Auditing and ML Privacy

Jun 3, 2020
Dr. Wei Wei, Prof. James Landay

CS 335: Fair, Accountable, and Transparent (FAccT) Deep Learning
Stanford University
Outline

- ML Auditing
  - Distill-and-Compare

- Privacy in ML
  - Differential Privacy with Deep Learning
  - Model Inversion Attack and Differential Privacy
  - Local Differential Privacy
  - Federated Learning
ML Auditing Using Model Distillation

Tan et al., 2018
General Additive Model

\[ g(y) = h_0 + \sum_i h_i(x_i) + \sum_{i \neq j} h_{ij}(x_i, x_j) \]

- Transformation e.g., logistic for classification
- Weights
Chicago Police “Strategic Subject”.

- A risk score for individuals being victims or offenders in a shooting incident
- 16 features
  - 8 reported being used by Chicago Police
Features Reported being Used

green - model being audited
red - mimic model
Features Reported Not Being Used

green - model being audited
red - mimic model
Auditing COMPAS

green - model being audited
red - mimic model
Auditing Lending Club

\[ g(y) = h_0 + \sum_i h_i(x_i) + \sum_{i \neq j} h_{ij}(x_i, x_j) \]
Outline

● ML Auditing
  ○ Distill-and-Compare

● Privacy in ML
  ○ Differential Privacy with Deep Learning
  ○ Model Inversion Attack and Differential Privacy
  ○ Local Differential Privacy
  ○ Federated Learning
Privacy in ML

Malicious Users

Training Data

Deep Learning Models
Inferring Sensitive Features from ML Models

Demographic Info
Medical History
Genetic Markers

Dose of Warfarin

Fredrikson et al, 2014
Inferring Training Data from Facial Recognition Models

Original Image

Inferred Image

Fredrikson et al, 2015
Centralized Setting

Compromised App

Healthy Apps

Data

Model
Distributed Setting
Outline

● **ML Auditing**
  ○ Distill-and-Compare

● **Privacy in ML**
  ○ Differential Privacy with Deep Learning
  ○ Model Inversion Attack and Differential Privacy
  ○ Local Differential Privacy
  ○ Federated Learning
Differential Privacy

M(D)

Compare

M(D')

M

Dataset D

Adjacent Inputs

Dataset D'
Differential Privacy

- A randomized mechanism satisfies \((\varepsilon, \delta)\) - differential privacy for adjacent inputs \(d\) and \(d'\) if

\[
\Pr[\mathcal{M}(d) \in S] \leq e^\varepsilon \Pr[\mathcal{M}(d') \in S] + \delta
\]

- Dataset \(d\)
- Adjacent dataset \(d'\)
- Amount of Information Leakage
- A small probability of failure

Abadi et al., 2016
Differential Privacy with Deep Learning

Differential Privacy

$$\Pr[\mathcal{M}(d) \in S] \leq e^\varepsilon \Pr[\mathcal{M}(d') \in S] + \delta$$

Solution to Differentially Private Deep Learning

$$\mathcal{M}(d) \triangleq f(d) + \mathcal{N}(0, S_f^2 \cdot \sigma^2)$$

Gradients of Deep Neural Networks

$$S_f = |f(d) - f(d')| \quad \delta \geq \frac{4}{5} \exp(-\sigma^2 / 2) \quad \varepsilon < 1$$

Abadi et al. 2016
Differentially Private SGD

Gradient Norm Bounds \( C \)

Step 1 Calculate Gradients \( \mathbf{g}_t(x_i) \leftarrow \nabla_{\theta_t} \mathcal{L}(\theta_t, x_i) \)

Step 2 Gradient Clipping \( \mathbf{\bar{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i) / \max\left(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C}\right) \)

Step 3 Adding Noise \( \mathbf{\tilde{g}}_t \leftarrow \frac{1}{L} \left( \sum_i \mathbf{\bar{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I}) \right) \)

Step 4 Parameter Updating \( \theta_{t+1} \leftarrow \theta_t - \eta_t \mathbf{\tilde{g}}_t \)

One noise added to each lot (group of data)

\[ \mathcal{M}(d) \overset{\Delta}{=} f(d) + \mathcal{N}(0, S_f^2 \cdot \sigma^2) \]

Abadi et al. 2016
Differentially Private SGD

Abadi et al, 2016
Composition Theorem

- If $f$ is $(\varepsilon_1, \delta_1)$ - DP (Differential Private) and $g$ is $(\varepsilon_2, \delta_2)$ - DP, then

  
  $f(D), g(D)$ is $(\varepsilon_1 + \varepsilon_2, \delta_1 + \delta_2)$ - DP

Abadi et al, 2016
Budget Analysis for Differentially Private SGD

- Bounds the amount of privacy leakage (budget)
- Each lot (group of data) with L samples is \((\varepsilon, \delta)\) - DP
- Using Composition theorem, our SGD is \((q \cdot \varepsilon, q \cdot \delta)\) - DP
  - \(q = L/N\) - sampling ratio per lot

Abadi et al. 2016
Moments Accountant

- Provides a tighter bounds for privacy leakage by considering the Gaussian distributed noise.
- Under Moments Accountant, there exist $c_1$ and $c_2$ such that Differentially Private SGD is

$$\left(O(q\varepsilon\sqrt{T}), \delta\right) - \text{Differentially Private}$$

$$\sigma \geq c_2 \frac{q\sqrt{T\log(1/\delta)}}{\varepsilon} \quad \varepsilon < c_1q^2T$$

- $q = L/N$ - sampling ratio per lot
- $T$ - number of time steps

Abadi et al. 2016
ε As A Function of Epoch E

- E - number of epochs
- q = 0.01
- σ = 4
- δ = 10^{-5}

Abadi et al., 2016
Performance and $(\varepsilon, \delta)$

Abadi et al. 2016
Performance and Noise Levels

(1) Large noise
\[ \sigma = 8 \]

(2) Medium noise
\[ \sigma = 4 \]

(3) Small noise
\[ \sigma = 2 \]

\[
(\tilde{O}(q\varepsilon\sqrt{T}), \delta) \quad \text{- Differentially Private}
\]

\[
\sigma \geq c_2 \frac{q\sqrt{T\log(1/\delta)}}{\varepsilon} \quad \varepsilon < c_1 q^2 T
\]

Abadi et al. 2016
Outline

- ML Auditing
  - Distill-and-Compare

- Privacy in ML
  - Differential Privacy with Deep Learning
  - Model Inversion Attack and Differential Privacy
  - Local Differential Privacy
  - Federated Learning
## Recap: Types of Adversarial Attack

<table>
<thead>
<tr>
<th>Attack Phase</th>
<th>Goal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evasion</td>
<td>Testing</td>
</tr>
<tr>
<td></td>
<td>Compromise Model Performance</td>
</tr>
<tr>
<td>Data Poisoning</td>
<td>Training</td>
</tr>
<tr>
<td></td>
<td>Compromise Model Performance</td>
</tr>
<tr>
<td>Exploratory</td>
<td>Testing</td>
</tr>
<tr>
<td></td>
<td>Explore Model Characteristics</td>
</tr>
<tr>
<td></td>
<td>Reconstruct User Data</td>
</tr>
</tbody>
</table>
Recap

- Exploratory Attack
  - Reverse engineer user data from a trained model
Model Inversion Attacks

\[ x = \arg \max_x f_y(x) \]

Fredrikson et al, 2015
Model Inversion Attack to Evaluate Differential Privacy

Park et al. 2019
Results

<table>
<thead>
<tr>
<th></th>
<th>non-private</th>
<th>( \sigma )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>( \epsilon )</td>
<td>2</td>
<td><img src="image1" alt="" /></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td><img src="image3" alt="" /></td>
</tr>
<tr>
<td></td>
<td>6</td>
<td><img src="image5" alt="" /></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td><img src="image7" alt="" /></td>
</tr>
</tbody>
</table>

Park et al. 2019
Outline

● ML Auditing
  ○ Distill-and-Compare

● Privacy in ML
  ○ Differential Privacy with Deep Learning
  ○ Model Inversion Attack and Differential Privacy
  ○ Local Differential Privacy
  ○ Federated Learning
Differential Privacy and Local Differential Privacy

\[ \Pr[\mathcal{M}(d) \in S] \leq e^\varepsilon \Pr[\mathcal{M}(d') \in S] + \delta \]

**Differential Privacy**
- d, d' are sets of data
- d and d' differ in one sample
- Centralized setting

**Local Differential Privacy**
- d and s' are single samples
- d and d' differ in one sample
- Distributed setting
Deployment of Local Differential Privacy

- RAPPOR by Google
  - Collect user data
  - Randomized Aggregatable Privacy-Preserving Ordinal Response

- Private Count Mean Sketch by Apple
  - Collect emoji usage data along with other information in iPhone
  - Learning with Privacy at Scale
Outline

- ML Auditing
  - Distill-and-Compare

- Privacy in ML
  - Differential Privacy with Deep Learning
  - Model Inversion Attack and Differential Privacy
  - Local Differential Privacy
  - Federated Learning
Distributed Optimization

Centralized Setting

Distributed Setting

Relies on distributed optimization
Federated Optimization

● Non-IID
  ○ User data is localized to their own usage
  ○ Hard to be a representative of the population

● Unbalanced Similarly
  ○ Some users will make much heavier on particular services than others

● Distributed Computing Capacity
  ○ Expect a large number of devices to be updated at the same time

● Limited communication
  ○ Mobile devices are frequently offline or on slow or expensive connections

McMahan et al, 2017
FedSGD

\[ g_k = \nabla F_k(w_t) \]

\[ w_{t+1} \leftarrow w_t - \eta \sum_{k=1}^{K} \frac{n_k}{n} g_k \]

McMahan et al, 2017
FedAvg

Gradient Model

\[ w_{t+1} \leftarrow w_t - \eta g_k \]

\[ w_{t+1} \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} w_{t+1}^k \]

McMahan et al, 2017
## Trade-offs Between Local and Global Iterations

- Number of rounds of communication necessary to achieve a test-set accuracy of 97% for the 2NN(MLP) and 99% for the CNN on MNIST

<table>
<thead>
<tr>
<th></th>
<th>2NN</th>
<th></th>
<th>CNN, $E = 5$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$C$</td>
<td>$B = \infty$</td>
<td>$B = 10$</td>
</tr>
<tr>
<td>0.0</td>
<td>1455</td>
<td>316</td>
<td>4278</td>
</tr>
<tr>
<td>0.1</td>
<td>1474 (1.0×)</td>
<td>87 (3.6×)</td>
<td>1796 (2.4×)</td>
</tr>
<tr>
<td>0.2</td>
<td>1658 (0.9×)</td>
<td>77 (4.1×)</td>
<td>1528 (2.8×)</td>
</tr>
<tr>
<td>0.5</td>
<td>—</td>
<td>(—)</td>
<td>75 (4.2×)</td>
</tr>
<tr>
<td>1.0</td>
<td>—</td>
<td>(—)</td>
<td>70 (4.5×)</td>
</tr>
</tbody>
</table>

- $C$ - ratio of clients updated to the server
- $B$ - batch size of clients
- $E$ - number of epochs client makes over its local dataset on each round

McMahan et al, 2017
Comparisons Between FedSGD and FedAvg

<table>
<thead>
<tr>
<th></th>
<th>CNN</th>
<th>MNIST CNN, 99% ACCURACY</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Non-IID</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>E</td>
<td>B</td>
<td>u</td>
<td>IID</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FedSGD</td>
<td>1</td>
<td>∞</td>
<td>1</td>
<td>626</td>
<td></td>
<td></td>
<td>483</td>
</tr>
<tr>
<td>FedAvg</td>
<td>5</td>
<td>∞</td>
<td>5</td>
<td>179 (3.5×)</td>
<td></td>
<td></td>
<td>1000 (0.5×)</td>
</tr>
<tr>
<td>FedAvg</td>
<td>1</td>
<td>50</td>
<td>12</td>
<td>65 (9.6×)</td>
<td></td>
<td></td>
<td>600 (0.8×)</td>
</tr>
<tr>
<td>FedAvg</td>
<td>20</td>
<td>∞</td>
<td>20</td>
<td>234 (2.7×)</td>
<td></td>
<td></td>
<td>672 (0.7×)</td>
</tr>
<tr>
<td>FedAvg</td>
<td>1</td>
<td>10</td>
<td>60</td>
<td>34 (18.4×)</td>
<td></td>
<td></td>
<td>350 (1.4×)</td>
</tr>
<tr>
<td>FedAvg</td>
<td>5</td>
<td>50</td>
<td>60</td>
<td>29 (21.6×)</td>
<td></td>
<td></td>
<td>334 (1.4×)</td>
</tr>
<tr>
<td>FedAvg</td>
<td>20</td>
<td>50</td>
<td>240</td>
<td>32 (19.6×)</td>
<td></td>
<td></td>
<td>426 (1.1×)</td>
</tr>
<tr>
<td>FedAvg</td>
<td>5</td>
<td>10</td>
<td>300</td>
<td>20 (31.3×)</td>
<td></td>
<td></td>
<td>229 (2.1×)</td>
</tr>
<tr>
<td>FedAvg</td>
<td>20</td>
<td>10</td>
<td>1200</td>
<td>18 (34.8×)</td>
<td></td>
<td></td>
<td>173 (2.8×)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>LSTM</th>
<th>SHAKESPEARE LSTM, 54% ACCURACY</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Non-IID</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>E</td>
<td>B</td>
<td>u</td>
<td>IID</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FedSGD</td>
<td>1</td>
<td>∞</td>
<td>1.0</td>
<td>2488</td>
<td></td>
<td></td>
<td>3906</td>
</tr>
<tr>
<td>FedAvg</td>
<td>1</td>
<td>50</td>
<td>1.5</td>
<td>1635 (1.5×)</td>
<td></td>
<td></td>
<td>549 (7.1×)</td>
</tr>
<tr>
<td>FedAvg</td>
<td>5</td>
<td>∞</td>
<td>5.0</td>
<td>613 (4.1×)</td>
<td></td>
<td></td>
<td>597 (6.5×)</td>
</tr>
<tr>
<td>FedAvg</td>
<td>1</td>
<td>10</td>
<td>7.4</td>
<td>460 (5.4×)</td>
<td></td>
<td></td>
<td>164 (23.8×)</td>
</tr>
<tr>
<td>FedAvg</td>
<td>5</td>
<td>50</td>
<td>7.4</td>
<td>401 (6.2×)</td>
<td></td>
<td></td>
<td>152 (25.7×)</td>
</tr>
<tr>
<td>FedAvg</td>
<td>5</td>
<td>10</td>
<td>37.1</td>
<td>192 (13.0×)</td>
<td></td>
<td></td>
<td>41 (95.3×)</td>
</tr>
</tbody>
</table>

K - number of clients
B - batch size
E - number of epochs
u - $(\mathbb{E}[n_k]/B)E$

McMahan et al, 2017
Effects of Number of Local Epoches

McMahan et al, 2017
Effects on $\eta$

McManan et al, 2017
Reading Assignments (ML Auditing)

- Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety, arXiv 2016
- Malgieri, Gianclaudio. The concept of fairness in the GDPR: a linguistic and contextual interpretation, FAccT 2020
- Goodman, Bryce, and Seth Flaxman. European Union regulations on algorithmic decision-making and a “right to explanation”, AI magazine 2017
- Bellamy, Rachel KE, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia et al. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, arXiv 2018
Reading Assignments (Privacy)

- Bonawitz, Keith, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning, ACM SIGSAC Conference on Computer and Communications Security, 2017
- Bonawitz, Keith, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon et al. Towards federated learning at scale: System design, SysML 2019
- Smith, Virginia, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S. Talwalkar. Federated multi-task learning, NeurIPS 2017
- Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith. "Calibrating noise to sensitivity in private data analysis, Theory of cryptography conference 2006