Membership Inference Attack in IoT services

Bo Cheng Chu
b04902044@ntu.edu.tw
National Taiwan University

Yun Da Tsai
b04902103@ntu.edu.tw
National Taiwan University

Pei Lun Tai
b04902105@ntu.edu.tw
National Taiwan University

ABSTRACT

We quantitatively investigate how machine learning models leak information about the individual data records. We then investigate the possible threats and applications in IoT security. We conducted several experiments and compared the inference techniques on different classification models trained on different datasets and show that these models can be vulnerable to membership inference attacks. We then investigate the factors that influence this leakage and evaluate mitigation strategies. We also investigate possible defense strategy and focus on adversarial regularization which acts as a strong regularizer and significantly generalizes the model. Finally, we analysis the possible threats and scenarios of this attack on IoT security.

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

KEYWORDS

membership privacy, machine learning

ACM Reference Format:
Bo Cheng Chu, Yun Da Tsai, and Pei Lun Tai. 2018. Membership Inference Attack in IoT services. In *Proceedings of ACM Conference (Conference'17)*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Machine learning is widely used to provide popular services such as image and speech recognition and natural language translation. They are also utilized by a great deal of companies to improve marketing and advertising, to recommend products and services to users. To train more accurate model for specific users, data such as speech record, GPS location, health data, preferences are collected. These collected data may contain private information, while the model is provided to other users. Since the model takes in the training data and learns the relationship between data and corresponding labels, it is possible that we can inversely infer the training data by examining the output label. We then investigate a **shadow training** method proposed in *Membership Inference Attacks Against Machine Learning Models*[5]. It's concept is to learn the distribution of the output probability of shadow models, which mimics the behavior of the target model, and predict whether the input data is in the training set. If the membership inference attack successes, the attacker may be able to obtain sensitive data. The emergence of the Internet of Things (IoT) has generated much concern over rapidly expanding privacy risks, including those related to the inference of users' personal behavior. Some IoT devices consistently collect users' data in order to make the model more robust, and provide more accurate, personalized recommendations. However, when users consider a particular characteristic private, they may wish to prevent inference algorithms from inferring such attributes. We focus on the potential private information leak on the automatic speech recognition (ASR) system. ASR systems are widely adopted on Internet of Things (IoT) devices and provide fascinating services. Since the ASR systems are used in daily life, the privacy concerns of accessing unauthorized user's audio are of great awareness for customers.

We further investigate several defensive methods to counter the membership inference attack. Our experiments show that *adding L2 regularization* significantly reduces the membership inference accuracy, making the accuracy on CIFAR-10 drop from 0.79% to 0.50%.

2 MACHINE LEARNING BACKGROUND

Machine learning algorithms is to learn the relationship between the data and the labels and construct a model that can generalize to data records beyond the training set. Model-training algorithms aim to minimize the model's prediction error on the training dataset and thus may overfit to this dataset. This would lead to privacy leak of the data.

2.1 Machine Learning as a Service

Internet giants such as Google, Microsoft and Amazon are already offering machine learning as a service. Any customer can upload their data to the service and pay it to construct a model. The details of the service is hidden and can only be queried through API. Therefore, we define 'machine learning as a service' as a black box.

2.2 Privacy in Machine Learning

A privacy breach occurs if an adversary can use the model's output to infer the values of unintended (sensitive) attributes used as input to the model. For example, knowing that a certain patient's clinical record was used to train a model associated with a disease (e.g, to determine the appropriate medicine dosage or to discover the genetic basis of the disease) can reveal that the patient has this disease.

3 PROBLEM DEFINITION

The objective of the attacker is to construct an attack model that can recognize the classification probability differences in the target

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

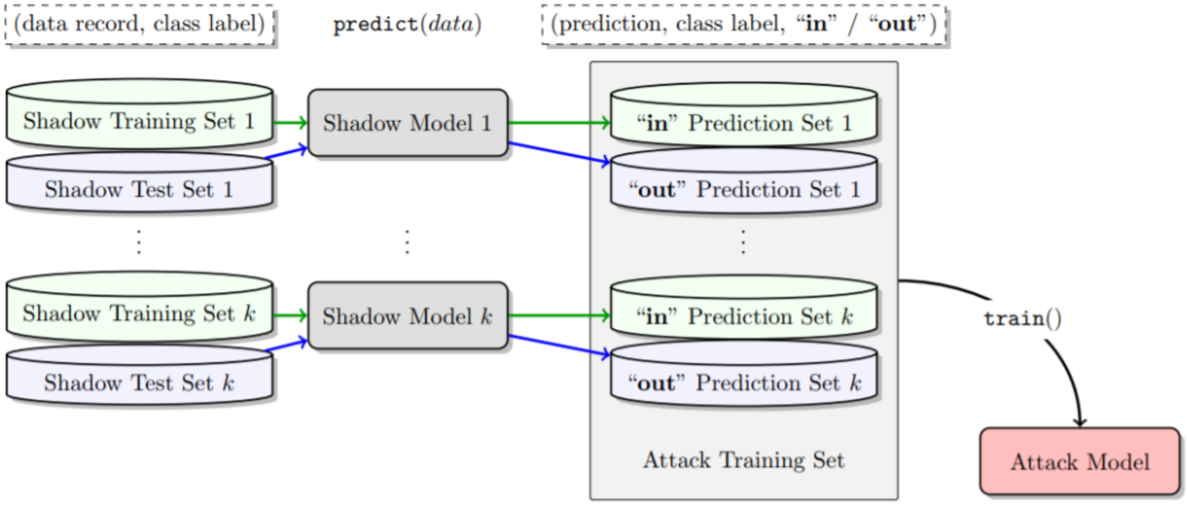


Figure 1: Shadow Training

This figure shows the training process of shadow models and attack model. Since we know the training and testing set used to train shadow models, we create a **attack training set** with X = 'the prediction output of shadow models' and Y = 'whether the data is in the training set or not'

model's behavior and use them to distinguish members from non-members of the target model's training dataset based solely on the target model's output. Privacy is breached if the attacker can successfully predict whether an input data is from the training set.

Formally, given a target model f_{target}, whose training data is D_{target}, we aim to train an attack model f_{attack} such that for an arbitrary input u, f_{attack} should determine whether $u \in D_{target}$ or not. We assume that the attacker only has black-box access to the target model. Given an input, the attacker can only obtain the final output and its probability. We also assume the attacker knows the learning algorithms used by f_{target}.

4 ATTACK

4.1 Membership Inference

The membership inference attack exploits the observation that machine learning models often behave differently on the data that they were trained on versus the data that they 'see' for the first time so that we could determine whether a given data record was part of the model's training dataset or not. The objective of the attacker is to construct an attack model that can recognize such differences in the target model's behavior and use them to distinguish members from non-members of the target model's training dataset based solely on the target model's output.

4.2 Experiment

4.2.1 setup. We conduct the experiment in the black-box scenario where the adversary can only supply inputs to the model and receive the model's output. First, we build multiple 'shadow' models intended to behave similarly to the target model. Second, to train the shadow models, we have to synthesis training data similar to the target model. Here we uses three different algorithms, Model-Based Synthesis, Statistics-Based Synthesis and Noisy Real Data.

4.2.2 dataset. We used CIFAR-10 as our benchmark dataset. CIFAR-10 is composed of 32x32 color images in 10 classes, with 6, 000 images per class. In total, there are 50, 000 training images and 10, 000 test images.

4.2.3 target model. We train a standard convolutional neural network (CNN) with two convolution and max pooling layers plus a fully connected layer of size 128 and a SoftMax layer. We use Tanh as the activation function. We set the learning rate to 0.001, the learning rate decay to 1e 07, and the maximum epochs of training to 100.

4.2.4 evaluation. We evaluate this attack by executing it on randomly reshuffled records from the target's training and test datasets. In our attack evaluation, we use sets of the same size (i.e, equal number of members and non-members) in order to maximize the uncertainty of inference, thus the baseline accuracy is 0.5.

We evaluate the attack using the standard precision and recall metrics. Precision is the fraction of the records inferred as members of the training dataset that are indeed members. Recall measures coverage of the attack, i.e., the fraction of the training records that the attacker can correctly infer as members.

4.2.5 results. We choose 2500, 5000. 10000, and 15000 for the training set size.

Table 1: Our inference attack results

Traning Set size	Precision	Recall	Accuracy
2500	0.77	1.0	0.84
5000	0.73	1.0	0.81
10000	0.73	1.0	0.81
15000	0.69	1.0	0.79

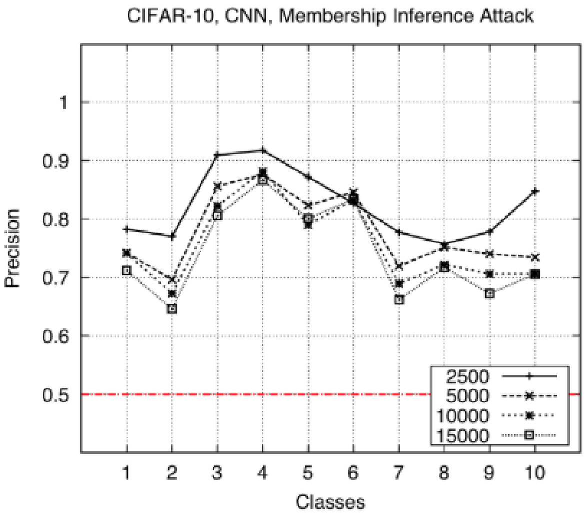


Figure 2: The inference attack result of the reference paper

4.3 Analysis

Table II shows the relationship between the accuracy of our membership inference attack and the (train-test) gap of the target models. Figure 12 also illustrates how the target models' outputs distinguish members of their training datasets from the non-members. This is the information that our attack exploits.

5 MITIGATION

There are two major groups of existing defense mechanisms. The first group includes simple mitigation techniques, such as limiting the model's predictions to top-k classes, therefore reducing the precision of predictions, or regularizing the model (e.g., using L2- norm regularizers). These techniques may impose a negligible utility loss to the model. However, they cannot guarantee any rigorous notion of privacy. The second major group of defenses use differential privacy mechanisms. These mechanisms do guarantee (membership) privacy up to their privacy parameter . However, the existing mechanisms may impose a significant classification accuracy loss for protecting large models on high dimensional data for small values of . This comes from not explicitly including utility into the design objective of the privacy mechanism.

5.1 Adversarial Regularization

Adversarial regularization is a mitigation method that achieve membership privacy with the minimum classification loss. It optimizes a composition of two conflicting objectives which is a min-max privacy game between the defense mechanism and the inference attack. This could be achieved by adversarial training similar to generative adversarial networks. To summarize, the solution will be a classification model with minimum classification loss such that the strongest inference attack against it cannot distinguish its training set members from non-members by observing the model's predictions on them.

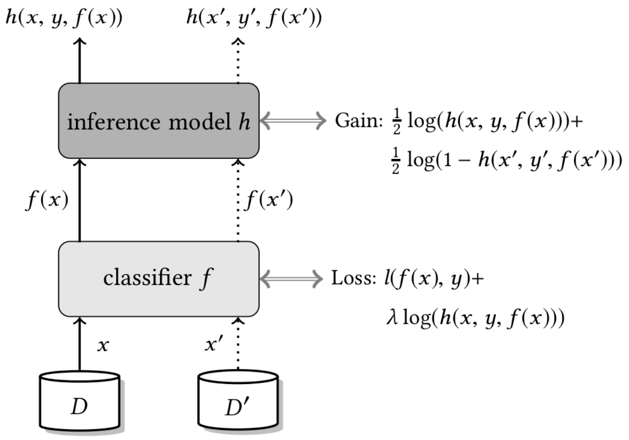


Figure 3: Adversarial Regularization framework

5.2 Experiment

5.2.1 dataset. We used CIFAR-10 as our benchmark dataset. CIFAR-10 is composed of 32x32 color images in 10 classes, with 6, 000 images per class. In total, there are 50, 000 training images and 10, 000 test images.

5.2.2 results. L2 regularization significantly reduces the inference attack accuracy while improving defence accuracy. In contrast, adversarial regularization achieve both.

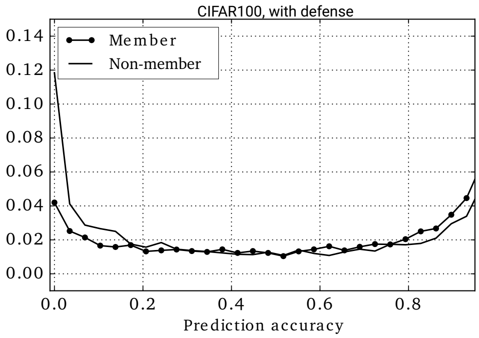
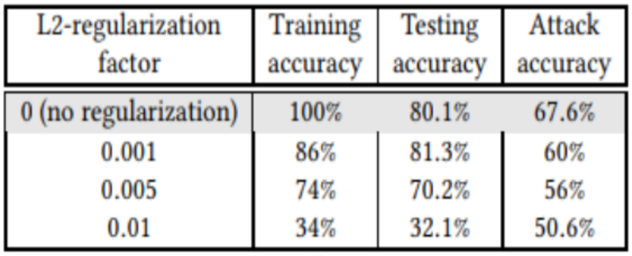


Table 2: Our results after adding L2 regularization

Traning Set size	Precision	Recall	Accuracy
2500	0.78	0.18	0.57
5000	0.61	0.06	0.52
15000	0.40	0.0008	0.50

6 IOT SECURITY

The automatic speech recognition (ASR) system is very common on IoT devices nowadays. With a audio input, a ASR system would transcribe it to written text. There are also plenty of speech assistants such as Google Assistant, Siri, Alexa, etc. Although they claim



L2-regularization factor	Training accuracy	Testing accuracy	Attack accuracy
0 (no regularization)	100%	80.1%	67.6%
0.001	86%	81.3%	60%
0.005	74%	70.2%	56%
0.01	34%	32.1%	50.6%

Figure 4: The L2 regularization result of the reference paper

that users can choose whether to allow them to collect user information such as speech data in order to improve model accuracy and enhance user experience, users can never know whether the assistants really follow the policy.

6.1 Experiment

We simulate a ASR system and try to do membership inference attack on it.

6.1.1 Target model. We use a open source state-of-the-art DNN-HMM-based ASR model *PyTorch-Kaldi*, which implements a well-known and standard algorithm for ASR systems, as the target model. Given a audio signal, the model would output a corresponding transcribed text and its probability. As the same as previous experiments, the attackers only have black-box access to the model.

6.1.2 Dataset. We use TIMIT dataset to train our ASR systems and shadow models. We split the model to three disjoint sets, and use them to train a target model and two shadow models respectively.

6.1.3 Attack. We conduct the experiment with a user-level membership attack. The attack aims to determine whether a user's data is within the target model's training dataset.

We randomly selected 25 users and 5 audios per user, and used random forest algorithm to train the attack model.

6.1.4 Results. We tried different numbers of users to train the attack model.

Table 3: Our attack results on a ASR model

Users	Accuracy	Precision	Recall
25	0.78	0.79	0.88
50	0.82	0.82	0.88

7 RELATED WORKS

7.1 Attack

Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays [4] propose the first membership inference attack on genomic data. This attack relies on the L1 distance between the allele frequencies and the victim's genomic data. *Membership Inference Attacks Against Machine Learning Models (2017 IEEE Symposium on Security and Privacy)*[5] invented a **shadow training** technique to train multiple shadow models that mimics the victim model. Then they train an attack model that takes in the output of shadow models and learns the distribution of classification probability, which would further be used to identify whether an input locates in the training set. *ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models*[1] proposes a **data transferring attack** for membership inference. Instead of training multiple shadow models, they train a single shadow model with a different dataset. This means the shadow model here is not used to mimic the target model's behavior but only to capture the membership status of data points in a machine learning training set. *The Audio Auditor: Participant-Level Membership Inference in Internet of Things Voice Services* [6] focus on a DNN-HMM-based automatic speech recognition model over the TIMIT audio data to address the membership inference attack issue on ASR systems and IoT Devices.

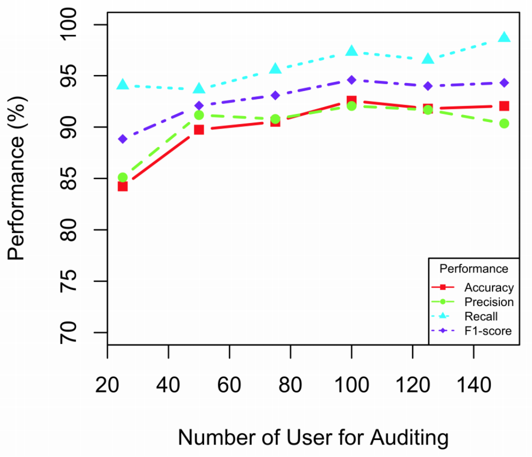
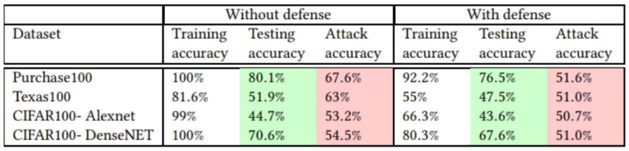


Figure 5: The inference attack result on ASR model of the reference paper

7.2 Defense

ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models[1] experiments on **Dropout** method, which prevents model from overfitting. They also propose a **Model Stacking** method, which significantly reduce inference attack accuracy while preserving model performance. *Machine Learning with Membership Privacy using Adversarial Regularization (2018 ACM SIGSAC)*[3] proposes an adversarial method that trains models that guarantees privacy while retaining predicting accuracy. They show that their training method **strongly regularizes the model**. Deep Learning with Differential Privacy (2016 ACM CCS)[2] develops new algorithmic techniques for learning and a refined analysis of privacy costs within the framework of differential privacy. They demonstrate that they can train deep neural networks with non-convex objectives, under a modest privacy budget, and at a manageable cost in software complexity, training efficiency, and model quality.



	Without defense			With defense		
Dataset	Training accuracy	Testing accuracy	Attack accuracy	Training accuracy	Testing accuracy	Attack accuracy
Purchase100	100%	80.1%	67.6%	92.2%	76.5%	51.6%
Texas100	81.6%	51.9%	63%	55%	47.5%	51.0%
CIFAR100- Alexnet	99%	44.7%	53.2%	66.3%	43.6%	50.7%
CIFAR100- DenseNET	100%	70.6%	54.5%	80.3%	67.6%	51.0%

Figure 6: The adversarial defense result of Machine Learning with Membership Privacy using Adversarial Regularization

8 CONCLUSIONS AND FUTURE WORK

Our results show that there exists potential security issues in machine learning models and IoT voice services. Overfitting on training data is the main reason that the membership inference attack can succeed. In our experiments on CIFAR-10 and TIMIT speech datasets, we can reach at least 80% accuracy if no defense techniques are applied, which means that we can almost reconstruct the whole training set.

However, several techniques can be used to successfully defend the membership inference attack. From simple methods such as *Dropout*, *L2 regularization*, to relatively complex methods that can preserve the target model's prediction accuracy such as *Model Stacking* and *Adversarial Regularization*.

We believe that as long as service providers put emphasis on these methods, membership inference attacks cannot easily succeed.

Aside from simple CIFAR-10 dataset and ASR systems, we would like to launch a deeper investigation on IoT Security such as IoT cameras. It is possible that the videos captured by surveillance cameras installed in the office or at home are being used to train better image recognition models and might impose privacy leaks.

REFERENCES

[1] Mathias Humbert Pascal Berrang Mario Fritz Michael Backes Ahmed Salem, Yang Zhang. 2019. ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models. In *NDSS 2019*. https://arxiv.org/abs/1806.01246

[2] Ian Goodfellow H. Brendan McMahan Ilya Mironov Kunal Talwar Li Zhang Martin Abadi, Andy Chu. 2016. Deep Learning with Differential Privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. https://dl.acm.org/citation.cfm?id=2978318

[3] Amir Houmansadr Milad Nasr, Reza Shokri. 2018. Machine Learning with Membership Privacy using Adversarial Regularization. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security Pages 634-646*. https://dl.acm.org/citation.cfm?id=3243855

[4] Margot Redman David Duggan Waibhav Tembe Jill Muehling John V. Pearson Dietrich A. Stephan Stanley F. Nelson David W. Craig Nils Homer, Szabolcs Szelinger. 2008. Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays. https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000167

[5] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership Inference Attacks Against Machine Learning Models. In *2017 IEEE Symposium on Security and Privacy (SP)*. https://ieeexplore.ieee.org/document/7958568

[6] Minhui Xue Chao Chen Lei Pan Jun Zhang Dali Kaafar Yang Xiang Yuantian Miao, Ben Zi Hao Zhao. 2019. The Audio Auditor: Participant-Level Membership Inference in Internet of Things Voice Services. https://arxiv.org/abs/1905.07082