fluent adversarial examples for natural languages. In: Korhonen A, Traum DR, Màrquez L (eds) Proceedings of the 57th conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28– August 2, 2019, volume 1: long papers. Association for Computational Linguistics, pp 5564–5569. https://doi.org/10.18653/v1/p19-1559

Zhao B, Mopuri KR, Bilen H (2020) idlg: improved deep leakage from gradients. arXiv preprint arXiv:2001.02610

Zheng Q, Chen S, Long Q, Su W (2021) Federated f-differential privacy. In: International conference on artificial intelligence and statistics. PMLR, pp 2251–2259

Zhu L, Han S (2020) Deep leakage from gradients. In: Federated learning. Springer, pp 17–31. https://doi.org/10.1007/978-3-030-63076-8_2

Zong B, Song Q, Min MR, Cheng W, Lumezanu C, Cho D, Chen H (2018) Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In: 6th international conference on learning representations, ICLR 2018, Vancouver, BC, Canada, April 30–May 3, 2018, conference track proceedings. OpenReview.net

Publisher's Note