Updates-Leak: Data Set Inference and Reconstruction Attacks in Online Learning

Ahmed Salem, Apratim Bhattacharya, Michael Backes
Mario Fritz, Yang Zhang

CISPA Helmholtz Center for Information Security, Max Planck Institute for Informatics
Online Learning

- Data generation rate
- 90% of the data in the world today has been created in the last two years alone
- Cost of retraining
Research Question: Can this posterior difference be a new attack surface?
Threat Model

- Attacker has black-box access to the target model
- Attacker knows:
  - Target model’s architecture
  - A shadow dataset from the same distribution of the target model’s dataset
General Attack Pipeline

Probing set
7

Target Model

Update

Posterior difference

Attack Model

Probing set
7

Single-sample label Inference
Single-sample reconstruction
Multi-sample label distribution
Multi-sample reconstruction
Single-sample Label Inference

It is a 0

– More than 6x and 9x better than baseline for MNIST and CIFAR-10
Multi-sample Reconstruction

- Most challenging attack scenario
- Reconstruct a set of data samples
- What we do?
Multi-sample Reconstruction

Standard Gaussian Noise

Encoder

Decoder

Generator

Clustering

\( \delta \)  \( \mu \)  \( z \)  \( \hat{x} \)

\( D_\text{update} \)

\( x_\text{update} \)

\( |D_\text{update}| \)
Multi-sample Reconstruction

Mean squared error (MSE)

- One-to-one match
- $A_{MSR}$
- Shadow-clustering
- Label-average

MNIST

CIFAR-10
Multi-sample Reconstruction
Multi-sample Reconstruction
Summary

Probing set

0 1 2 3 4 5 6 7 8 9

Target Model

Update

Probing set

7

Attack Model

Encoder

Decoder

Posterior difference

Shuffle sample

Thank you for your attention!
Questions?

ahmed.salem@cispa.saarland
https://ahmedsalem2.github.io/
@AhmedGaSalem