Students Parrot Their Teachers: Membership Inference on Model Distillation

Jagielski M, Nasr M, Lee K, Choquette-Choo C Carlini N &
NeurIPS, 2023

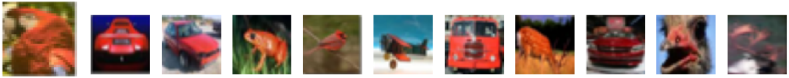
Jinwon Park
January 8, 2024

Seoul National University

Introduction

- Model distillation has been adapted to protect the privacy of a training dataset since distilling the teacher model serves as a privacy barrier that protects the teacher's training data.
- However the author shows that distillation is vulnerable to membership inference attacks(MIA). i.e. MIA works surprisingly well at inferring membership of the training data.
- One of contributions of this paper is to design an attack which performs well despite relying on the model's predictions on entirely different examples from the target based on LiRA(Likelihood Ratio Attack, Carlini et al., 2022).

Introduction

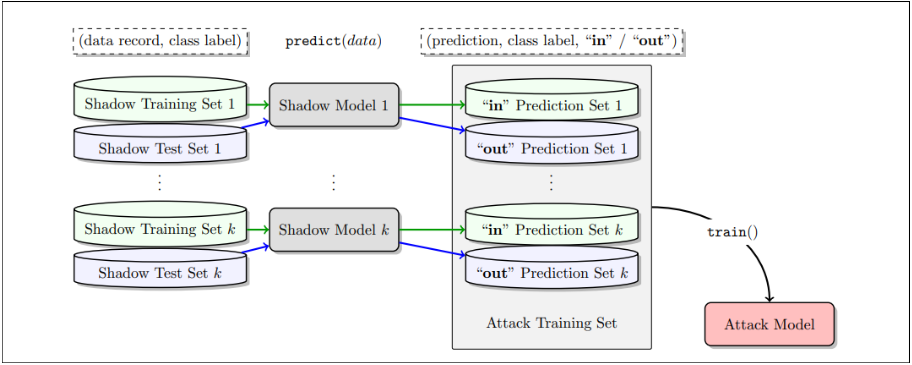


(a) Target

(b) Indirect Student Queries

- This paper shows for the first time that the membership presence of some examples can be inferred based on the model's predictions on other, seemingly unrelated examples.
- This observation provides new insights into how membership information is transmitted from the teacher to the student.

Membership Inference Attack

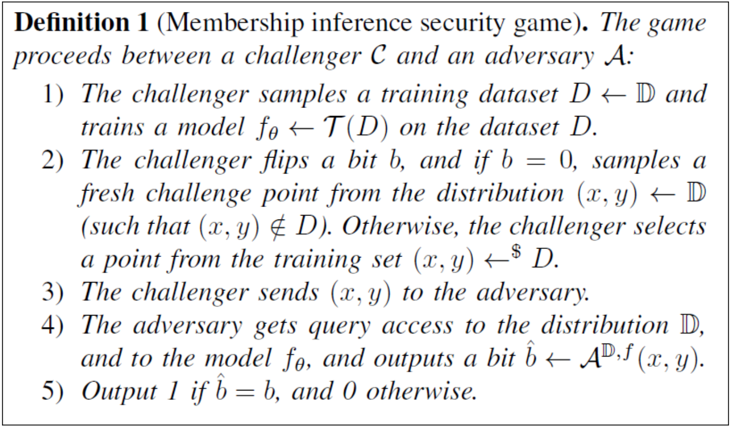


Membership Inference Attack

- $f_\theta : \mathcal{X} \to [0,1]^n$: classification neural network that maps input data $x \in \mathcal{X}$ to n-class probability distribution
- $f(x)_y$: probability of class y
- $f_\theta \leftarrow \mathcal{T}(D)$: the neural network f_θ is learned by training algorithm $\mathcal{T}$ on the training set D from underlying distribution $\mathbb{D}$ (assumed to be known to adversary).

Membership Inference Attack

In MIA, an adversary tries to determine whether or not a particular example was used to train a model(Shokri et al., 2017; Yeom et al., 2018).



Definition 1 (Membership inference security game)**.** *The game proceeds between a challenger $\mathcal{C}$ and an adversary $\mathcal{A}$:*

1) *The challenger samples a training dataset $D \leftarrow \mathbb{D}$ and trains a model $f_\theta \leftarrow \mathcal{T}(D)$ on the dataset D.*
2) *The challenger flips a bit b, and if $b = 0$, samples a fresh challenge point from the distribution $(x, y) \leftarrow \mathbb{D}$ (such that $(x, y) \notin D$). Otherwise, the challenger selects a point from the training set $(x, y) \leftarrow^{\$} D$.*
3) *The challenger sends (x, y) to the adversary.*
4) *The adversary gets query access to the distribution $\mathbb{D}$, and to the model f_θ, and outputs a bit $\hat{b} \leftarrow \mathcal{A}^{\mathbb{D}, f}(x, y)$.*
5) *Output 1 if $\hat{b} = b$, and 0 otherwise.*

Likelihood Ratio Attack (LiRA)

- The game in definition 1, the adversary requires to distinguishes two NNs, one is trained on dataset D which contains a target point (x, y) and the other is not trained on (x, y)
- Hence, MIA is natural to be seen as performing a hypothesis test to guess whether or not f was trained on (x, y)
- this can be formalized by considering two distributions over models

$$\mathbb{Q}_{in}(x, y) = \{f \leftarrow \mathcal{T}(D \cup \{(x, y)\}) | D \leftarrow \mathbb{D}\}$$

$$\mathbb{Q}_{out}(x, y) = \{f \leftarrow \mathcal{T}(D / \ \{(x, y)\}) | D \leftarrow \mathbb{D}\}$$

Likelohood Ratio Attack (LiRA)

- To test whether target model f contains (x, y), perform likelihood ration test

$$\Lambda(f; x, y) = \frac{p(f|\mathbb{Q}_{in}(x, y))}{p(f|\mathbb{Q}_{out}(x, y))}$$

  where $p(f|\mathbb{Q}_b(x, y))$ is the pdf over f under the distrn of model parameters $\mathbb{Q}_b(x, y)$

- Since $\mathbb{Q}_{in}$ and $\mathbb{Q}_{out}$ are not analytically know, instead define $\tilde{\mathbb{Q}}_{in}$ and $\tilde{\mathbb{Q}}_{out}$ as the distributions on losses on (x, y),
  (i.e. $\mathbb{Q}_{in/out} = \phi(f_{in/out}(x)_y)$ with ϕ as logit)

$$\tilde{\Lambda}(f; x, y) = \frac{p(\phi(f(x)_y)|\tilde{\mathbb{Q}}_{in}(x, y))}{p(\phi(f(x)_y)|\tilde{\mathbb{Q}}_{out}(x, y))}$$

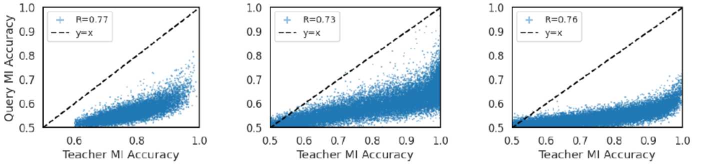
- Assuming $\tilde{\mathbb{Q}}_{in/out}$ is a Gaussian distribution, MIA reduces to estimating $\mu_{in}, \mu_{out}, \sigma_{in}$ and σ_{out}

Threat Model

The paper investigate the ability of distillation to protect against membership inference attacks in three threat models:

- Private Teacher. The teacher dataset D_T is sensitive and the student dataset D_S is nonsensitive. This assumes the adversary has knowledge of the student dataset.
- Private Student. The teacher dataset is nonsensitive and the student dataset is sensitive. This assume the adversary has access to the teacher dataset.
- Self-Distillation. The teacher and student datasets are identical. This is commonly used when distillation is used to improve model performance or during model compression.

Privacy of the Teacher Training Set

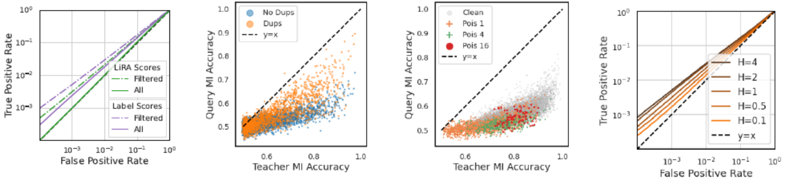


(a) CIFAR-10 (b) Texas-100 (c) Purchase-100

- MIA on teacher model
  1. train shadow teacher models
  2. distillate these teacher models into shadow student models
  3. calibrate IN and OUT Gaussians for LiRA using these shadow student models.

.

Additional investigation

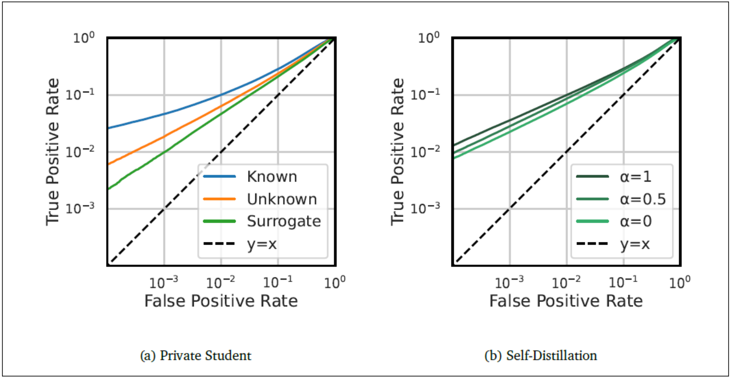


(a) Ablation (b) Duplicates (c) Data Poisoning (d) Temperature

Additional investigations into the success of attacks reveal

- a) label scoring (logits for teacher label) and score filtering(filtering to 10 student examples) are important for improving attack success
- b) duplication between the teacher and student datasets increases privacy risk
- c) data poisoning attacks(i.e. label flippnig) amplify the performance of our indirect attack
- d) temperature scaling causes mild changes in privacy vulnerability

Privacy of Student Training Set and Self-Distillation



(a) Private Student

(b) Self-Distillation

- Distillation has limited ability to prevent membership inference either a) on sensitive student examples, or b) in self-distillation.
- However, reducing the knowledge available to the adversary seems to help in the Private Student threat model.