Efficient passive membership inference attack in federated learning

Oualid Zari
Inria, Univ. Côte d'Azur,
Sophia Antipolis, France

Chuan Xu
Univ. Côte d'Azur, Inria, CNRS, I3S
Sophia Antipolis, France

Giovanni Neglia
Inria, Univ. Côte d'Azur
Sophia Antipolis, France
firstname.lastname@inria.fr

October 31, 2021

Abstract

In cross-device federated learning (FL) setting, clients such as mobiles cooperate with the server to train a global machine learning model, while maintaining their data locally. However, recent work shows that client's private information can still be disclosed to an adversary who just eavesdrops the messages exchanged between the client and the server. For example, the adversary can infer whether the client owns a specific data instance, which is called a passive membership inference attack [9]. In this paper, we propose a new passive inference attack that requires much less computation power and memory than existing methods. Our empirical results show that our attack achieves a higher accuracy on CIFAR100 dataset (more than 4 percentage points) with three orders of magnitude less memory space and five orders of magnitude less calculations.

1 Introduction

In recent years, it has been demonstrated that a machine learning model is vulnerable to different attacks, e.g., membership inference attacks [10, 12], model inversion attacks [3], attribute inference attacks [5], and property inference attacks [2], which leak sensitive information present in the training dataset. The performance of these attacks depend on various factors, such as the complexity of the trained model (and then its propensity to overfit the data) [12] and the adversary's capabilities [10], including the adversary's access to auxiliary information, like a dataset statistically similar to the client's training one [11].

Federated learning (FL) [7] allows clients to participate to training without sharing their local data, but the iterative exchange of models between clients and the orchestrator can disclose additional private information. For instance, an adversary who has access to the mini-batch gradients computed on the client's local dataset may recover some data instances at the client [13, 18, 17, 4, 15], e.g., it can reconstruct up to 97.3% of a client's images to a recognizable level [15]. The adversary can also detect when new samples with a certain property (even unrelated to the learning task) are added during training to the the client's local dataset [8].[1] Finally, the adversary can exploit the FL model exchanges to perform advanced *client-level membership inference attacks* [9, 16]. As a consequence, when hospitals participate to the FL process (an increasingly popular FL use case), the adversary may infer whether a patient has visited a particular hospital.

In this paper, we a consider a *passive* attacker who does not interfere with the FL training process and only eavesdrops the exchanged messages. This attacker is also called *honest-but-curious* and *passive global*

[1] Leaking this information is dangerous, for example the adversary may infer when a person starts visiting a special type of doctor.

attacker in [9]. We show how the adversary may perform the membership inference attack with much less computational load (five orders of magnitude) and memory space (two/three orders of magnitude) than the state-of-the-art procedure proposed in [9]. Moreover, the proposed attack can be easily adapted to the case when the auxiliary dataset only contains incomplete records (e.g., labels are missing).

2 Background: Passive membership inference attack for FL

Federated learning In a cross-device federated learning setting, the clients (e.g., mobiles or IoT devices) cooperate with the server to train a global ML model $\theta \in \mathbb{R}^d$, which minimizes the following (weighted) empirical risk over all the data owned by clients:

$$\min_{\theta \in \mathbb{R}^d} \mathcal{L}(\theta) = \sum_{c \in \mathcal{C}} p_c \mathcal{L}_c(\theta) = \sum_{c \in \mathcal{C}} p_c \left(\frac{1}{|\mathcal{D}_c|} \sum_{(x,y) \in \mathcal{D}_c} \ell(\theta, x, y) \right),$$

where $\mathcal{C}$ denotes the set of clients and $\mathcal{D}_c$ the local dataset of client $c \in \mathcal{C}$ with size $|\mathcal{D}_c|$, $(x, y) \in \mathcal{D}_c$ is a sample consisting of an input object x and its associated label y, $\ell(\theta, x, y)$ measures the loss of the model on the sample and p_c is the positive weight of client c, s.t. $\sum_{c \in \mathcal{C}} p_c = 1$. To accomplish the above learning task, many distributed learning algorithms were proposed [7, 6] with FedAvg [7, Algo. 1] being the earliest and the most popular one, which we also consider in this paper. Shortly, at each communication round t, the selected client c receives the global model θ^t from the server, updates the model following some local stochastic gradient descent updates on its dataset $\mathcal{D}_c$ and sends this updated model θ_c^t back to the server, who averages all the received models.

Adversary capabilities The adversary targets a specific client c and trains an *attack model* (often a neural network) to infer whether a data point belongs to the *target dataset* $\mathcal{D}_c$ [3, 12, 9]. To this purpose, the adversary needs an auxiliary dataset $\mathcal{D}_a$. As in [9], the auxiliary dataset contains both points which do and do not belong to the target dataset (called respectively *member* and *non-member* points), i.e., $\mathcal{D}_a \cap \mathcal{D}_c \neq \emptyset$ and $\mathcal{D}_a \backslash \mathcal{D}_c \neq \emptyset$. The samples in $\mathcal{D}_a \backslash \mathcal{D}_c$ are generated from the same distribution of $\mathcal{D}_c$. In the FL setting, it is natural to assume that the adversary knows the architecture of the model under training, and this information is indeed needed by the attack proposed in [9]. Our attack could instead work under a black-box model [3], where the adversary can query the targeted models with any input and receive the corresponding output (e.g., the score vector for classification problems).

Attack strategy for classification [9] In [9], the authors consider m-ary classification problems. During training the adversary collects the models updated by c at specific time instances in the set $\mathcal{T}$. The collected models $\Theta_c = \{\theta_c^t, t \in \mathcal{T}\}$ are called the *target models*.

Since the adversary knows the architecture of target models Θ_c, for every data sample $(x, y) \in \mathcal{D}_a$, it can compute the corresponding gradients by back-propagation, i.e., $\{\partial \ell(\theta, x, y), \forall \theta \in \Theta_c\}$. Besides, it can access the loss values $\{\ell(\theta, x, y), \forall \theta \in \Theta_c\}$ and per-layer output values (including the last-layer output which is the prediction vector), i.e., $\{\theta^{[l]}(x) \in \mathbb{R}^{s(l)}, \forall l \in \{1, ..., L\}, \forall \theta \in \Theta_c\}$ where L is the number of layers in θ and $s(l)$ denotes the output size of layer l. These values together with a one hot encoding $\mathbf{e}_y \in \{0, 1\}^m$ of the label y constitute the input $\mathcal{I}(x, y)$ to a convolutional neural network ω used for membership inference prediction. The network is trained on $\{\mathcal{I}(x, y), \forall (x, y) \in \mathcal{D}_a\}$ by minimizing the mean square loss.

Note that the size of $\mathcal{I}(x, y)$ is $(d + 1 + \sum_{l=1}^{L} s(l)) \times |\mathcal{T}| + m$, which is extremely large for deep neural networks. For instance, when training the ResNet-110 in [9], even when the authors consider only the last four layers' gradients and last three layers' outputs in $\mathcal{I}(x, y)$, there are still more than 1.6 million parameters for each target model. Correspondingly, the neural network ω needs to be large as well, e.g., the original implementation requires 256 MB.[2] In addition, the adversary incurs a high computational load to compute

[2]The open source code is available at `https://github.com/SPIN-UMass/MembershipWhiteboxAttacks/blob/master/ATTACK-ALEXNET-grad_fed_local.py`.

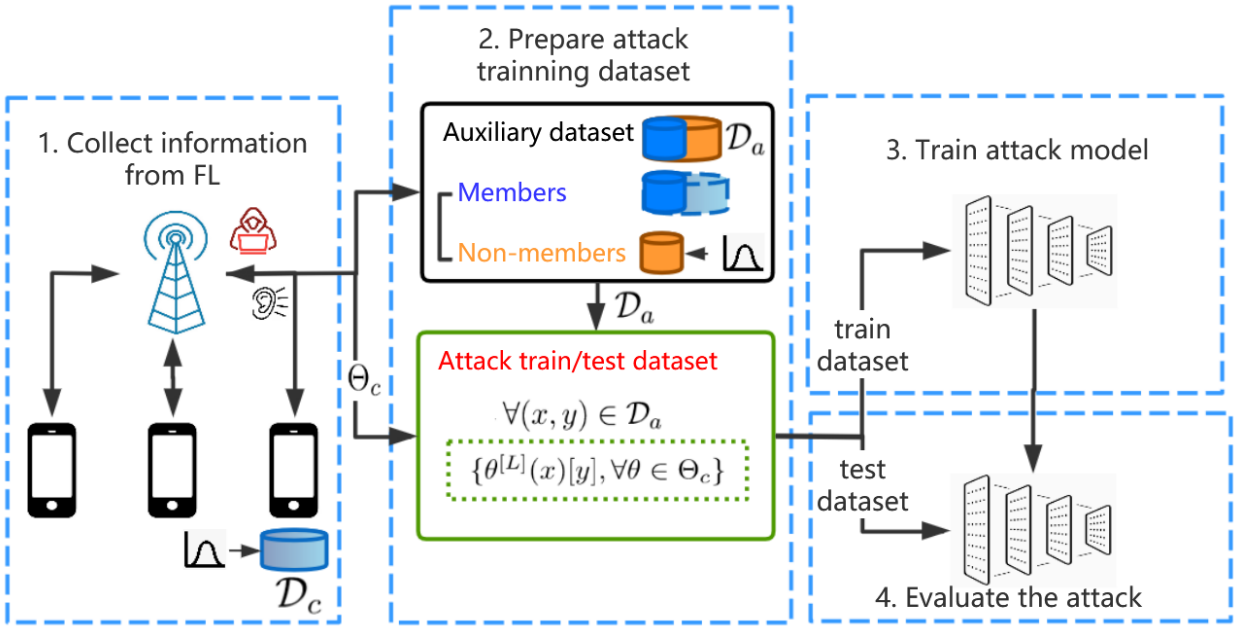


Figure 1: The procedure of our membership inference attack on federated learning. Θ_c are the set of target models, $\mathcal{D}_c$ is the target dataset, $\mathcal{D}_a$ is the auxiliary dataset and $\theta^L(x)[y]$ denotes the score of instance x for label y.

the $|\mathcal{T}| \times |\mathcal{D}_a|$ gradients in $\{\mathcal{I}(x,y), \forall(x,y) \in \mathcal{D}_a\}$, as gradients' back-propagation computation is much slower than the forward pass for deep neural networks. For instance, to achieve the attack on FL training of AlexNet on CIFAR100, the adversary computes 5×20000 gradients [9, Table XI and XII], which takes at least 1.5 hours on a NVIDIA GeForce GTX 1050 Ti. [3]

3 Efficient passive membership inference attack

In this section, we propose an efficient passive inference attack for FL which releases the adversary from the high computational burden and large memory space requirement. Our attack is depicted in Figure 1.

If the adversary attacks client c, it starts by collecting the target models $\Theta_c = \{\theta_c^t, t \in \mathcal{T}\}$ exchanged between client c and the server. Then, for every data sample $(x,y) \in \mathcal{D}_a$, it computes the score assigned by target models to the correct label y for the input x. The input of the attack model only includes $\{\theta^{[L]}(x)[y], \forall\theta \in \Theta_c\}$, where $\alpha[i]$ indicates the i^{th} element in the vector α. This choice is motivated by the empirical observation that the temporal evolutions of true label scores for members and non-members data points (i.e., those in) are easily distinguishable (see Fig. 3 in Appendix). Since target models corresponds to different time instants in the training process, the input can be a time series. We choose then as attack model a fully convolutional network, which is a suited network architecture for classifying time series [14]. The attack model's architecture is shown in Fig. 4 in Appendix and it is trained minimizing the usual cross-entropy loss.

In comparison to the state-of-the-art attack in [9], our approach requires a much smaller input of size $|\mathcal{T}|$, independently of the size of the target model. As a result, while experiments in [9] are limited to consider 5 target models, because of memory constraints, our attack has not such limit and can take into account a finer-grained temporal evolution. Also the attack model architecture in [9] considers the input as a flat

[3] We used the optimized per-sample gradient calculation package offered by Opacus [1].

Dataset	Model type	θ^T accuracy		Attack accuracy		Memory (MB)		MACs	
		Train	Test	**Ours**	Baseline	**Ours**	Baseline	**Ours**	Baseline
CIFAR100	AlexNet	99%	36%	89.5%	85.1%	1.06	1053	6.66K	1.44G
	DenseNet	100%	55%	84.2%	79.2%	1.06	1405	6.66K	1.93G
Purchase100	Fully connected	93%	82%	60.1%	72.4%	1.06	527	6.66K	0.72G

Table 1: Performance comparison of our passive membership inference attack with the baseline [9]. FL training with 4 clients spanning 300 epochs for different datasets and model architectures. Attack accuracy is averaged over all the clients.

vector, while our architecture is designed to explicitly capture the FL training dynamics, which may expose more information about data point's membership.

As most of the related work, we have assumed that samples in the auxiliary dataset $\mathcal{D}_a$ consist of input-label pairs, but labels may contain particularly sensitive information and then be better protected. Our attack can be easily adapted to deal with the case when the adversary has no access to labels. Indeed, it is sufficient to replace the attack input $\theta^{[L]}(x)[y]$ with the entropy of the score vector $\theta^{[L]}(x)$ or with its maximum value. The size of the attack model's input does not change, but one can expect the attack accuracy to decrease as less information is available to the adversary. This is confirmed by our experiments in Table 3 in the appendix.

4 Experiments

We evaluate our attacks on two datasets: one is CIFAR100 which contains 60,000 images for 100 different classes; the other one is Purchase100 which contains 197,324 shopping records for 600 products with customers clustered into 100 classes on the basis of the similarity of their purchases.

For a fair comparison with the baseline in [9], we consider the same FL scenario with 4 clients and data distributed uniformly at random among the clients. The observed epochs are $\mathcal{T} = \{100, 150, 200, 250, 300\}$ for the CIFAR100 dataset, and $\mathcal{T} = \{40, 60, 80, 90, 100\}$ for the Purchase100 dataset.[4] More details on our experimental setup are provided in the appendix.

The attack performance is evaluated in terms of accuracy of membership inference on a test auxiliary dataset. Table 1 shows that our attack requires 2 to 3 orders of magnitued less memory space (the attack model's size) and at least 5 orders of magnitude less computation (Multiply–Accumulate Operations, MACs). For CIFAR100, the accuracy of our attack is at least 5% higher than the baseline. Although our attack is less accurate on Purchase100, we can take advantage of the smaller memory footprint to increase the number of epochs considered $|\mathcal{T}|$. For example, for $|\mathcal{T}| = 30$, the accuracy of our attack increases to 62.3% with the same memory requirement (1.06 MB) and 31 KMACs, still more than 20,000 times less operations than the baseline.

Figure 2 shows the the membership inference attack accuracy decreases when labels are not available, but it is still larger than 75% when the adversary trains the model over the latest 10 epochs. The figure also illustrates client model's train and test accuracy over time and suggests that, in the setting considered in [9], the model is overfitting the dataset and the FL orchestrator may have stopped the training earlier. Intuitively, overfitting can help the adversary as the model memorizes the training samples. Table 2 confirms this intuition as the accuracy of both attacks decreases when training stops earlier. When overfitting is prevented, our attack is even better than the baseline with up to 11.3% accuracy increase (6.6 percentage points) (see Table 2).

[4]The number of communication rounds and training epochs coincide, as each client processes one local epoch at each communication round.

Observed Epochs	Attack Accuracy	
	Ours	Baseline
5, 10, 15, 20, 25	64.0%	57.4%
10, 20, 30, 40, 50	82.2%	76.5%
50, 100, 150, 200, 250	86.4%	79.5%
100, 150, 200, 250, 300	89.5%	85.1%

Table 2: Effect of the adversary's observed epochs on attack accuracy. Four clients trains AlexNet to classify CIFAR100 dataset.

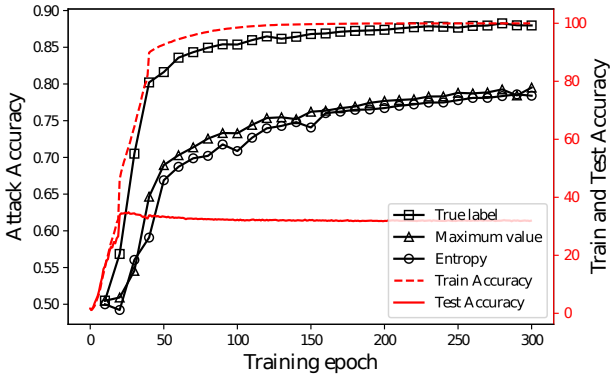


Figure 2: Attack accuracy and train/test client's model accuracy over time t. Four clients train AlexNet to classify CIFAR100 dataset. The attack model is trained with $\mathcal{T} = \{t-9, t-8, \ldots, t\}$.

References

[1] Opacus, https://github.com/pytorch/opacus.

[2] Giuseppe Ateniese, Luigi V. Mancini, Angelo Spognardi, Antonio Villani, Domenico Vitali, and Giovanni Felici. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers. *Int. J. Secur. Netw.*, 10(3):137–150, September 2015.

[3] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1322–1333, 2015.

[4] Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients–how easy is it to break privacy in federated learning? *NIPS*, 2020.

[5] Shiva Prasad Kasiviswanathan, Mark Rudelson, and Adam Smith. The power of linear reconstruction attacks. In *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*, pages 1415–1433. SIAM, 2013.

[6] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *MLSys*, 2020.

[7] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017.

[8] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 691–706. IEEE, 2019.

[9] Milad Nasr, Reza Shokri, and Amir Houmansadr. 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)*, pages 739–753, 2019.

[10] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2017.

[11] Shakila Mahjabin Tonni, Dinusha Vatsalan, Farhad Farokhi, Dali Kaafar, Zhigang Lu, and Gioacchino Tangari. Data and model dependencies of membership inference attack. *arXiv preprint arXiv:2002.06856*, 2020.

[12] Stacey Truex, Ling Liu, Mehmet Emre Gursoy, Lei Yu, and Wenqi Wei. Demystifying membership inference attacks in machine learning as a service. *IEEE Transactions on Services Computing*, 2019.

[13] Zhibo Wang, Mengkai Song, Zhifei Zhang, Yang Song, Qian Wang, and Hairong Qi. Beyond inferring class representatives: User-level privacy leakage from federated learning. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, pages 2512–2520. IEEE, 2019.

[14] Zhiguang Wang, Weizhong Yan, and T. Oates. Time series classification from scratch with deep neural networks: A strong baseline. *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 1578–1585, 2017.

[15] Hongxu Yin, Arun Mallya, Arash Vahdat, Jose M Alvarez, Jan Kautz, and Pavlo Molchanov. See through gradients: Image batch recovery via gradinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16337–16346, 2021.

[16] Jingwen Zhang, Jiale Zhang, Junjun Chen, and Shui Yu. Gan enhanced membership inference: A passive local attack in federated learning. In *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, pages 1–6, 2020.

[17] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. idlg: Improved deep leakage from gradients. *CoRR*, abs/2001.02610, 2020.

[18] Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. In *Advances in Neural Information Processing Systems*, pages 14774–14784, 2019.

Appendix

Figures for Section 3

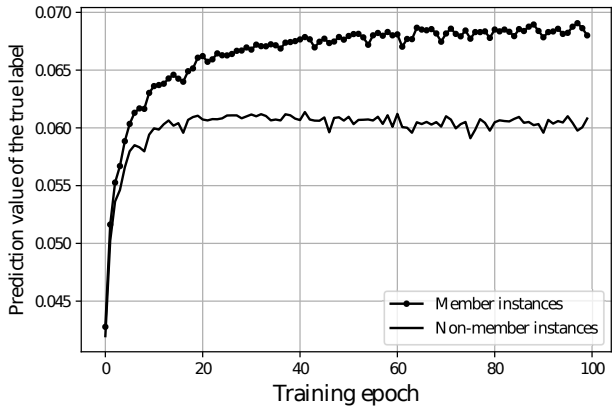


Figure 3: Average prediction value of the true label of Purchase100 dataset for members and non-members.

Table for Section 3

The knowledge of y	Attack method	Attack accuracy
✓	True label	87.9%
✗	Entropy	78.6%
✗	Maximum prediction value	77.4%

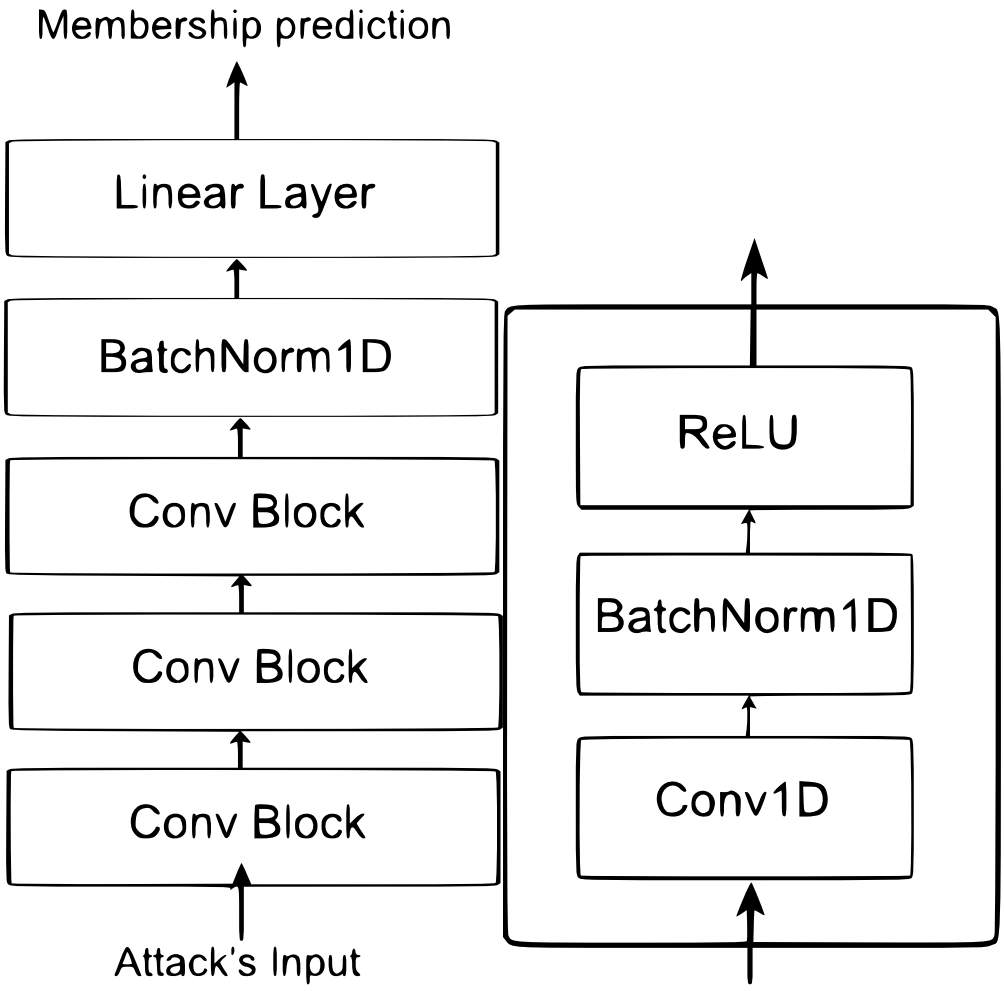
Table 3: The performance of passive membership inference attacks in FL with 4 participants after T=300 epochs for classifying CIFAR100 datset using AlexNet under different assumptions on the adversary capability to access the label y in $\mathcal{D}_a$. The observed epochs are the last 10 epochs.

Experiment details

Federated learning setup The federated learning setup is set to the same as the open-source code provided for [9]. [5] For CIFAR100 dataset, at each communication round, each client performs one local epoch update using SGD optimizer with batch size 100. The learning rate is set to 0.05 for the first 20 epochs, 0.005 for epochs from 21 to 40 and 0.0005 for the epochs from 41 to 300. For Purchase100 dataset, at each communication round, each client performs one local epoch update using Adam optimizer with batch size 100 and learning rate 0.001.

Our attack setup To train the attack model, we use the Adam optimizer with batch size 100 and learning rate of 0.001. The model is trained for 100 epochs. For both CIFAR100 and Purchase100 dataset, the model is trained on 2000 members and 2000 non-members data samples and tested on 5000 members and 5000 non-members data samples. Notice that, compared with the baseline [9], we train on less samples but test on the same numbers of data samples.

[5] `https://github.com/SPIN-UMass/MembershipWhiteboxAttacks/blob/master/ATTACK-ALEXNET-grad_fed_local.py`.



(a) model's architecture (b) Conv block

Figure 4: The architecture of the attack model consisting of three convolution blocks, one-dimensional batch normalization and a final linear layer with binary membership prediction.