Towards End-to-End Learning with Differential Privacy (and Fairness)

Yu-Xiang Wang

Computer Science Department
UC Santa Barbara
Challenges in ML for All-Inclusive Finance

• Privacy Challenge
  • Sensitive financial data coming from users directly!
  • How to share data across different parties?
  • How to learn from the data while ensuring the privacy of all users?

• Fairness Challenge
  • Who to lend to? The ML model should not discriminate.
  • Who to decline insurance?
  • Are we making the rich richer?
Outline

1. Differential Privacy and Private Learning

2. Learning Fair Representations (a preliminary result)
Lessons from privacy breaches

“Robust De-anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset)”

• Anonymization doesn’t work!
• Need robust / provable approaches.

“Only You, Your Doctor, and Many Others May Know”

Vijay Pandurangan. tech.vijayp.ca, 2014
Recent/upcoming legislations on privacy forces companies to revise their data practice

- I can’t keep personal data for more than three weeks?
- I will have to delete all traces of a user upon request?

How about my machine learning models trained on user data?
ML models memorize training datasets, even though they are generalizing well!

Membership Inference Attacks Against Machine Learning Models

Reza Shokri
Cornell Tech

Marco Stronati
INRIA

Congzheng Song
Cornell

Vitaly Shmatikov
Cornell Tech

Abstract—We quantitatively investigate how machine learning models leak information about the individual data records on which they were trained. We focus on the basic membership inference attack: given a data record and black-box access to a model, determine if the record was in the model’s training dataset. To perform membership inference against a target model, we make adversarial use of machine learning and train our own inference model to recognize differences in the target model’s predictions on the inputs that it trained on versus the inputs that it did not train on.

We empirically evaluate our inference techniques on classification models trained by commercial “machine learning as a service” providers such as Google and Amazon. Using realistic datasets and classification tasks, including a hospital discharge dataset whose membership is sensitive from the privacy perspective, we show that these models can be vulnerable to membership inference attacks. We then investigate the factors that influence this leakage and evaluate mitigation strategies.

The Secret Sharer: Measuring Unintended Neural Network Memorization & Extracting Secrets

Nicholas Carlini
University of California, Berkeley

Chang Liu
University of California, Berkeley

Jernej Kos
National University of Singapore

Ulfar Erlingsson
Google Brain

Dawn Song
University of California, Berkeley

This paper presents exposure, a simple-to-compute metric that can be applied to any deep learning model for measuring the memorization of secrets. Using this metric, we show how to extract those secrets efficiently using black-box API access. Further, we show that unintended memorization occurs early, is not due to over-fitting, and is a persistent issue across different types of models, hyperparameters, and training strategies. We experiment with both real-world models (e.g., a state-of-the-art translation model) and datasets (e.g., the Enron email dataset, which contains users’ credit card numbers) to demonstrate both the utility of measuring exposure and the ability to extract secrets.

Finally, we consider many defenses, finding some ineffective (like regularization), and others to lack guarantees. However, by instantiating our own differentially-private recurrent model, we validate that by appropriately investing in the use of state-of-the-art techniques, the problem can be resolved, with high utility.

Security and Privacy, 2017

ArXiv, 2018
History of privacy research

• Statistical disclosure control
  • [Duncan et. al., Hundepool et. al., since 1970s ]

• k-anonymity, l-divergence, t-closeness
  • [Sweeny, Machanavajjhala et. al., Li et. al., 2002-2007]

• Differential privacy
  • [Dwork, McSherry, Nissim, Smith, 2006++]
Differential Privacy (Dwork et. al., 2006; Gödel Prize 2017) makes no assumption on adversaries

- Almighty Adversary
  - Arbitrary side info, arbitrary computational power.
- Interpretable, quantifiable, composable.
Deployments of Differential Privacy

Aggregate via Differential Privacy

Learn from crowd while protecting individual privacy
Strong mathematical guarantees
iOS and macOS

Chrome Settings

- History: Automatically send usage statistics and crash reports to Google
- Extensions: Send RAPPOB statistics to Google
- Settings: Send a “Do Not Track” request with your browsing traffic

Uber Releases Open Source Project for Differential Privacy
Katie Tezapsidis, Software Engineer, Privacy Engineering

The U.S. Census Bureau Adopts Differential Privacy
John M. Abowd
Chief Scientist and Associate Director for Research and Methodology
U.S. Census Bureau
2018 International Methodology Symposium
Ottawa, Ontario, Canada
November 9, 2018
Formal definition of DP

• Let $Z, Z'$ be any two datasets that differ only by one user, and A is a randomized algorithm. We say A is $\varepsilon$-DP if for all output $h$

$$\sup_{Z, Z' : d(Z, Z') \leq 1} \sup_{h \in \mathcal{H}} \log \frac{p_{h \sim A(Z)}(h)}{p_{h \sim A(Z')}(h)} \leq \varepsilon$$

small $\delta$ $(\varepsilon, \delta)$ – DP
Example: Who voted for Trump?

• How many people in this room voted for Trump?

• Let’s say the answer is 15.
  • If I know the political view of everybody else except Le Song.
  • Then I can easily infer his choice.

• DP releases: 15 + noise.
Differentially Private Machine Learning

Data
- Feature-label pairs
- Unlabeled features
- Feature points

Learning algorithm
- Supervised Learning
- Clustering
- Kernel Density Estimation

Classifier
- Cluster centers
- Estimated density function

Data
- Feature-label pairs
- Unlabeled features
- Feature points

Differentially private learning algorithm
- Supervised Learning
- Clustering
- Kernel Density Estimation

Classifier
- Cluster centers
- Estimated density function
Example: Recommender System

Deep Neural Network based Recommendation Engine
“If your recommendation engine is private, then an adversary can’t infer whether a particular user was present”
A closely related setting: Federated Learning

Additional considerations
- Communication cost
- Size of the model
- Rounds of adaptivity

Illustration extracted from McMahan and Ramage (2017)

Really a very scalable and practical setting!
Two generic algorithms for DP Learning

• **Posterior Sampling:** (Dimitrakakis et al. 14, W., Fienberg, Smola, 15)
  • Sample from a (possibly modified) posterior distribution
  • Essentially exponential mechanism (McSherry & Talwar 07)
  • Privacy for free: Rate-optimal, reuse existing implementation

• **Noisy SGD:** (Bassily et al., 14; Abadi et al. 16)
  • Randomly select a minibatch
  • Add noise to the (clipped) stochastic gradients.
In practice (Liu, W., Smola, RecSys 2015)

• Worked well on Netflix/Yahoo data sets.
  
  • Netflix: $\epsilon = 0.25$, RMSE = 1.1 (non-private: 0.9)
  • Yahoo: $\epsilon = 0.25$, RMSE = 1.5 (non-private 1.3)

• About 15% loss in performance

Privacy-Utity Tradeoff

Larger $\epsilon$ >>> Stronger privacy >>> Smaller $\epsilon$
Privacy can be **for free**!

• In the Trump example, we are interested to see the mean of the population, but all we get is $n$ iid samples.

\[
\text{Statistical error: } O_P \left( \frac{1}{\sqrt{n}} \right)
\]

\[
\text{Error due to DP: } O_P \left( \frac{1}{n\epsilon} \right)
\]
The list of problems where DP is almost free

• Linear regression (W., 2018)
  • Statistical Rate: \( n^{-1} \) error from DP \((n\epsilon)^{-2}\)

• Nonparametric regression: (Wang, Baby, W., 2019)
  • Statistical Rate: \( n^{-\frac{2s}{2s+1}} \) error from DP \((n\epsilon)^{-\frac{4s}{2s+1}}\)

• Many more in the fast-growing literature:
  • Strongly convex ERM, hypothesis testing... and so on.
Practical challenges in Scaling-up DP computation!

- Need to add too much noise.
  - Ruin inferences. e.g., Contingency Table (Fienberg et. al. 2010),
  - GWAS data (Yu et. al., PSD’14), etc.

- Need a lot of tricks/hacks to work
  - E.g., “clipping” “rescaling” as in the Netflix data.

- Require advanced mathematical calculations
  - The SOTA algorithmic tools in DP (e.g., Moments Accountant (Abadi et al., 16),
    CDP (Bun et al., 17) and RDP (Mironov, 18)) are not easy to use / extend.
Our Solution: Analytical moments accountant

(W., Balle, Kasiviswanathan, 2019), (Zhu and W., 2019)

- Tracking RDP for all order as a symbolic functions.
- Numerical calculations for $(\varepsilon, \delta)$-DP guarantees.
- Automatically DP calculations for complex algorithms.
- Enable state-of-the-art DP for non-experts.

ε = ?, δ = 1e-8

Open source project: https://github.com/yuxiangw/autodp

pip install autodp
AutoDP often leads to even stronger composition than the optimal composition!

More details in Yuqing Zhu’s talk / poster on Tuesday.
Outline

1. Differential Privacy and Private Learning

2. Learning Fair Representations
Fairness challenges

Google’s image recognition system
Illustration borrowed from Ziyuan Zhong’s Blogpost.
Fairness definitions are still being debated and it will take some time to settle.

• **Individual fairness:** *(Dwork et al. 2012, Fairness Through Awareness)*
  - Similar individuals should be treated similarly
  - Can be achieved by Exponential Mechanism!

• **Statistical fairness:***
  - Unawareness
  - Demographic parity
  - Equalized odds/Equality of Opportunity
  - And more…
Rawls’s Veil of Ignorance

(Curtesy Subhash Suri, figure from Wikipedia)

• How would you develop of social contract fairly?
• How to minimize personal biases and prejudices?
• Rawls: Imagine yourself behind a veil of ignorance...

• Technical formalism:
  • My decision is based on a feature embedding statistically independent to Race/Gender/Age and so on.
Learning a fair representation with HSIC
(Wang, W. and Wang, 2019)

• Let $s$ be protected attributes and $z$ be the learned embedding.

• We minimize

$$\mathcal{L}(x, y) = -\mathbb{E}_{z \sim \text{enc}(x)}[\log f(y|z)] + HSIC_{z \sim \text{enc}(x)}(z, s)$$

• Penalize the Hilbert-Schmidt Independence Criterion (Gretton et al, 2015)
  • Induce statistical independence between $z$ and $s$. 

Figure 1: Network Architecture of our model for learning informative representation $z$ with respect to $y$ which factors out protected factors $s$. This theorem makes HSIC an ideal tool in our problem of learning invariant representations, as we can include it as a constraint or penalty term during training to restrain $z$ from correlating with the protected factor $s$. To see the easiness of using HSIC, we now review an efficient way to calculate it.

Lemma 1. Empirical HSIC

Given $m$ sample points, we can compute HSIC empirically in $O(m^2)$ with the following formula

$$HSIC(F, G) = \frac{1}{m(m-1)} \text{Tr}(KHLH)$$

where $K = T^T$, $L = L^T$, and $H = I_{11}^T/m$.
Preliminary Empirical Results: Predicting “Income” when “gender” is protected.

Table 1: Fair classification on Adult and German.

<table>
<thead>
<tr>
<th></th>
<th>Adult</th>
<th>German</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Accuracy on $s$</td>
<td>Accuracy on $y$</td>
<td>Accuracy on $s$</td>
<td>Accuracy on $y$</td>
</tr>
<tr>
<td>Major line</td>
<td>0.67</td>
<td>0.75</td>
<td>0.78</td>
<td>0.70</td>
</tr>
<tr>
<td>2-layer NN (baseline)</td>
<td>0.78</td>
<td>0.88</td>
<td>0.78</td>
<td>0.91</td>
</tr>
<tr>
<td>VFAE [21]</td>
<td>0.67</td>
<td>0.84</td>
<td>0.78</td>
<td>0.72</td>
</tr>
<tr>
<td>Xie et al. [29]</td>
<td>0.67</td>
<td>0.83</td>
<td>0.78</td>
<td>0.70</td>
</tr>
<tr>
<td>Moyer et al. [23]</td>
<td>0.67</td>
<td>0.84</td>
<td>0.78</td>
<td>0.72</td>
</tr>
<tr>
<td>min-max COCO [9]</td>
<td>0.71</td>
<td>0.88</td>
<td>0.79</td>
<td>0.88</td>
</tr>
<tr>
<td>min-max HSIC (Ours)</td>
<td><strong>0.67</strong></td>
<td><strong>0.89</strong></td>
<td><strong>0.78</strong></td>
<td><strong>0.92</strong></td>
</tr>
</tbody>
</table>
Take home messages

• There is a rich and growing literature on addressing the problem of “privacy” and “Fairness” in machine learning.

• They may not necessarily lead to (significant) accuracy loss in the resulting ML models.

• Crucial challenges to overcome in realizing the ideal of being All-Inclusive in financial services.
Thank you for your attention!

Acknowledgement: