MIA-Leak: Exploring Membership Inference Attacks in Federated Learning Systems

Chengcheng Zhu, Jiale Zhang(✉), Xiang Cheng, Weitong Chen, and Xiaobing Sun

School of Information Engineering, Yangzhou University, Yangzhou 225009, China
mx120220554@stu.yzu.edu.cn, {jialezhang,wtchen,xbsun}@yzu.edu.cn

Abstract. Federated learning has achieved significant success in both academia and industry scenarios since it can train a joint model among unbalanced datasets while protecting the training data privacy. Recent research has shown that, by inferring whether a given data record belongs to the model's training dataset, the membership information could be leaked by malicious participants. However, when deploying member inference attacks in federated learning, the core problem is how to obtain the membership inference attack data with the same distribution as the training data. In this paper, to tackle this problem, we mainly focus on exploring membership inference attacks in federated learning based on the data augmentation method. Specifically, we present two types of membership inference attacks based on the generative adversarial nets, in which a class-level attack aims to infer the global model and a user-level attack tries to focus on a specific victim. We conduct extensive experiments to evaluate the effectiveness of our proposed two types of membership inference attacks on two benchmark datasets. The experimental results have shown that both class-level and user-level attacks can achieve extraordinary attack accuracy on federated learning.

Keywords: Federated learning · Membership inference · Generative adversarial nets · Privacy leakage

1 Introduction

To prevent privacy leakage, federated learning has been introduced into distributed computing systems due to its specific privacy-preserving structure,

This work is partially supported by the National Natural Science Foundation of China (62206238), Natural Science Foundation of Jiangsu Province (Grant No. BK20220562), Natural Science Foundation of Jiangsu Higher Education Institutions of China (Grant No. 22KJB520010), Future Network Scientific Research Fund Project (FNSRFP-2021-YB-47), Yangzhou City-Yangzhou University Science and Technology Cooperation Fund Project (YZ2021158).

W. Meng and W. Li (Eds.): BlockTEA 2022, LNICST 498, pp. 140–154, 2023.
https://doi.org/10.1007/978-3-031-31420-9_9

where the central server trains a joint global model via uploaded gradients from participants instead of the raw data [1–3]. Unlike the traditional centralized distributed computing system which aggregates raw data from participants, federated learning distributes the training models to local devices, and only transmits the parameters between the central server and participants to update the global model [4]. Hence, data privacy can be large extend preserved, while participants keep their local datasets on their own end.

However, recent researches reveal that the federated learning framework is vulnerable to various inference attacks, such as membership inference [5], representatives inference [6], properties inference [7], and gradients inference [8]. Among these inference attacks, membership inference is one of the most powerful active attacks against private training datasets. Shokri et al. [9] first proposed the membership inference attack against machine learning models through a black-box API, which reveals the fact that membership information can be leaked by distinguishing the difference between model predictions from training and non-training inputs. Notably, the purpose of membership inference attacks is to determine whether a certain data sample is used to train the model (Fig. 1).

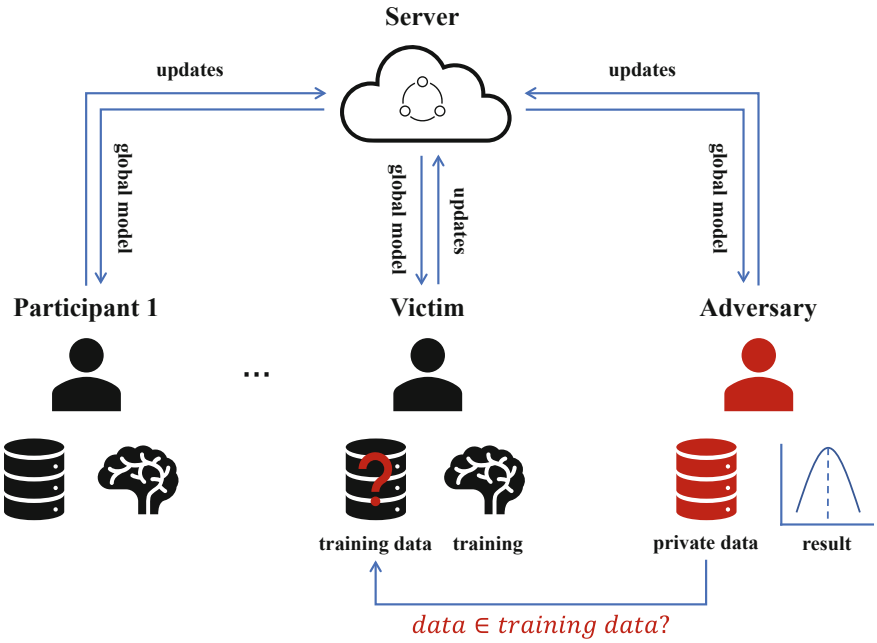


Fig. 1. Membership inference attack in federated learning.

Membership inference attacks are ongoing threats to the data of participants in federated learning, which lead to user privacy leakage issues [10]. For example, a membership inference attack can be initialized by an attacker who aims to

reveal the medical records of specific patients [11]. To initial certain attacks, the attacker train a binary model and take the confidence score vector of the Victims as input [9]. In federated learning, the attacker may play the role of a benign participant or central server to perform the attacks. The attackers can either join the federated learning as a participant to observe the latest model parameters from the server and perform the membership inference attacks or play the server role to collect the parameters uploaded from participants to modify the model [12]. The user data is exposed to attackers in either scenario, where existing defense mechanisms have little effect on membership inference attacks.

Therefore, the defense mechanism on membership inference attacks attracts more attention in the research area. It is worth noticing that Generative adversarial networks (GAN) could be the reason for the recent success of membership inference attacks [13]. GAN is designed and widely applied in the computer vision area due to its excellent data augmentation features. By using the discriminator and generator in GAN, attackers may leverage the parameters from the training models to generate the fake samples or obtain the data from other participants which has the same distribution in the training dataset [14]. State-of-the-art membership inference attack defense mechanisms in federated learning focus on preventing attackers obtain unprotected model parameters such as multi-party aggregation [15], homomorphic cryptosystem [16] and differential privacy preserving [17]. It has been proven that adding a crafted noise vector to the attack model can successfully maximize the effects of privacy-utility tradeoffs [18].

In this paper, we firstly investigate two types of membership inference attacks based on GAN from the class-level and user-level perspectives, and further propose the defense mechanisms of the attack cases. The purpose of the class-level membership inference attack is to train a binary classification model which can infer the information from the global model of federated learning. To achieve the attack, attackers must play the role of local participants and overcome the insufficient attack data which causes a low accuracy problem of the binary classification model. Therefore, GAN is used to increase and fill the diversity of the attack data, where the class-level attack can be successfully launched. Unlike the class-level attacks which aim to reveal the membership information from the global model, the user-level attack aims to infer the information from a specific participant in federated learning. We make an assumption that the attacker is also a local participant but does not need the knowledge of the training datasets. The attacker relies on the local-deployed GAN to generate high-quality fake samples to launch the attack. To defend against the proposed attacks, we further propose the defence mechanism namely DefMIA which focuses on local attackers and applies the adversarial samples against the membership inference attacks in federated learning.

The main contributions of this paper are as follows.

– We first demonstrate that constructing membership inference attacks in federated learning faced the problem of lacking attack training data. Then, we point out that GAN is a promising technology to generate fake samples with the same distributions as participants' training data.

- We present two types of membership inference attacks, in which class-level attack aims to attack the global model and user-level aims to attack the specific victim. We explore the weakness point of the current federated learning and initialize the attacks which are enhanced by GAN. We prove that both of the attacks are efficient and can achieve excellent accuracy when attacking federated learning.
- Exhaustive experimental evaluations on two benchmark datasets show that both class-level and user-level attacks achieve extraordinary attack accuracy on federated learning. However, the attack accuracy reduced dramatically after we explore the DefMIA method in the federated learning system.

The rest of this paper is organized as follows. Section 2 reviews the related works. The membership inference attacks in federated learning are analyzed in Sect. 3. Follow by the experiment and evaluation in Sect. 4. Finally, Sect. 5 summaries the whole paper and gives future directions.

2 Related Work

2.1 Attacks in Federated Learning

Compared to traditional machine learning approaches, federated learning does not require participants to upload their local raw data to the central server. Therefore, federated learning has its native advantage of privacy-preserving, where participants only need to upload the parameters of their local trained model to the central server [19]. Although federated learning can efficiently handle unbalanced data while protecting training data privacy, security issues still exist. Firstly, the central server does not know the local training data, hence, the server cannot verify the uploaded parameters are correct or not. Furthermore, parameters can be easily leaked by the malicious server or external adversaries, which leads to privacy leaking problems.

Attackers target the aforementioned drawbacks in federated learning, and launch the different types of attacks such as model inversion attack [20], poisoning attack [21,22], and adversarial attack [23]. The types attack can be classified by different purposes, which are confidentiality, integrity, and availability [24]. The purpose of confidentiality is to protect the sensitive data from users. Confidentiality attacks are not only trying to steal the local training data, but also trying to expose the privacy data or infer the training models [3]. Attackers in confidentiality attacks will not interfere the training progress and the training models, they just act like participants to initialize the attacks. Integrity attacks aims to destroy the model outputs by poisoning the model. Typical integrity attacks such as label-flipping [25] and backdoor attack [26], which mislead the target model to a specific direction which given by the attackers [27]. Availability attack aims to attack the availability in classification including errors, false positives and false negatives [28]. The main purpose of availability attack is to make the target model in federated learning unusable.

2.2 Membership Inference

Membership inference refers to attacks on machine learning models to determine the certain data is from the training set or not [11]. Membership inference attacks can target both traditional machine learning and federated learning which is a severe security threat to user information [29]. For traditional machine learning membership inference attacks, Shokri et al. [9] designed a shadow model which can simulate the target model to give results. Additionally, attackers also set up a testing dataset that has the same distribution as the training dataset to train the inference model.

Membership inference against federated learning systems has been introduced by Nasr et al. [12]. The attackers can either play the role of server who collects the uploaded parameters from participants or play the role of participants to obtain the aggregated models. Attackers also can launch the attacks actively, where the malicious server and participants can generate adversary data to realize the attacks [2].

2.3 Defense Proposals

Previous researches on attacks show that federated learning is vulnerable to membership inference attacks, where the adversarial example can mislead the prediction results of the attack models [30]. To defend the adversarial attacks, MemGuard [18] is proposed, which applies formal utility-loss guarantees to defend membership inference attacks under the black-box setting. The carefully designed noise has been added into the confidence score vector of the model, where the attack models can be misleading to a random result. In summary, Federated learning is vulnerable to membership inference attacks since the parameters can be easily observed by malicious participants. To tackle this problem, the proposed DefMIA adds the crafted noises to the model, where the main challenge is to restrict the loss of the target model in multiple training iterations.

3 Membership Inference Attacks in Federated Learning

3.1 Threat Models

Adversary's Objectives. The objective of the threat model is to obtain indirect information about the target models. Therefore, the classification task of the threat model will be measured by the following metrics: 1) membership inference accuracy: which is the performance of the classification on the target dataset; 2) main task accuracy: which represents the performance of the global model of the federated learning. The threat model expects to achieve the high performance of the membership inference accuracy and keep the high performance of the main task accuracy.

Adversary's Observations. The threat model will be set under a white-box setting, where the adversary observes can initialize the inference attack. The attacker can observe the latest model which the server distributed to every participant each iteration in federated learning. Therefore, the attack obtains every detail of the distributed global model, including structure, learning algorithm L, and the parameter θ of the model. This information can be used to train a GAN to generate the samples and initialize the membership inference attack.

Adversary's Capabilities. The capabilities of the adversary will be listed in this part. The adversary can: 1) obtaining details of the global model for each training iteration, 2) controlling the local training and local data as a participant, 3) modifying the hyper-parameters of the model, 4) updating the select parameters randomly. But cannot: 1) observing the parameters from other participants because the global model averaging all the uploaded gradients on the server, 2) accessing the data from other participants.

3.2 Case1: Class-Level Attack

Overview of Attack. The goal of the class-level attack is to wreak the confidentiality of the models and obtain the membership information, where the attacker plays the role of a participant in federated learning and passively trains the attack model. The attack model will be considered as a white-box setting to the attacker since it can observe all the model structures and gradients of each layer of the model. Based on the aforementioned white-box setting, the attacker joins the federated learning as a participant in the proposed class-level attack method, where the attacker train a binary classification model with GAN-generated data to distinguish members from non-members in the training data. To achieve this target, We propose a two-phase GAN enhanced membership inference attack, which includes data augmentation and attack model training phases.

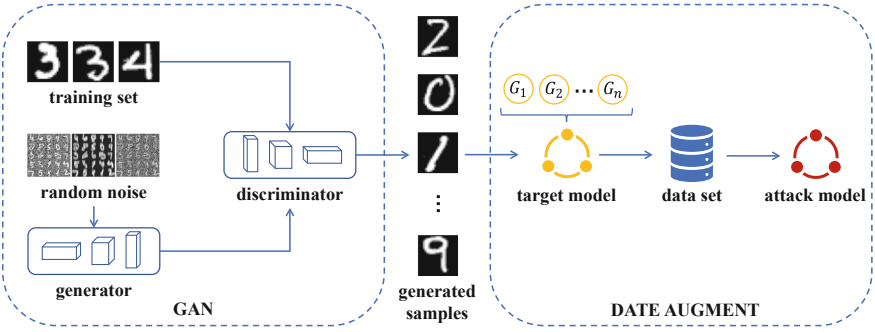


Fig. 2. Class-level membership inference attack in federated learning.

Figure 2 illustrates the architecture of the class-level membership inference attack in federated learning. We leverage the true labels and prediction results from the target model to train the attack model, where the attack model can learn the distribution of prediction to distinguish members from non-members of the target model. We define $f_{target}()$ as the target model, and D^{train}_{target} as the training set of the target model, where the labeled data $(x\{i\}, y\{i\})_{target}$ belongs to D^{train}_{target}. $(x\{i\}$ represents the target model input, and $(y\{i\}$ represents the true label of $(x\{i\}$ which is from the label classes c_{target}. The prediction of target model $Y = f_{target}(x)$ is computed and applied to train the attack model, due to Y is highly dependent on the true label. Therefore, the attack model can distinguish that if the data is from the training dataset of the target model or not.

However, the lack of diversity in D^{train}_{target} is challenging. The attacker only has limited data as a participant, and a data augmentation model is designed to overcome the challenge which uses GAN to generate new data. In the data augmentation phase, the GAN sets the target model as the discriminator and generates new data which follows the same distribution as the original training set. The enhanced dataset will be used to train the attack model which will have the capacity to launch the membership inference attack with high inference accuracy.

Augment Training Data with GAN. We apply GAN to overcome the problem of low training data diversity in class-level attacks, which can generate the extra data x_{gen} with the same distribution of original data. We will detail the GAN in this part with model structure and augmentation progress. As shown in Fig. 2, the generator G is initialized by a random noise $g(z; \theta_G)$, in the meantime, the discriminator D is initialized by the target federated learning model $f(x; \theta_D)$. The goal of the GAN is to generate the data to increase the training set which shares the same distribution of the original data. Hence, the discriminator will lead the generator to generate the training data. After certain training iterations, the quality generated data x_{gen} is close to the original data. The generating process can summarize as Eq. 1, where x_i indicates the original data and x_{gen} indicates the generated data.

$$\min_{\theta_G} \max_{\theta_D} \left(\sum_{i=1}^{n_+} \log f(x_i; \theta_D) + \sum_{j=1}^{n_-} \log(1 - f(g(x_{gen}; \theta_G); \theta_D))\right) \tag{1}$$

In the augmentation phase, the GAN firstly initializes the generator G and apply the target model $f_{target}()$ to initialize the discriminator D. Then, D will determine the generated samples x_{gen} is from original dataset or not until D cannot distinguish x_{gen} is generated by G. There are two ways to label the x_{gen} which are labels that can be recognized by a person or run the target model to label the data. In the context of federated learning, we can apply the target global model to label the generated samples x_{gen} easily. In the classification phase, the original data and generated data are combined as one training dataset to train the attack model.

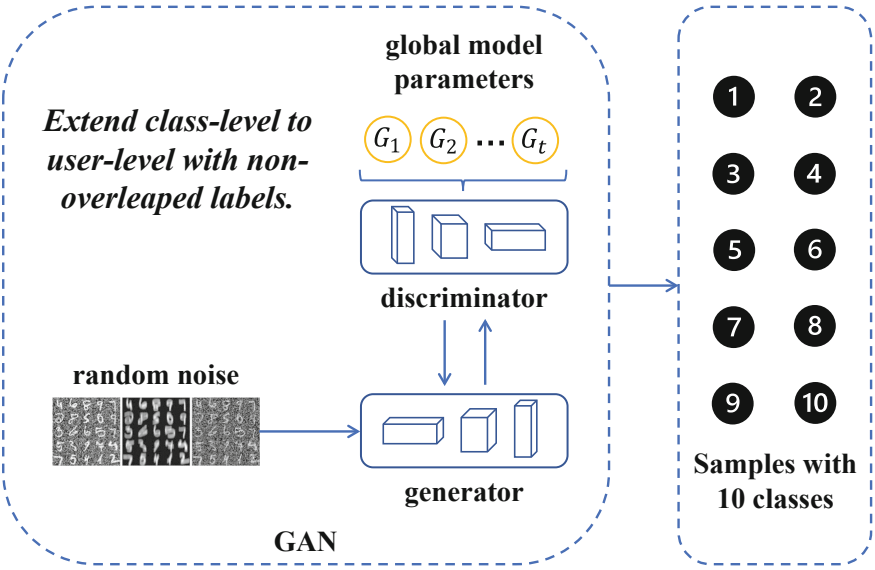


Fig. 3. User-level membership inference attack in federated learning.

Train Attack Model. The object of the attack model is leveraging the GAN to reveal the data distribution of the target model and apply the generated data to train the attack model. As shown in Fig. 2, the training data is integrated by original and generated data after the data augmentation phase. The integrated data includes predictions, true labels, and member states, which will be learnt by the attack model. The final training data of the attack model $x\{i\}_{attack}$ consists of prediction, true label and two attributes "in" or "out" which indicates a member or non-member of the target model.

According to GAN, we query the predictions Y from D^{train}_{target}. Then, the dataset D^{gen}_{target} with $(record, label)$ will be generated. Therefore, the enriched training set D^{train}_{target} is the combination of D^{ori}_{target} and D^{gen}_{target}.

The object of the classification model is the classify the member state through the distribution of prediction around the true label. Therefore, the GAN enhanced membership inference attack model is trained based on the labels. We further divide training dataset D^{train}_{attack} into n categories, where $f_{attack}()$ represents the attack mode and the input of the model x_{attack} is $(y, Y, in/out)$. Each category will be used to train one attack model, where the attack model can classify the membership state for giving a certain data record. The reasons for launching a successful membership inference attack are generalizability and training data diversity, where GAN is a good way to increase the diversity due to the outstanding performance on data augmentation.

3.3 Case2: User-Level Attack

Assumptions. To initialize a user-level attack, there some assumptions need to be done in the first place. As done in previous researches, participants need to declare the labels of the local training data to the server before they start training, which will not expose the local training data. The reason is that the label is not a reflection of the data features. The proposed scheme assumes that labels of the data will not be overlapped by owned participants. Take MNIST for example, participant P_1 has labeled "0", "1" and all other participants will not have the data with the same labels. And the label "1" is not reflected in the handing writing features of the digit "1". The purpose of this assumption is facilitating to compare the results of the attack model with the previously declared information to launch the membership inference attack. For instance, medical information analysis will integrate the diagnosis data from different hospitals with different labels to enrich the dataset. Therefore, federated learning may have more diagnose class labels, and data with the same labels should follow the same distribution from different hospitals.

Attack Construction. The user-level attack is illustrated in Fig. 3. Let N participants join the federate learning, and V represents the victim. A represents the attacker who also join the federated learning as a participant. After k training iterations, A and V have the same global model with parameter θ_d downloaded from the central server. Normally, A and V will use the global model and local data to training a new local model, and upload the updated parameters θ_u to the server. The federated learning central server will firstly average all the parameter updates from all participants, then update the global model. Therefore, it is hard for A to launch the membership inference on the V directly. We take the same structure of GAN in Sect. 3.2, where the θ_d is used to train the discriminator D, and generator G can generate the fake data which similar to real data. We use the generated data to train a binary classifier, once we obtained the target dataset, the classifier can predict the result as "in" which means the results are consistent with the declaration information, otherwise mark as "out". As shown in Fig. 3, we train the classifier with the generated data with all labels. As we described in Sect. 3.2, we can choose inference algorithm after analyzing the generated data. In our experiment, we choose MNIST as dataset and CNN as the classifier.

4 Experimental Evaluation

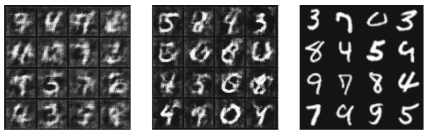
This section firstly introduce the dataset and experimental setup, then we evaluate the performance of class-level attacks, user-level attack and defense methods.

4.1 Datasets and Experimental Setup

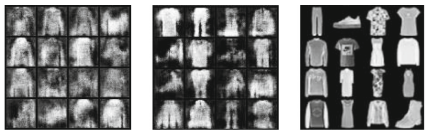
Dataset. We apply two famous datasets MNIST and Fashion MNIST to evaluate our experiment.

MNIST is a handwritten digits dataset that consists of 60000 training data and 10000 testing data from digits "0" to "9". Each image is a 28×28 image with white text on black background [31].

Fashion MNIST (F-MNIST) is a Zalando's article images dataset which consists of 60,000 training data and 10000 testing data with 10 classes of clothes, pants, and shoes, etc. Each example is also a 28×28 grayscale image [32].



(a) Reconstructed samples of MNIST dataset



(b) Reconstructed samples of F-MNIST dataset

Fig. 4. Data reconstruction results on MNIST and F-MNIST datasets based on GAN.

Experimental Setup. All the experiments are conducted on an Ubuntu 16.04 Linux server with 32 GB RAM and NVidia Quadro P4000 GPU. All the codes are written under Python 3.6 with Pytorch and Tensorflow with Keras framework.

Class-Level Attack Configuration: We set 100 participants in federated learning, and each of them has 60 MNIST samples. The participants will train 10 epochs and the learning rate is set to 0.001.

User-Level Attack Configuration: We apply a basic CNN model for the membership inference model on both datasets. The model of MNIST includes two convolutional layers and two dense layers. The kernel size is set to 5×5. The model for F-MNIST has four convolutional layers and the kernel size is set to 3×3. Each participant will train MNIST for 30 epochs and the learning rate is 0.01. For F-MNIST, each participant will train 60 epochs with a learning rate of 0.0001. All the experiments are run for 400 communication rounds of federated learning.

Defense Configuration: The number of participants of the target federated learning model is set to five, one of the participants is considered as an attacker. Each of the participants has 12000 training samples and train the model for 10 epochs with a learning rate of 0.01. The class-level and user-level attack models will be implemented in the experiment. The settings of the defense model are based on the white-box scenario, and we assume that the defense model does not share the same network structure as the attack models.

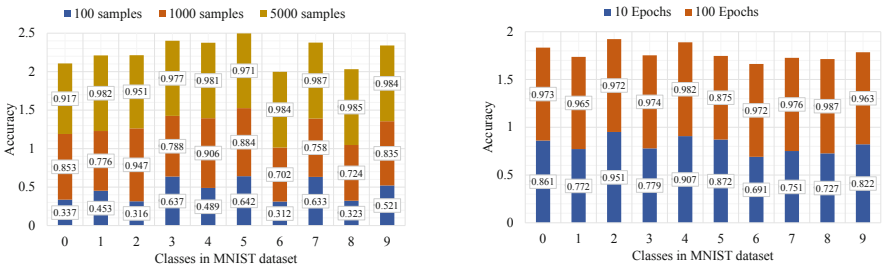
4.2 Effectiveness of Data Augmentation

To evaluate the effectiveness of the data augmentation by using GAN in federated learning. The reconstruction process is being monitored, where the number of

participants and samples remains unchanged. The generator G will generate the samples with 100 lengths and reshape them to 28×28. The generator will start to generate samples when the accuracy of the global model reaches 0.93.

Figure 4 shows that sample visualization during different iterations. The comparison of the reconstruction results between 150 iterations, 300 iterations, and original samples of the MNIST dataset are shown on the left, and the reconstruction results between 200 iterations, 400 iterations, and original samples of the F-MNIST dataset are shown on the right. The generated samples are getting more clear once the iteration increases. Hence, GAN can successfully simulate the original samples of all participants.

4.3 Evaluation of Class-Level Attack



(a) Performance under different generated data size

(b) Performance under different learning epochs

Fig. 5. Evaluation results of class-level inference attack under different variables.

The performance of the class-level attack is measured by prediction accuracy and recall metrics. The prediction accuracy represents the attack accuracy directly, and the coverage of the attack method is measured by the recall. To evaluate the prediction accuracy, the training set and testing set have been reshuffled to train and test the attack model. We also set the random conjecture accuracy as 0.5 for comparison purposes. Table 1 shows the performance of the proposed membership inference attack model. By increasing the training set from 100 samples to 5000 samples, the proposed membership inference attack achieves 97.63% test accuracy on the MNIST dataset, and the recall is 88.36% We also use the F1 score to measure the ability of classification. The result shows that the F1 score reaches 0.94, which means the proposed attack method has good generalization and membership inference capability.

Figure 5(a) shows how the different size of training data affects the prediction of membership inference attack method on each class of MNIST dataset. Results show that the overall accuracy is 52.9% when the attacker only has 100 data samples. However, the accuracy has increased dramatically to 97.9% after the training set is reached 1000 s samples. To explore how the impact of the

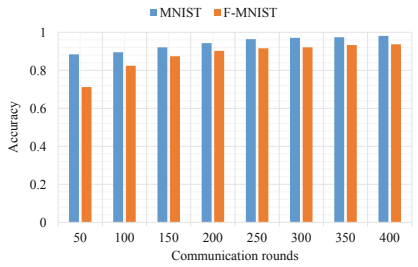
Table 1. Performance of class-level inference attack

Attack Accuracy	Recall Ratio	F1 Score
97.63%	88.36%	0.94

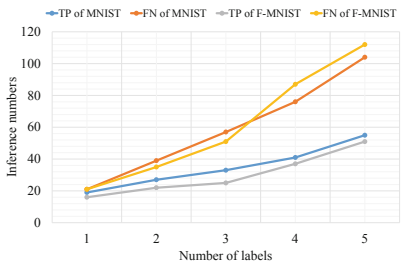
settings of the target model affects the accuracy of the proposed attach model, we implement experiments on different learning epochs. Figure 5b show that the attack accuracy of 100 epochs is much higher than 10 epochs, and the attack precision is close to 1, which means overfitting could lead to the model being more vulnerable to membership inference attacks.

4.4 Evaluation of User-Level Attack

According to Fig. 6(a), the accuracy of models achieves 99.45% and 93.71% on MNIST and F-MNIST, respectively, which is high enough to complete all the classification tasks on the testing dataset. In the meantime, the attacker also generates enough samples via a local deployed GAN to train the attack model. After the membership inference attack, we further measure the attack effectiveness from the label point of view. Figure 6(b) shows the attack effectiveness on the two datasets, where TP represents the true positive and FN represents the false negative. We compare the number of classes held by each participant, and we assume that the victim has more than one class of the dataset, which may affect the membership inference. We compare the effectiveness of membership inference between different numbers of classes held by the victim, where the effectiveness of the attack model getting worse the more classes held by the victim.



(a) Performance of benchmark federated learning



(b) Performance of user-level inference attack

Fig. 6. Evaluation results of benchmark federated learning and user-level inference attack.

To elaborate on the advantage of the proposed user-level attack, we compare our user-level attack with the active inference attacks using the SGA algorithm

proposed by Nasr et al. [12]. The attack accuracy of SGA can reach about 76% on the F-MNIST dataset, which is close to our experiment settings while the victim holds just one class of the data of F-MNIST. However, the novelty of our user-level attack is the attack objective. Nasr et al. [12] claim that their attack methods against all the participants in federated learning, which means their attack only aim the whole training set. However, the proposed user-level attack method can launch an inference attack on a specific victim who joins the federated learning.

5 Summary and Future Work

In this paper, we give a comprehensive study on exploring membership inference attacks and mitigation methods in federated learning systems. We firstly proposed two types of membership inference attacks based on GAN, which are class-level attack and user-level attack. For class-level attack, GAN is used to increase and fill the diversity of the attack data, so as to increase the accuracy of the binary classification attack model. For user-level attack, it aims to infer the membership information from a specific participant, which is a more deep-level attack method. The experimental results have shown that both class-level and user-level attacks can achieve extraordinary attack accuracy on federated learning. In future work, we plan to explore the membership inference attacks in an untrusted federated learning environment, where a part of participants tries to jeopardize the global model through poisoned local model updates. In this situation, how to guarantee the attack accuracy of inferring the membership information of certain data records is becoming a big challenge.

References

1. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial Intelligence and Statistics, pp. 1273–1282. PMLR (2017)
2. Li, T., Sahu, A.K., Talwalkar, A., Smith, V.: Federated learning: challenges, methods, and future directions. IEEE Signal Process. Mag. **37**(3), 50–60 (2020)
3. Yang, Q., Liu, Y., Chen, T., Tong, Y.: Federated machine learning: concept and applications. ACM Trans. Intell. Syst. Technol. (TIST) **10**(2), 1–19 (2019)
4. Sattler, F., Wiedemann, S., Müller, K.R., Samek, W.: Robust and communication-efficient federated learning from non-IID data. IEEE Trans. Neural Netw. Learn. Syst. **31**(9), 3400–3413 (2019)
5. Truex, S., Liu, L., Gursoy, M.E., Yu, L., Wei, W.: Demystifying membership inference attacks in machine learning as a service. IEEE Trans. Serv. Comput. **14**, 2073–2089 (2019)
6. Wang, Z., Song, M., Zhang, Z., Song, Y., Wang, Q., Qi, H.: Beyond inferring class representatives: user-level privacy leakage from federated learning. In: IEEE INFOCOM 2019-IEEE Conference on Computer Communications, pp. 2512–2520. IEEE (2019)

7. Melis, L., Song, C., De Cristofaro, E., Shmatikov, V.: Exploiting unintended feature leakage in collaborative learning. In: 2019 IEEE Symposium on Security and Privacy (SP), pp. 691–706. IEEE (2019)
8. Zhu, L., Liu, Z., Han, S.: Deep leakage from gradients. Adv. Neural. Inf. Process. Syst. **32**, 14774–14784 (2019)
9. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: 2017 IEEE Symposium on Security and Privacy (SP), pp. 3–18. IEEE (2017)
10. Chen, J., Zhang, J., Zhao, Y., Han, H., Zhu, K., Chen, B.: Beyond model-level membership privacy leakage: an adversarial approach in federated learning. In: 2020 29th International Conference on Computer Communications and Networks (ICCCN), pp. 1–9. IEEE (2020)
11. Hayes, J., Melis, L., Danezis, G., De Cristofaro, E.: LOGAN: membership inference attacks against generative models. arXiv preprint arXiv:1705.07663 (2017)
12. Nasr, M., Shokri, R., Houmansadr, A.: 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), pp. 739–753. IEEE (2019)
13. Goodfellow, I., et al.: Generative adversarial nets. Adv. Neural Inf. Process. Syst. **27** (2014)
14. Qu, Y., Yu, S., Zhang, J., Binh, H.T.T., Gao, L., Zhou, W.: GAN-DP: generative adversarial net driven differentially privacy-preserving big data publishing. In: ICC 2019–2019 IEEE International Conference on Communications (ICC), pp. 1–6. IEEE (2019)
15. Jónsson, K.V., Kreitz, G., Uddin, M.: Secure multi-party sorting and applications. IACR Cryptol. ePrint Arch. **2011**, 122 (2011)
16. Aono, Y., Hayashi, T., Wang, L., Moriai, S., et al.: Privacy-preserving deep learning via additively homomorphic encryption. IEEE Trans. Inf. Forensics Secur. **13**(5), 1333–1345 (2017)
17. Abadi, M., et al.: Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp. 308–318 (2016)
18. Jia, J., Salem, A., Backes, M., Zhang, Y., Gong, N.Z.: MemGuard: defending against black-box membership inference attacks via adversarial examples. In: Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, pp. 259–274 (2019)
19. Zhou, Y., Ye, Q., Lv, J.C.: Communication-efficient federated learning with compensated overlap-FedAvg. IEEE Trans. Parallel Distrib. Syst. **33**, 192–205 (2021)
20. Fredrikson, M., Jha, S., Ristenpart, T.: Model inversion attacks that exploit confidence information and basic countermeasures. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, pp. 1322–1333 (2015)
21. Yang, C., Wu, Q., Li, H., Chen, Y.: Generative poisoning attack method against neural networks. arXiv preprint arXiv:1703.01340 (2017)
22. Zhang, J., Chen, J., Wu, D., Chen, B., Yu, S.: Poisoning attack in federated learning using generative adversarial nets. In: 2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE), pp. 374–380. IEEE (2019)
23. Lyu, L., Yu, H., Yang, Q.: Threats to federated learning: a survey. arXiv preprint arXiv:2003.02133 (2020)

24. Proudfoot, D.: Anthropomorphism and AI: turing's much misunderstood imitation game. Artif. Intell. **175**(5-6), 950957 (2011)
25. Zhang, J., Chen, B., Cheng, X., Binh, H.T.T., Yu, S.: PoisonGAN: generative poisoning attacks against federated learning in edge computing systems. IEEE Internet Things J. **8**(5), 3310–3322 (2020)
26. Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., Shmatikov, V.: How to backdoor federated learning. In: International Conference on Artificial Intelligence and Statistics, pp. 2938–2948. PMLR (2020)
27. Xu, G., Li, H., Liu, S., Yang, K., Lin, X.: VerifyNet: secure and verifiable federated learning. IEEE Trans. Inf. Forensics Secur. **15**, 911–926 (2019)
28. Lu, Y., Huang, X., Dai, Y., Maharjan, S., Zhang, Y.: Blockchain and federated learning for privacy-preserved data sharing in industrial IoT. IEEE Trans. Industr. Inf. **16**(6), 4177–4186 (2019)
29. Salem, A., Zhang, Y., Humbert, M., Berrang, P., Fritz, M., Backes, M.: ML-leaks: model and data independent membership inference attacks and defenses on machine learning models. arXiv preprint arXiv:1806.01246 (2018)
30. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: high confidence predictions for unrecognizable images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 427–436 (2015)
31. Deng, L.: The MNIST database of handwritten digit images for machine learning research [best of the web]. IEEE Signal Process. Mag. **29**(6), 141–142 (2012)
32. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747 (2017)