EACL 2023 Tutorial on Privacy-Preserving NLP
Block 2b: federated learning and other privacy enhancing methods

Fatemehsadat Mireshghallah
May 2023
So far …

- We heard about the **attack** landscape and **leakage** in language models
Large Models are Leaky

LONG LIVE THE REVOLUTION.
OUR NEXT MEETING WILL BE
AT THE DOCKS AT MIDNIGHT
ON JUNE 28

AHA, FOUND THEM!

WHEN YOU TRAIN PREDICTIVE MODELS
ON INPUT FROM YOUR USERS, IT CAN
LEAK INFORMATION IN UNEXPECTED WAYS.

xkcd.com/2169/
Large Models are Leaky: Data Extraction

Prompt
East Stroudsburg
Stroudsburg...

Large Language Model
(GPT-2)

Memorized Text
Corp. Name: **** Corp. Seabank Centre
Person’s Name: Peter W****
Email:****@****. com
Phone Number: +****7 5****
Large Models are Leaky: Data Extraction

- Github CoPilot

Title:
Hi everyone, my name is Anish Athalye and I'm a PhD student at Stanford University.
Large Models are Leaky: Data Extraction

- Github CoPilot

Title:

Hi everyone, my name is Anish Athalye and I'm a PhD student at Stanford University.

https://www.anish.io

Anish Athalye

I am a PhD student at MIT in the PDOS group. I'm interested in formal verification, systems, security, and machine learning.

GitHub: @anishathalye   Blog: anishathalye.com
So far …

- We heard about the attack landscape and leakage in language models
- We discussed privacy protection methods with **formal guarantees**:  
  - Differential Privacy
Differential Privacy

A randomized algorithm $A$ satisfies $\epsilon$-DP, if for all databases $D$ and $D'$ that differ in data pertaining to one user, and for every possible output value $Y$: 

$$\frac{\Pr[A(D) = Y]}{\Pr[A(D') = Y]} \leq e^\epsilon.$$
So far …

- We heard about the attack landscape and leakage in language models.
- We discussed privacy protection methods with **formal guarantees**:
  - Differential Privacy
  - Encryption-based methods
In this talk …

- There are also **privacy-enhancing paradigms and execution modes** that do not necessarily have formal guarantees.
- These methods are designed to **limit access to raw data**, but provide **no worst-case guarantees**:
  - Federated Learning
  - Split Learning
  - Privacy Regularizers
In this talk …

- There are also **privacy-enhancing paradigms and execution modes** that do not necessarily have formal guarantees.
- These methods are designed to **limit access to raw data**, but provide **no worst-case guarantees**:
  - Federated Learning
  - Split Learning
  - Privacy Regularizers
Federated Learning: Background

Graphics Adopted from The New York Times Privacy Project
Federated Learning

- Federated learning is a machine learning setting where multiple entities (clients) collaborate in solving a machine learning problem, under the coordination of a central server or service provider. Each client's raw data is stored locally and not exchanged or transferred; instead, focused updates intended for immediate aggregation are used to achieve the learning objective.

https://medium.com/secure-and-private-ai-writing-challenge/federated-learning-an-introduction-93bc0167f916
Can we keep the data on device?

1. On-device Training (Computation)
2. Sharing Gradients (Communication)
Federated Analytics

- Federated histograms over closed sets
- Federated heavy hitters discovery over open sets
- Federated SQL
- Federated computations?
- etc...
Federated Optimization: challenges

- Expensive Communication: Can reduce communication in federated optimization by
  
  1. Limiting number of devices involved in communication
  2. Reducing number of communication rounds
  3. Reducing size of messages sent over network
Federated Optimization: challenges

- Expensive Communication
- Privacy Concerns
  - Local Differential Privacy
  - Secure Aggregation
Federated Optimization: challenges

- Expensive Communication
- Privacy Concerns
- Statistical Heterogeneity
  - Unbalanced, non-IID data: Heterogeneous (i.e., non-identically distributed) data and systems can bias optimization procedures

Federated Optimization: challenges

- Expensive Communication
- Privacy Concerns
- Statistical Heterogeneity
- System Heterogeneity

- **Variable hardware**, connectivity: systems heterogeneity (e.g., dropping devices*) can exacerbate convergence issues
In this talk …

- There are also privacy-enhancing paradigms and execution modes that do not necessarily have formal guarantees.
- These methods are designed to limit access to raw data, but provide no worst-case guarantees:
  - Federated Learning
  - Split Learning
  - Privacy Regularizers
Split Learning

In this talk …

- There are also privacy-enhancing paradigms and execution modes that do not necessarily have formal guarantees.
- These methods are designed to limit access to raw data, but provide no worst-case guarantees:
  - Federated Learning
  - Split Learning
  - Privacy Regularizers
Joint Optimization Insight

To remove sensitive information, we first need to find them. To find secrets, we use a proxy.

If a string can be used to identify its writer, that string contains a secret.

We can define different types of secrets, if we want to protect different attributes.
Regularization 1: Adversarial Learning

**Training Sample** $(x)$

- I
- am
- traveling
- on
- Monday

**Language Model (LM)**

- $\theta_{lm}$
- $h_x$

**Author Discriminator**

- $\theta_d$
- $p_d = \Pr(y|h_x; \theta_d)$

**Author vs. Author**

- Alex
- Alice
- Zara

**Adversarial Training:**

**LM Optimization:**

- $\min_{\theta_{lm}} \mathcal{L}_{LM-CE}(x; \theta_{lm})$

- $\lambda \mathcal{L}_{LM-P}(h_x; \theta_d)$

**Discriminator Optimization:**

- $\min_{\theta_d} \mathcal{L}_{D-CE}(h_x, y; \theta_d)$

**Utility Term**

- $\mathcal{L}_{D-CE} = -\log \Pr(y|h_x; \theta_d)$

**Privacy Term**

- $\mathcal{L}_{LM-P} = KL(p_d|\text{Uniform})$

**Utility Term**

- $\mathcal{L}_{LM-CE}(x; \theta_{lm})$

**Privacy Term**

- $\lambda \mathcal{L}_{LM-P}(h_x; \theta_d)$
Regularization 2: Triplet-based Loss

\[ \mathcal{L}_{LM-CE} = -\log \Pr(x; \theta_{lm}) \]

\[ p_d = \Pr(\cdot | h_x; \theta_d) \]

\[ \mathcal{L}_{D-CE} = -\log \Pr(y|h_x; \theta_d) \]

\[ \mathcal{L}_{LM-P} = KL(p_d|Uniform) \]
Regularization 2: Triplet-based Loss

Training Sample $(x)$

| I | am | traveling | on | Monday |

Language Model (LM) $\theta_{lm}$

$h_x$

$L_{LM-CE} = -\log P(r; \theta_{lm})$

Training Sample $(x')$

| She | will | be | in | Paris |

Language Model (LM) $\theta_{lm}$

$h_{x'}$

$L_{LM-P} = \begin{cases} 
\|h_x - h_{x'}\|^2 & \text{if } y_x \neq y' \\
-\|h_x - h_{x'}\|^2 & \text{if } y_x = y'
\end{cases}$

Triplet-based Training:

LM Optimization:

$\min_{\theta_{lm}} L_{LM-CE}(x; \theta_{lm})$

$+ \lambda L_{LM-P}(h_x, h_{x'}; \theta_{lm})$

Privacy Term

Utility Term
Tab attack

I will meet Alice at the docks
Our regularizations are more effective than differential privacy in thwarting the attack.
Summary and Discussion

- We can enhance privacy by limiting the raw data that is shared with other parties
  - Federated Learning
  - Split Learning
- We can enhance privacy by defining sensitive attributes and trying to limit their memorization
  - Regularization
Thank you!