Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data

Nicolas Papernot

joint work with
Martín Abadi, Úlfar Erlingsson, Ian Goodfellow, Kunal Talwar

Google Brain

Nicolas is at Penn State, was an intern in Brain; Ian did part of the work at OpenAI.
Some challenges of learning from private data

Training-data extraction attacks
Fredrikson et al. (2015) *Model Inversion Attacks*

Membership attacks
Shokri et al. (2016) *Membership Inference Attacks against ML Models*
Types of adversaries and our threat model

Model querying (black-box adversary)
- Shokri et al. (2016) Membership Inference Attacks against ML Models
- Fredrikson et al. (2015) Model Inversion Attacks

Model inspection (white-box adversary)
- Zhang et al. (2017) Understanding DL requires rethinking generalization

In our work, the threat model assumes:
- Adversary can make a potentially unbounded number of queries
- Adversary has access to model internals
Quantifying privacy

Randomized Algorithm

Answer 1
Answer 2
...  
Answer n

Randomized Algorithm

Answer 1
Answer 2
...  
Answer n

???
Our design goals

Problem
Preserve **privacy of training data** when learning **classifiers**

Goals
**Differential privacy** protection guarantees
**Intuitive privacy** protection guarantees
**Generic*** (independent of learning algorithm)

*This is a key distinction from previous work, such as
Pathak et al. (2011) *Privacy preserving probabilistic inference with hidden markov models*
Jagannathan et al. (2013) *A semi-supervised learning approach to differential privacy*
Shokri et al. (2015) *Privacy-preserving Deep Learning*
Abadi et al. (2016) *Deep Learning with Differential Privacy*
Hamm et al. (2016) *Learning privately from multiparty data*
The PATE approach:

Private Aggregation of Teacher Ensembles
Teacher ensemble

Sensitive Data

Partition 1 → Teacher 1
Partition 2 → Teacher 2
Partition 3 → Teacher 3
... → ...
Partition n → Teacher n

Aggregated Teacher

Training → Data flow
Aggregation

Count votes

\[ n_j(x) = |\{i : i \in 1..n, f_i(x) = j\}| \]

Take maximum

\[ f(x) = \arg \max_j \left\{ n_j(x) \right\} \]
Intuitive privacy analysis

If most teachers agree on the label, it does not depend on specific partitions, so the privacy cost is small.

If two classes have close vote counts, the disagreement may reveal private information.
Noisy aggregation

\[ n_j(x) = |\{i : i \in 1..n, f_i(x) = j\}| \]

Count votes

\[ \text{Lap} \left( \frac{1}{\epsilon} \right) \]

Add Laplacian noise

\[ f(x) = \arg \max_j \left\{ n_j(x) + \text{Lap} \left( \frac{1}{\epsilon} \right) \right\} \]

Take maximum
Teacher ensemble

Sensitive Data

Partition 1 → Teacher 1
Partition 2 → Teacher 2
Partition 3 → Teacher 3
... →...
Partition n → Teacher n

Aggregated Teacher

Training → Data flow
Student training

Sensitive Data

Partition 1 → Teacher 1
Partition 2 → Teacher 2
Partition 3 → Teacher 3
... → ...
Partition n → Teacher n

Aggregated Teacher → Student

Not available to the adversary
Available to the adversary
Queries

Sensitive Data → Public Data

Training → Inference → Data flow
Why train an additional “student” model?

The aggregated teacher violates our threat model:

1. Each prediction increases total privacy loss.
   Privacy budgets create a tension between the accuracy and number of predictions.

2. Inspection of internals may reveal private data.
   Privacy guarantees should hold in the face of white-box adversaries.
Student training

- Sensitive Data
- Partition 1 → Teacher 1
- Partition 2 → Teacher 2
- Partition 3 → Teacher 3
- Partition n → Teacher n
- Aggregated Teacher
- Student
- Queries

Data flow:
- Training
- Inference
- Data flow

- Not available to the adversary
- Available to the adversary

Sensitive Data
Public Data
Differential privacy analysis

**Differential privacy:**
A randomized algorithm $M$ satisfies $(\varepsilon,\delta)$ differential privacy if for all pairs of neighbouring datasets $(d,d')$, for all subsets $S$ of outputs:

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

Application of the **Moments Accountant** technique (Abadi et al, 2016)

Strong **quorum** $\Rightarrow$ Small privacy cost

Bound is **data-dependent**: computed using the empirical quorum
PATE-G: the generative variant of PATE
Generative Adversarial Networks (GANs)

2 competing models trying to game each other:

**Generator:**
- Input: noise sampled from random distribution
- Output: synthetic input close to the expected training distribution

**Discriminator**
- Input: output from generator OR example from real training distribution
- Output: in distribution OR fake

Goodfellow et al. (2014) Generative Adversarial Networks
GANs for semi-supervised learning

2 competing models trying to game each other:

**Generator:**
- Input: noise sampled from random distribution
- Output: synthetic input close to the expected training distribution

**Discriminator**
- Input: output from generator OR example from real training distribution
- Output: in distribution *(which class)* OR fake

Gaussian sample → Fake sample

Sample → P(real 0)=... P(real 1)=... ...
             P(real n)=... P(fake)=...

Salimans et al. (2016) *Improved techniques for training GANs*
Student training in PATE-G

Sensitive Data

Partition 1
Teacher 1

Partition 2
Teacher 2

Partition 3
Teacher 3

Partition n
Teacher n

Aggregated Teacher

Not available to the adversary
Available to the adversary

Generator

Discriminator

Public Data

Queries

Sensitive Data

Training

Inference

Data flow
Deployment of PATE-G
Experimental results
# Experimental setup

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Teacher Model</th>
<th>Student Model</th>
<th>Student Public Data</th>
<th>Testing Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST</td>
<td>2 conv + 1 relu</td>
<td>GANs (6 fc layers)</td>
<td>test[:1000]</td>
<td>test[1000:]</td>
</tr>
<tr>
<td>SVHN</td>
<td>2 conv + 2 relu</td>
<td>GANs (7 conv + 2 NIN)</td>
<td>test[:1000]</td>
<td>test[1000:]</td>
</tr>
<tr>
<td>UCI Adult</td>
<td>RF (100 trees)</td>
<td>RF (100 trees)</td>
<td>test[:500]</td>
<td>test[500:]</td>
</tr>
<tr>
<td>UCI Diabetes</td>
<td>RF (100 trees)</td>
<td>RF (100 trees)</td>
<td>test[:500]</td>
<td>test[500:]</td>
</tr>
</tbody>
</table>

/TensorFlow/models/tree/master/differential_privacy/multiple_teachers
Aggregated teacher accuracy
Trade-off between student accuracy and privacy

- Pathak et al., 2011
- Papernot et al., 2017
- Shokri et al., 2015
- Abadi et al., 2016
Trade-off between student accuracy and privacy

UCI Diabetes

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$\epsilon$</td>
<td>1.44</td>
</tr>
<tr>
<td>$\delta$</td>
<td>$10^{-5}$</td>
</tr>
<tr>
<td>Non-private baseline</td>
<td>93.81%</td>
</tr>
<tr>
<td>Student accuracy</td>
<td>93.94%</td>
</tr>
</tbody>
</table>
Come check out our poster C13 (PATE-G swag will be distributed while supplies last)

nicolas@papernot.fr
www.cleverhans.io
@NicolasPapernot