Data Privacy in Machine Learning

Reza Shokri

Data Privacy and Trustworthy ML Research Lab
National University of Singapore

enze@comp.nus.edu.sg  @rzshokri
Threats to Data Privacy
Threats to Data Privacy

- Unauthorized access to data, and data breaches
- Massive data collection
Threats to Data Privacy

Direct and intentional leakage

- Unauthorized access to data, and data breaches
- Massive data collection
Threats to Data Privacy

Direct and intentional leakage

- Unauthorized access to data, and data breaches
- Massive data collection

Indirect and unintentional leakage
Threats to Data Privacy

Direct and intentional leakage

• Unauthorized access to data, and data breaches
• Massive data collection

Indirect and unintentional leakage

• Meta-data: Data about data
• Data correlated with data
• Computations on data
Privacy Risks in Machine Learning
Privacy Risks in Machine Learning

Direct Leakage

\[ X \xrightarrow{\text{training}} W \xrightarrow{\text{prediction}} f(x; W) \]
Privacy Risks in Machine Learning

Direct Leakage

Training phase

\[ f(x; W) \]

Input

Prediction
Privacy Risks in Machine Learning

Direct Leakage

training phase

inference phase

Training Set

User

\( f(x; W) \)
Privacy Risks in Machine Learning

How to prevent leakage? Secure multi-party computation, homomorphic encryption, trusted hardware, ...
Privacy Risks in Machine Learning

Indirect Leakage
Privacy Risks in Machine Learning

Indirect Leakage
Privacy Risks in Machine Learning

Indirect Leakage

predictions

parameters

Indirect Leakage

Training Set

W

X

User

f(x; W)

prediction

input
Privacy Risks in Machine Learning

What is leakage? Inferring information about members of $X$, beyond what can be learned about its underlying distribution

[Shokri, Stronati, Song, Shmatikov] Membership Inference Attacks against Machine Learning Models, SP’17
Privacy Risks in Machine Learning

What is leakage? Inferring information about members of $X$, beyond what can be learned about its underlying distribution

How to mitigate the risk?
Differential privacy

Indirect Leakage

[Shokri, Stronati, Song, Shmatikov] Membership Inference Attacks against Machine Learning Models, SP’17
How to Quantify the Leakage?

• Indistinguishability game: Can an adversary distinguish between two models that are trained on two neighboring datasets (one includes an extra data point $x$)?

• **Membership inference:** Given a model, can an adversary infer whether data point $x$ is part of its training set?

[Shokri, Stronati, Song, Shmatikov] Membership Inference Attacks against Machine Learning Models, SP’17
How to Quantify the Leakage?

- **Indistinguishability game**: Can an adversary distinguish between two models that are trained on two neighboring datasets (one includes an extra data point $x$)?

- **Membership inference**: Given a model, can an adversary infer whether data point $x$ is part of its training set?

---

[Shokri, Stronati, Song, Shmatikov] Membership Inference Attacks against Machine Learning Models, SP’17
Membership Inference Attacks against Classification Models

Machine Learning as a Service

[Shokri, Stronati, Song, Shmatikov] Membership Inference Attacks against Machine Learning Models, SP’17
Privacy Leakage due to Overfitting

Overfitted models and classes are more vulnerable

[Shokri, Stronati, Song, Shmatikov] Membership Inference Attacks against Machine Learning Models, SP’17
Disparate Privacy Vulnerability

Smaller groups are potentially more vulnerable

[Shokri, Stronati, Song, Shmatikov] Membership Inference Attacks against Machine Learning Models, SP’17
White-box Privacy Analysis

- Leakage through parameters (white-box) vs. predictions (black-box)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Architecture</th>
<th>Train Accuracy</th>
<th>Test Accuracy</th>
<th>Mem inference attack accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIFAR100</td>
<td>Alexnet</td>
<td>99%</td>
<td>44%</td>
<td></td>
</tr>
<tr>
<td>CIFAR100</td>
<td>ResNet</td>
<td>89%</td>
<td>73%</td>
<td></td>
</tr>
<tr>
<td>CIFAR100</td>
<td>DenseNet</td>
<td>100%</td>
<td>82%</td>
<td></td>
</tr>
</tbody>
</table>

White-box Privacy Analysis

- Leakage through parameters (white-box) vs. predictions (black-box)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Architecture</th>
<th>Train Accuracy</th>
<th>Test Accuracy</th>
<th>Mem inference attack accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIFAR100</td>
<td>Alexnet</td>
<td>99%</td>
<td>44%</td>
<td>Black-box</td>
</tr>
<tr>
<td>CIFAR100</td>
<td>ResNet</td>
<td>89%</td>
<td>73%</td>
<td></td>
</tr>
<tr>
<td>CIFAR100</td>
<td>DenseNet</td>
<td>100%</td>
<td>82%</td>
<td></td>
</tr>
</tbody>
</table>

High generalizability to test data

White-box Privacy Analysis

- Leakage through parameters (white-box) vs. predictions (black-box)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Architecture</th>
<th>Train Accuracy</th>
<th>Test Accuracy</th>
<th>Black-box</th>
<th>White-box (Outputs)</th>
<th>White-box (Gradients)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIFAR100</td>
<td>Alexnet</td>
<td>99%</td>
<td>44%</td>
<td>74.2%</td>
<td>74.6%</td>
<td>75.1%</td>
</tr>
<tr>
<td>CIFAR100</td>
<td>ResNet</td>
<td>89%</td>
<td>73%</td>
<td>62.2%</td>
<td>62.2%</td>
<td>64.3%</td>
</tr>
<tr>
<td>CIFAR100</td>
<td>DenseNet</td>
<td>100%</td>
<td>82%</td>
<td>67.7%</td>
<td>67.7%</td>
<td>74.3%</td>
</tr>
</tbody>
</table>

High generalizability to test data

White-box Privacy Analysis

- Leakage through parameters (white-box) vs. predictions (black-box)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Architecture</th>
<th>Train Accuracy</th>
<th>Test Accuracy</th>
<th>Mem inference attack accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIFAR100</td>
<td>Alexnet</td>
<td>99%</td>
<td>44%</td>
<td>Black-box: 74.2%</td>
</tr>
<tr>
<td>CIFAR100</td>
<td>ResNet</td>
<td>89%</td>
<td>73%</td>
<td>White-box (Outputs): 74.6%</td>
</tr>
<tr>
<td>CIFAR100</td>
<td>DenseNet</td>
<td>100%</td>
<td>82%</td>
<td>White-box (Gradients): 75.1%</td>
</tr>
</tbody>
</table>

High generalizability to test data

Low privacy (Significant leakage through parameters)

### White-box Privacy Analysis

- Leakage through parameters (white-box) vs. predictions (black-box)

#### Most accurate pre-trained models

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Architecture</th>
<th>Train Accuracy</th>
<th>Test Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIFAR100</td>
<td>Alexnet</td>
<td>99%</td>
<td>44%</td>
</tr>
<tr>
<td>CIFAR100</td>
<td>ResNet</td>
<td>89%</td>
<td>73%</td>
</tr>
<tr>
<td>CIFAR100</td>
<td>DenseNet</td>
<td>100%</td>
<td>82%</td>
</tr>
</tbody>
</table>

#### Mem inference attack accuracy

<table>
<thead>
<tr>
<th></th>
<th>Black-box</th>
<th>White-box (Outputs)</th>
<th>White-box (Gradients)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIFAR100</td>
<td>74.2%</td>
<td>74.6%</td>
<td>75.1%</td>
</tr>
<tr>
<td>CIFAR100</td>
<td>62.2%</td>
<td>62.2%</td>
<td>64.3%</td>
</tr>
<tr>
<td>CIFAR100</td>
<td>67.7%</td>
<td>67.7%</td>
<td>74.3%</td>
</tr>
</tbody>
</table>

- **Large capacity**
- **High generalizability to test data**
- **Low privacy** (Significant leakage through parameters)

White-box Privacy Analysis

- Leakage through parameters (white-box) vs. predictions (black-box)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Architecture</th>
<th>Train Accuracy</th>
<th>Test Accuracy</th>
<th>Mem inference attack accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIFAR100</td>
<td>Alexnet</td>
<td>99%</td>
<td>44%</td>
<td>Black-box: 74.2%, White-box (Outputs): 74.6%, White-box (Gradients): 75.1%</td>
</tr>
<tr>
<td>CIFAR100</td>
<td>ResNet</td>
<td>89%</td>
<td>73%</td>
<td>Black-box: 62.2%, White-box (Outputs): 62.2%, White-box (Gradients): 64.3%</td>
</tr>
<tr>
<td>CIFAR100</td>
<td>DenseNet</td>
<td>100%</td>
<td>82%</td>
<td>Black-box: 67.7%, White-box (Outputs): 87.7%, White-box (Gradients): 74.3%</td>
</tr>
</tbody>
</table>

Large capacity

High generalizability to test data

Low privacy (Significant leakage through parameters)


Decentralized (Federated) Learning

[Shokri and Shmatikov] Privacy-Preserving Deep Learning, CCS’15


[Melis, Song, De Cristofaro, Shmatikov] Exploiting Unintended Feature Leakage in Collaborative Learning, SP’19
Decentralized (Federated) Learning

Adversary can observe multiple snapshots of the model

<table>
<thead>
<tr>
<th>Observed Epochs</th>
<th>Attack Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>5, 10, 15, 20, 25</td>
<td>57.4%</td>
</tr>
<tr>
<td>10, 20, 30, 40, 50</td>
<td>76.5%</td>
</tr>
<tr>
<td>50, 100, 150, 200, 250</td>
<td>79.5%</td>
</tr>
<tr>
<td>100, 150, 200, 250, 300</td>
<td>85.1%</td>
</tr>
</tbody>
</table>

CIFAR100-Alexnet

[Shokri and Shmatikov] Privacy-Preserving Deep Learning, CCS’15
[Melis, Song, De Cristofaro, Shmatikov] Exploiting Unintended Feature Leakage in Collaborative Learning, SP’19
Decentralized (Federated) Learning

Adversary can observe multiple snapshots of the model

<table>
<thead>
<tr>
<th>Observed Epochs</th>
<th>Attack Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>5, 10, 15, 20, 25</td>
<td>57.4%</td>
</tr>
<tr>
<td>10, 20, 30, 40, 50</td>
<td>76.5%</td>
</tr>
<tr>
<td>50, 100, 150, 200, 250</td>
<td>79.5%</td>
</tr>
<tr>
<td>100, 150, 200, 250, 300</td>
<td>85.1%</td>
</tr>
</tbody>
</table>

CIFAR100-Alexnet

[Shokri and Shmatikov] Privacy-Preserving Deep Learning, CCS’15
[Melis, Song, De Cristofaro, Shmatikov] Exploiting Unintended Feature Leakage in Collaborative Learning, SP’19
Decentralized (Federated) Learning

Active Attack: Gradient Ascent

Aggregate

Decentralized (Federated) Learning

Active Attack: Gradient Ascent

Increase loss on a particular data point $x$. 

Decentralized (Federated) Learning

Active Attack: Gradient Ascent

Increase loss on a particular data point \( x \).

A participant correct it back (by running gradient descent locally) only if \( x \) is part of its training set. => membership leakage

AI Regulations - Data Protection

• “… membership inferences show that AI models can inadvertently contain personal data”

• “Attacks that reveal confidential information about the data include membership inference whereby …”

• “….. ensuring that privacy and personal data are adequately protected during the use of AI”

• “……. ensuring that AI systems are resilient to overt attacks and subtle attacks that manipulate data or algorithms....”

• “...should consider the risks to data throughout the design, development, and operation of an AI system”

On Artificial Intelligence - A European Approach to excellence and trust - Feb 2020
The White House Memo on Guidance for Regulation of Artificial Intelligence Applications - Jan 2020
Guidance on the AI auditing framework Draft guidance for consultation. Information Commissioner’s Office
Data Protection Impact Assessment

- Systematic description of data collection, storage and processing
- Assess necessity and proportionality
- Likelihood and impact of the threats on individuals
- Assess potential threats to the data
- Identify and analyze possible risk mitigation measures

https://gdpr-info.eu/art-35-gdpr/
Tool: ML Privacy Meter

ML Privacy Meter is a Python library (ml_privacy_meter) that enables quantifying the privacy risks of machine learning models. [https://github.com/privacytrustlab/ml_privacy_meter](https://github.com/privacytrustlab/ml_privacy_meter)
ML Privacy Meter
Example: NLP Models

- How much does the model leak about the sentences of a particular author/speaker? What about the membership of the author in the training set (based on known samples)?
- Which samples are leaked?
Membership Inference

SATED (Speaker Annotated TED talks) dataset

Membership Inference

SATED (Speaker Annotated TED talks) dataset

Membership Inference

SATED (Speaker Annotated TED talks) dataset

Examples of Vulnerable Training Data

But it gets worse. And this is very important, what I'm about to say is very generic. It doesn't have anything to do, in specific terms, with Stuxnet. It would work as well, for example, in a power plant or in an automobile factory. It is generic. And you don't have -- as an attacker -- you don't have to deliver this payload by a USB stick, as we saw it in the case of Stuxnet. You could also use conventional worm technology for spreading. Just spread it as

Chris Anderson: I've got a question. Ralph, it's been quite widely reported that people assume that Mossad is the main entity behind this. Is that your opinion?

Ralph Langner: Okay, you really want to hear that? Yeah. Okay. My opinion is that the Mossad is involved, but that the leading force is not Israel. So the leading force behind that is the cyber superpower. There is only one, and that's the United States -- fortunately, fortunately. Because otherwise, our problems would even be bigger.
Examples of Vulnerable Training Data

This year, Germany is celebrating the 25th anniversary of the peaceful revolution in East Germany. In 1989, the Communist regime was moved away, the Berlin Wall came down, and one year later, the German Democratic Republic, the GDR, in the East was unified with the Federal Republic of Germany in the West to found today's Germany. Among many other things, Germany inherited the archives of the East German secret police, known as the Stasi. Only two years after its dissolution, its documents were opened to the public, and historians such as me started to study these documents to learn more about how the GDR surveillance state functioned.
Privacy as a Learning Objective

\[ \Pr((x, y) \in D) = h(x, y, f(x)) \]

[Reza Shokri, 2020] Machine Learning with Membership Privacy using Adversarial Regularization, CCS ’18
Privacy as a Learning Objective

\[ \Pr((x, y) \in D) = h(x, y, f(x)) \]

maximize prediction accuracy

minimize inference accuracy

minimize \[ L_D(f) + \lambda \max_h G_{f,D,D'}(h) \]

optimal privacy-preserving classification

optimal inference

[Nasr, Shokri, Houmansadr] Machine Learning with Membership Privacy using Adversarial Regularization, CCS’18
Privacy as a Learning Objective

\[ \text{maximize} \ \text{prediction accuracy} \]

\[ \text{minimize} \ \text{inference accuracy} \]

\[ \Pr((x, y) \in D) = h(x, y, f(x)) \]

\[ \min_f \left( L_D(f) + \lambda \max_h G_{f,D,D'}(h) \right) \]

optimal privacy-preserving classification

[Nasr, Shokri, Houmansadr] Machine Learning with Membership Privacy using Adversarial Regularization, CCS'18
# Privacy and Generalization

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Without defense</th>
<th>With defense</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training accuracy</td>
<td>Testing accuracy</td>
</tr>
<tr>
<td>Purchase100</td>
<td>100%</td>
<td>80.1%</td>
</tr>
<tr>
<td>Texas100</td>
<td>81.6%</td>
<td>51.9%</td>
</tr>
<tr>
<td>CIFAR100- Alexnet</td>
<td>99%</td>
<td>44.7%</td>
</tr>
<tr>
<td>CIFAR100- DenseNET</td>
<td>100%</td>
<td>70.6%</td>
</tr>
</tbody>
</table>

- Smaller gap
- Random guess

[Nasr, Shokri, Houmansadr] Machine Learning with Membership Privacy using Adversarial Regularization, CCS’18
Bound the Worst-case Privacy Loss

- Differential Privacy: Ensure the indistinguishability between two models which are trained on two neighboring datasets.
- Randomize the training algorithm to bound the privacy loss mechanism (randomized model).

\[
\mathcal{L}(\xi) = \ln \left( \frac{\Pr[M(x) = \xi]}{\Pr[M(y) = \xi]} \right)
\]

[Dwork, McSherry, Nissim, Smith] Calibrating noise to sensitivity in private data analysis, TCC'06
DP Stochastic Gradient Descent

Randomize the gradient function

[Bassily, Smith, Thakurta] Private empirical risk minimization: Efficient algorithms and tight error bounds, FOCS’14
[Shokri and Shmatikov] Privacy-Preserving Deep Learning, CCS’15
DP Stochastic Gradient Descent

Randomize the gradient function

Loss

Lower accuracy!

[Bassily, Smith, Thakurta] Private empirical risk minimization: Efficient algorithms and tight error bounds, FOCS’14
[Shokri and Shmatikov] Privacy-Preserving Deep Learning, CCS’15
Causes of Performance Loss

- Computation of total privacy loss is not exact (i.e., the upper bound of the privacy loss (epsilon) is not tight). By overestimating the privacy loss, the added noise is larger than what is really needed to achieve the same true level of privacy.
Causes of Performance Loss

• Computation of total privacy loss is not exact (i.e., the upper bound of the privacy loss (epsilon) is not tight). By overestimating the privacy loss, the added noise is larger than what is really needed to achieve the same true level of privacy

• Gaussian mechanism is not a utility-preserving mechanism for DP SGD
Causes of Performance Loss

- Computation of total privacy loss is not exact (i.e., the upper bound of the privacy loss (epsilon) is not tight). By overestimating the privacy loss, the added noise is larger than what is really needed to achieve the same true level of privacy.

- Gaussian mechanism is not a utility-preserving mechanism for DP SGD.

- All randomized gradient vectors are treated equally (but, the signal to noise ratio is not the same across all, and their influence on the parameter vector should not be the same).
Observation

- Gradients follow a symmetric distribution, concentrated around zero

\[\text{fraction} \quad \begin{array}{c}
\text{gradient value} \\
-0.1 - 5 \cdot 10^{-2} & 0 & 5 \cdot 10^{-2} & 0.1
\end{array}\]

(a) CIFAR

\[\text{fraction} \quad \begin{array}{c}
\text{gradient value} \\
-0.1 - 5 \cdot 10^{-2} & 0 & 5 \cdot 10^{-2} & 0.1
\end{array}\]

(b) MNIST

- The DP noise would dominate the gradient values

[Nasr, Shokri, Houmansadr] Improving Deep Learning with Differential Privacy using Gradient Encoding and Denoising, 2020
Gradient Coding and De-noising

• Randomize gradients using a student-t distribution

• To compute DP parameters, encode gradient values into a finite number of samples from a Gaussian distribution
Gradient Coding and De-noising

- Randomize gradients using a student-t distribution
  - To compute DP parameters, encode gradient values into a finite number of samples from a Gaussian distribution
- Weighted update of model parameters
  - Lower the weight if noise dominates the signal
Significant Privacy Improvement for the Same Test Accuracy

\[\delta = 10^{-5}\]

Nasr, Shokri, Houmansadr] Improving Deep Learning with Differential Privacy using Gradient Encoding and Denoising, 2020
Data Privacy and Trustworthy ML Research Lab

National University of Singapore

reza@nus.edu.sg  @rzshokri

https://www.comp.nus.edu.sg/~reza/