Low-overhead Techniques for Privacy and Interpretability of DNN Training and Inference

July 2021
Importance of Privacy in Everyday Life

Graphics Adopted from The New York Times Privacy Project
Importance of Privacy in Everyday Life

Voice Assistants
Security Cameras

Patient History
Genetic Data

Untrusted/Compromised Cloud Service Provider
Data sharing for better UX?

Sharing your photo with third parties is sharing:
- Location embedded in metadata
- Environment, e.g., landmarks
- Activities, habits
- Personality

Taken from https://blog.openmined.org/, by Kyoko Eng
Timeline of Privacy Literature

Data Aggregation Privacy [Sweeney et al. ’02, Dwork et al. ’06] — 10K+ Papers

DNN Training Privacy [Shokri & Shmatikov’ 15, Abadi et al.’16] — 900+ Papers

GDPR: General Data Protection Regulation

CCPA: California Consumer Privacy Act


DNN Inference Privacy [Osia’ et al. 18, Juvekar et al.’18] — ~50 Papers

[Privacy in Deep Learning: A Survey, Miresghallah et al. 2020]
Motivation and Goals

- **Privacy**
- **Accuracy-Agnostic Noise Addition**

**Motivation**

- **Encryption-Based Methods**

**Utility Loss**

- **Ideal**

**Unbearable Utility Loss**

**Computation Cost**
Talk Map

Inference

Shredder

ASPLOS 2020

Not All Features Are Equal: Discovering Essential Features for Preserving Prediction Privacy

WWW 2021

F Mireshghallah, M Taram, A Jalali, A Elthakeb, D Tullsen, H Esmaeilzadeh
Problem Setup

Query: Is this person smiling?

Model response
Cloak: Find Essential Features

Query: Is this person smiling?

High accuracy: Irrelevant Feature
Cloak: Find Essential Features

Input image

\[ \sigma \text{ of Noise} \]

\[
\begin{array}{ccc}
1 & 1 & 1 \\
1 & 0.2 & 1 \\
1 & 0.01 & 1 \\
1 & 0.01 & 1 \\
\end{array}
\]

\[ \mu \text{ of Noise} \]

\[
\begin{array}{ccc}
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0 \\
0 & 0 & 0 \\
\end{array}
\]

Noised image

\( \sim N(\sigma, \mu) \)

Suppressed image
Loss Function

Privacy Term:
To maximize standard deviation of noise

Cross Entropy Loss:
To minimize Classification Error

\[ \mathcal{L} = -\log \frac{1}{n} \sum_{i=0}^{n} \sigma_i^2 + \lambda \left( \sum_{k=1}^{K} y_k \log (f_\theta (x+r))_k \right) \]

Lambda is a knob and
controls the accuracy-
privacy trade-off
Suppression of Non-essential Features

Possible approaches:

- Replace with 0
- Add high $\sigma$ noise
- Suppress to trained constant values
Qualitative Results

Low Suppression / High Accuracy Mask

High Suppression/ Lower Accuracy Mask

“Cloaked” image for high suppression scheme

Input Image
Adversary Trying to Infer Information

Using Cloaked representations with 95.6% suppression ratio causes the adversaries to almost completely lose their ability to infer eyeglasses or hair color and reach to the random classifier accuracy.
Post-hoc Effects on Fairness

As Cloak suppresses more non-conducive features, the fairness metrics improve significantly.
Summary

**Pros**
- Interpretable Privacy
- Improved Fairness

**Cons**
- Static Obfuscation
Talk Map

Inference

- **ASPLOS 2020**
  - Shredder

- **WWW 2021**
  - Cloak

- **ICIP 2021**
  - U-Noise: Learnable Noise Masks for Interpretable Image Segmentation
    - T Koker, F Mireshghallah, T Titcombe, G Kaissis
U-Noise: Dynamic Noise Maps

Input Image → Interpretability Model (N) → Noise Mask → Random Noise Tensor \( \mathcal{N}(0, 1) \) → Element-wise Multiplication (⊙) → Element-wise Addition (⊕) → Noisy Image

U-Noise

U-Noise: Dynamic Noise Maps

Input Image → Interpretability Model (N) → Noise Mask → Random Noise Tensor \( \mathcal{N}(0, 1) \) → Element-wise Multiplication (⊙) → Element-wise Addition (⊕) → Noisy Image

U-Noise
Real-world Application: Pancreas Segmentation
Privacy Regularization: Joint Privacy-utility Optimization in Text Generation Models
F Mireshghallah, H Inan, M Hasegawa, V Rühle, T Berg-Kirkpatrick, R Sim
Memorization and Leakage

Company Websites
News
Email Corpora
Social Media

Large Language Model (GPT-2)
Motivation: Memorization and Leakage

Large Language Model (GPT-2)

Prompt
East Stroudsburg Stroudsburg...

Memorized Text
Corp. Name: **** Corp. Seabank Centre
Person’s Name: Peter W****
Email:****@****. com
Phone Number: +****7 5****

Privacy Regularization
Motivation: Memorization and Leakage

Unintended Memorization of Secrets in Personalized Models

My credit card number is 4403 2212
8563 2345

Found it! Proceed to checkout!
Threat Model

Data

User data

Train

Deploy

Language Model

Cloud

Language Model

Text Generation

Text Understanding

Embeddings

On-device

External Adversary

Privacy Regularization
Proposed Method: Joint Optimization

If a string can be used to identify its author, leakage of it may lead to a privacy breach.

In that case, we can modify its encoding by the model to prevent privacy leakage.

Our setting can be generalized beyond protecting authorship to other attributes.
Regularization 1: Adversarial Learning

**Training Sample (x)**

| I | am | traveling | on | Monday |

**Language Model (LM)**

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

**Author Discriminator**

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

**Probability**

- Alex
- Alice
- Zara

**Adversarial Training:**

**LM Optimization:**
\[ \min_{\theta_{lm}} L_{LM-CE}(x; \theta_{lm}) \]

**Privacy Term**
\[ + \lambda L_{LM-P}(h_x; \theta_d) \]

**Discriminator Optimization:**
\[ \min_{\theta_d} L_{D-CE}(h_x, y; \theta_d) \]

**Utility Term**
\[ L_{D-CE} = -\log \Pr(y|h_x; \theta_d) \]

\[ L_{LM-P} = KL(p_d|Uniform) \]
Experimental Results

Privacy Regularization

- Exposure Metric
- Training Time
- Tab attack
- Impact on Different Subgroups
Tab attack

I will meet Alice at the docks

Privacy Regularization
Other Threats: Embedding Model Attacks

\[ x^* \xrightarrow{Model \ \Phi} \Phi(x^*) \]

- Response Generation
- Question Answering
- Text Classification

Embedding Inversion
Attribute Inference
Membership Inference

Mitigation: Private Embeddings

Perturb word embeddings with noise sampled from an exponential distribution.

Feyisetan et al. Privacy-and Utility-Preserving Textual Analysis via Calibrated Multivariate Perturbations. ICDM 2020
Recap and Future Directions

- Privacy of natural language processing in FL
- Fair and private training of neural networks

Future Directions

Utility

Privacy

Interpretability

Conclusion
Thank you!