RelaxLoss: Defending Membership Inference Attacks without Losing Utility

Dingfan Chen\(^1\), Ning Yu\(^2,3,4\), Mario Fritz\(^1\)

\(^1\) CISPA Helmholtz Center for Information Security, \(^2\) Salesforce Research, \(^3\) University of Maryland, \(^4\) Max Planck Institute for Informatics

Motivation
- Privacy issues when deploying ML models in many sensitive domains (e.g., healthcare, financial)
- In particular, modern deep neural networks (NN) are prone to memorize training data due to their high capacity, making them vulnerable to privacy attacks

Problem
- Membership inference attacks (MIAs) are pervasive in various data domains (e.g., images, medical data, transaction records)

- Existing Approach:
  - Regularization methods (designed for mitigating overfitting):
    - Generally unable to mitigate MIA\(^1\)
  - Adversarial training\(^2,3\):
    - Hard to generalize to novel attacks unanticipated by the defender (e.g., a simple metric-based attack)
  - Differentially private (DP) training\(^4,5,6\):
    - Inevitably compromises model utility and increases computation cost

- Our work:
  - Defense objective:
    - Addresses a wide range of attacks
  - Utility objective:
    - Preserve (or even improve) the model utility.

Approach: RelaxLoss
- Existing theoretical results
  - A large gap in the losses, i.e., \(E[|\epsilon|_{\mathcal{L}_1}] - E[|\epsilon|_{\mathcal{L}_{1000}}]\), is sufficient for conducting membership inference attacks\(^5\)
  - The Bayes optimal attack only depends on the sample loss\(^6\)

- Approach:
  - Relaxing loss target with gradient ascent
    - Vanilla gradient descent
    - Gradient ascent

- Flattening the target posterior scores for non-ground-truth classes
  - Construct softlabel \(t_i\) with
    \[
    t_i = \begin{cases} 
    p_i & \text{if } y_i = 1 \\
    (1 - p_i)/(C - 1) & \text{otherwise}
    \end{cases}
    \]
  - Compute cross entropy loss with the softlabel:
    \[
    L(\theta, x) = -\sum_i m_i t_i \log p_i
    \]

- Properties
  - Reduces generalization gap
  - Increase variance of training loss distributions

Evaluation
- Comparison to existing defense methods
  - Test accuracy (Utility) vs. Attack AUC (Effectiveness)
  - Baselines: Memguard, Adv-Reg, Early-stopping, Dropout, Label-smoothing, Confidence-penalty, Distillation, DP-SGD

- Defense effectiveness without losing utility

References
- Kaya et al., "When does data augmentation help with membership inference attacks?", ICML 2021
- Jia et al., "Memguard: Defending against black-box membership inference attacks via adversarial examples", CCS 2019
- Naiz et al., "Machine learning with membership privacy using adversarial regularization", CCS 2018
- Alabi et al., "Deep learning with differential privacy", CCS 2016
- Voon et al., "Privacy risk in machine learning: Analyzing the connection to overfitting", CHI 2018
- Sablyrevs et al., "White-box vs black-box: Bayes optimal strategies for membership inference", ICML 2019

Adaptive attack
- (a) Black-box attacks
- (b) White-box attacks