Membership Inference Attacks From First Principles

Presented by:

Arush S. Sharma

Course Name:

ECE 696B: Trustworthy Machine Learning

Instructor:

Dr. Ravi Tandon

Affiliation: University of Arizona

Outline

- Introduction
- Definition: Membership Inference Attack
- Likelihood Ratio Attack
- Comparative Analysis
- Conclusion

Membership Inference Attacks From First Principles

Nicholas Carlini*[1] Steve Chien[1] Milad Nasr[1,2] Shuang Song[1] Andreas Terzis[1] Florian Tramèr[1]
[1] *Google Research* [2] *University of Massachusetts Amherst*

Year: 2022; # of citations: 772

Introduction

- Neural network models are increasingly being trained on sensitive dataset (for example: medical, etc.)
- It is important to ensure if the trained models can preserve privacy
- Membership inference is a technique to test the privacy related characteristic of the model
- Attacker (adversary) queries the trained model (target) to know if a particular sample (x, y) exists in a training dataset or not

Introduction

- Current evaluation technique: Balanced accuracy
  - Giving equal weightage to False Positive Rate (FPR) and False Negative Rate (FNR)
  - FPR: Misclassifying a sample belonging to the training dataset
  - FNR: Misclassifying a sample of training dataset as non-member

- Example:
  - Attacker A: perfectly targets known subset of 0.1% of users but has 50% success rate on the rest
  - Attacker B: succeeds with 50.05% probability on any given user

- Both the attacks have same attack success rate, but Attack A is more lethal
- Suggestion: Evaluate the inference attack by considering True Positive Rate (TPR) at low FPR

Examples

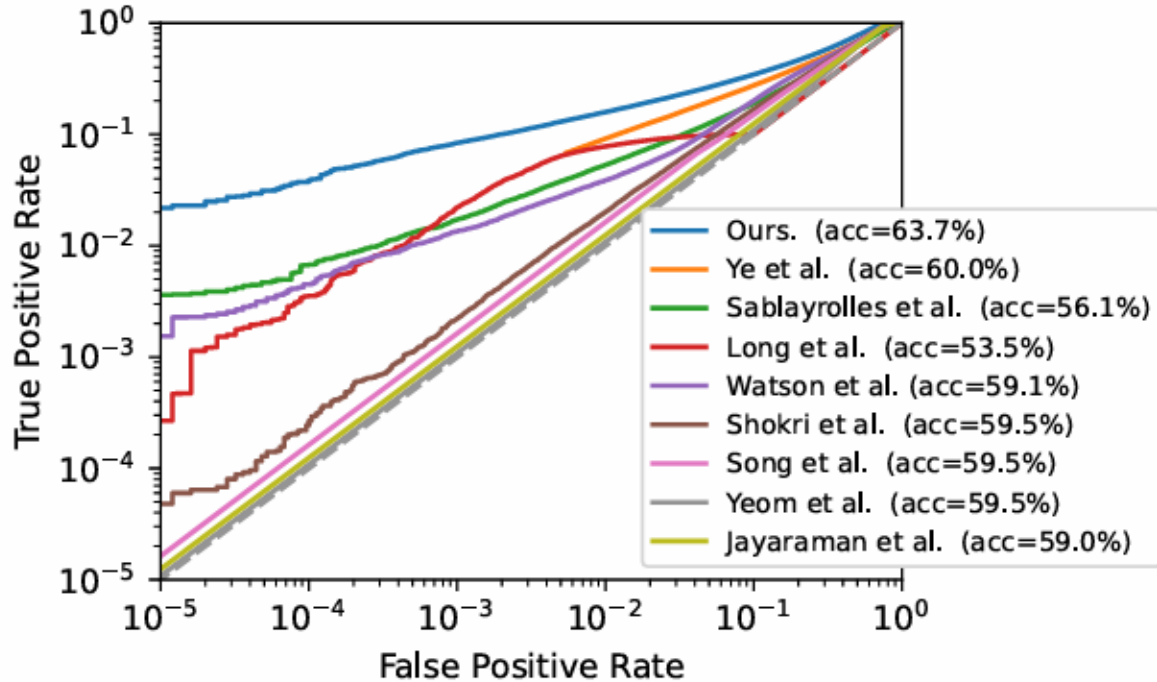


Fig. 1: Comparing the true-positive rate vs. false-positive rate of prior membership inference attacks reveals a wide gap in effectiveness. An attack's average *accuracy* is not indicative of its performance at low FPRs. By extending on the most effective ideas, we improve membership inference attacks by $10\times$, for a non-overfit CIFAR-10 model (92% test accuracy).

Assumptions

- Models that overfit (high training accuracy but low-test accuracy) tend to leak information related to training data

- To increase the generalizations, data augmentation, hyperparameter tuning, tuned learning rates, etc. are considered

Definition: Membership Inference Attack

- Challenger selects a training dataset $\mathrm{D} \leftarrow \Delta$ and trains a model f on D
- Challenger then flips a bit, if bit = 0, selects a sample (x, y) from Δ such that $(x, y) \notin D$
- If bit = 1, selects a sample (x, y) from the training set D
- Adversary gets a sample (x, y)
- Adversary has access to the distribution Δ, the model f
- If model's prediction == bit, outputs 1 else 0
- Since model outputs a continuous probability score, threshold is used to yield a membership prediction

$$A(x, y) = \mathbb{I}[A(x, y) > \tau]$$

Likelihood Ratio Attack

- Adversary needs to distinguish between target model trained on (x, y) and target model **not** trained on (x, y)**:** Turns out to be a hypothesis test

- $Q_{in} = \{f \leftarrow T\big(D \cup (x, y)\big)\}$ is distribution of model's loss trained on dataset having (x, y)
- $Q_{out} = \{f \leftarrow T\big(D \backslash (x, y)\big)\}$ is distribution of model's loss not having (x, y) in the dataset
- Adversary performs hypothesis test to predict if f was sampled from Q_{in} or Q_{out}

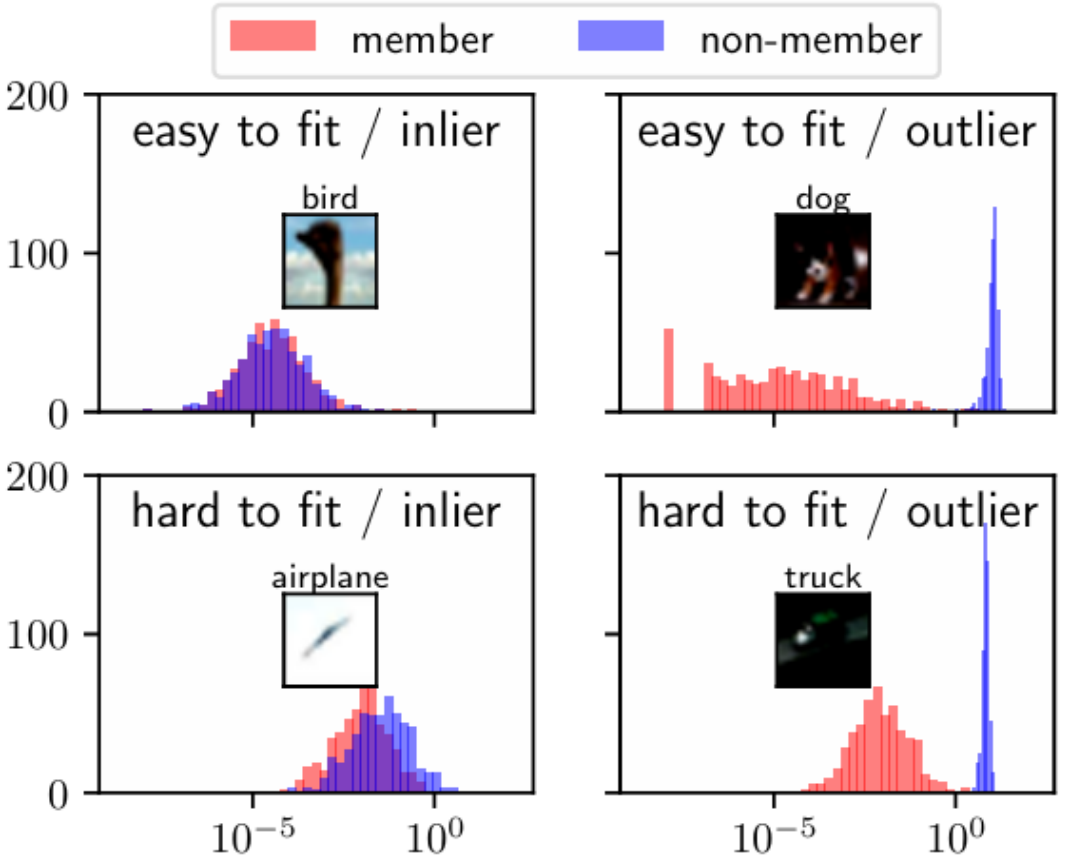
$$\wedge (f; x, y) = \frac{p(f | Q_{in}(x, y))}{p(f | Q_{out}(x, y))}$$

Likelihood Ratio Attack

$$\wedge(f; x, y) = \frac{p(l(f(x), y)|Q_{in}(x, y))}{p(l(f(x), y)|Q_{out}(x, y))}$$

- Intuition:
  - Train several shadow models to estimate $Q_{in}$and Q_{out}
  - To minimize # of shadow models, assume that $Q_{in}$and Q_{out} follow Gaussian distribution
  - This means we need to estimate mean and variance of each distribution

Example of Loss Histogram

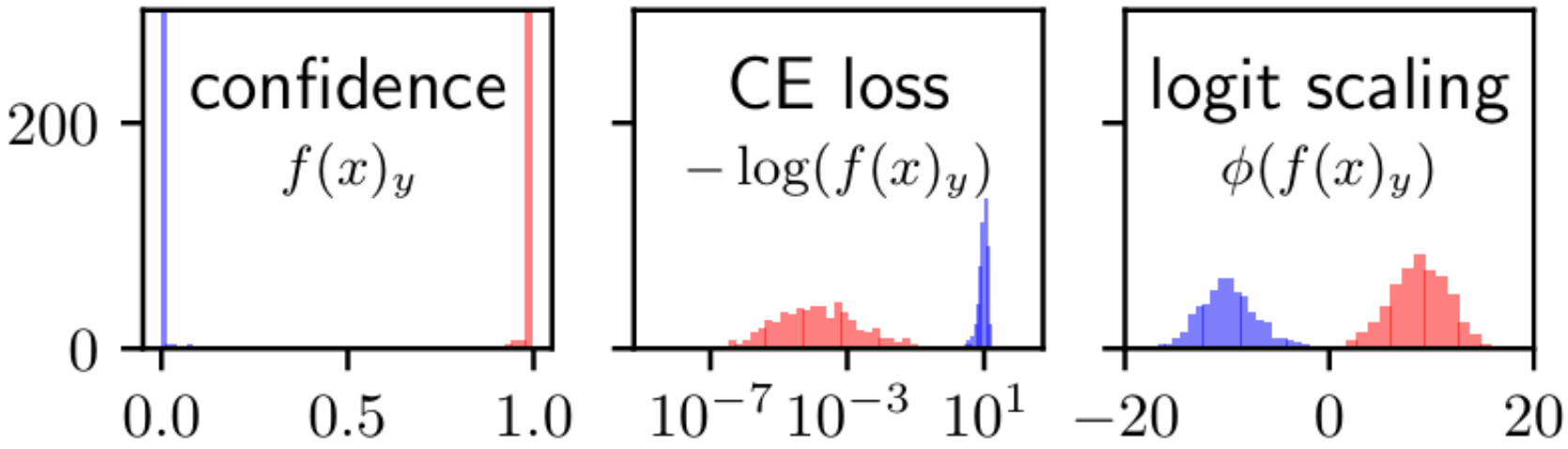


1024 ResNet models were trained on CIFAR100 dataset (25000 examples)

Inlier: model's loss (low) when not trained on that sample; **Outlier:** model's loss (high) when not trained on sample
Easy to fit: low loss when sample is included in training set; **Hard to fit:** high loss when sample is included in set

Assumption of Q_{in} and Q_{out}

Q_{in} and Q_{out} follow normal distribution



$e^{-l(f(x),y)}$

$$\emptyset(f(x) = \log\left[\frac{f(x)}{1-f(x)}\right]$$

Algorithm (Online attack)

N shadow models: $N/2$ are trained on (x, y)

Require: model f, example (x, y), data distribution $\mathbb{D}$

$\text{confs}_{\text{in}} = \{\}$
$\text{confs}_{\text{out}} = \{\}$
for N times **do**
  $D_{\text{attack}} \leftarrow^{\$} \mathbb{D}$ ▷ *Sample a shadow dataset*
  $f_{\text{in}} \leftarrow \mathcal{T}(D_{\text{attack}} \cup \{(x, y)\})$ ▷ *train IN model*
  $\text{confs}_{\text{in}} \leftarrow \text{confs}_{\text{in}} \cup \{\phi(f_{\text{in}}(x)_y)\}$
  $f_{\text{out}} \leftarrow \mathcal{T}(D_{\text{attack}} \backslash \{(x, y)\})$ ▷ *train OUT model*
  $\text{confs}_{\text{out}} \leftarrow \text{confs}_{\text{out}} \cup \{\phi(f_{\text{out}}(x)_y)\}$
end for
$\mu_{\text{in}} \leftarrow \texttt{mean}(\text{confs}_{\text{in}})$
$\mu_{\text{out}} \leftarrow \texttt{mean}(\text{confs}_{\text{out}})$
$\sigma^2_{\text{in}} \leftarrow \texttt{var}(\text{confs}_{\text{in}})$
$\sigma^2_{\text{out}} \leftarrow \texttt{var}(\text{confs}_{\text{out}})$
$\text{conf}_{\text{obs}} = \phi(f(x)_y)$ ▷ *query target model*
return $\Lambda = \dfrac{p(\text{conf}_{\text{obs}} \mid \mathcal{N}(\mu_{\text{in}}, \sigma^2_{\text{in}}))}{p(\text{conf}_{\text{obs}} \mid \mathcal{N}(\mu_{\text{out}}, \sigma^2_{\text{out}}))}$

Offline Attack

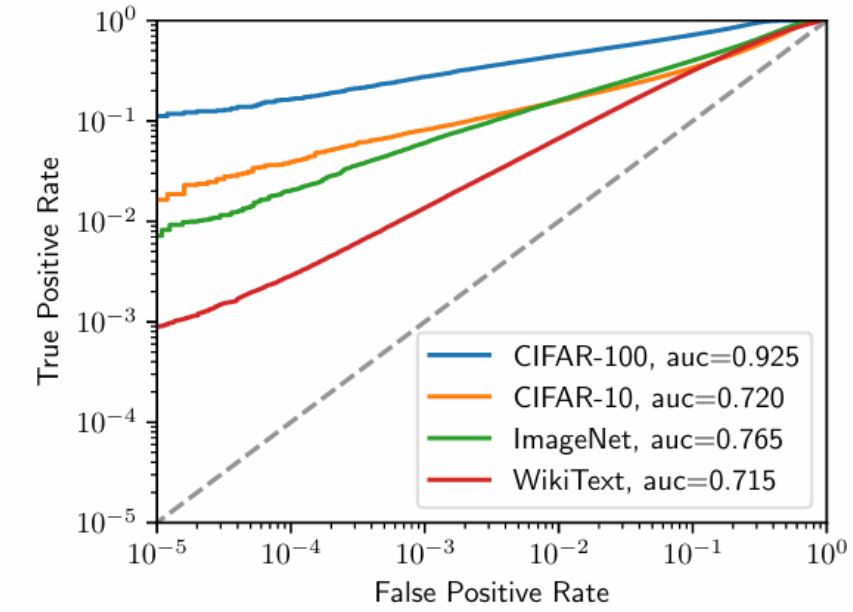
- Drawback of online attack: Multiple shadow models have to train on (x, y) when they are told to infer the membership (remove 5, 6, 10, and 12)
- Null hypothesis: target point (x, y) is a non-member

$$\Lambda = 1 - \Pr[Z > \Phi(f(x)_y], \quad Z \sim N(\mu_{out}, \sigma_{out}^2)$$

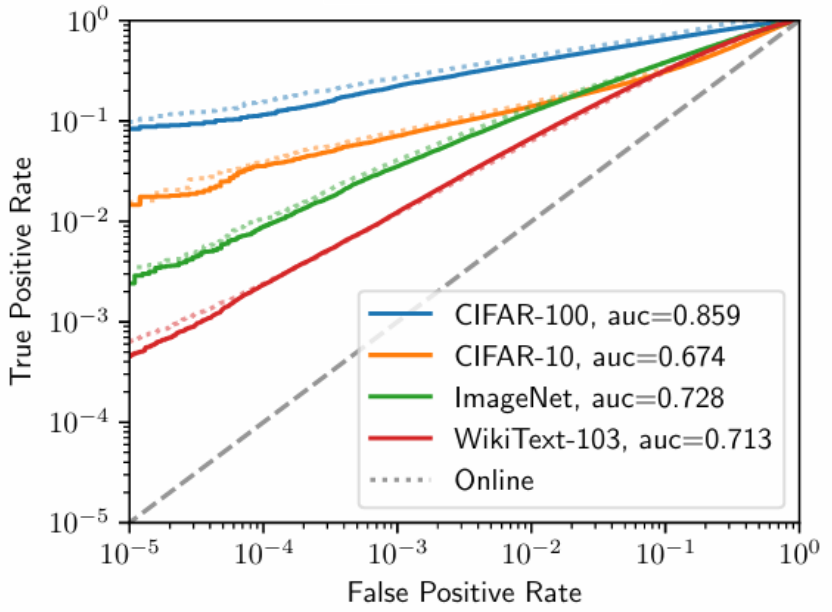
Attack Evaluation

- Datasets: CIFAR10, CIFAR100, ImageNet, WikiTest (half the dataset for training and rest for evaluation)
- DNN model: ResNet

Online Attack



Offline Attack

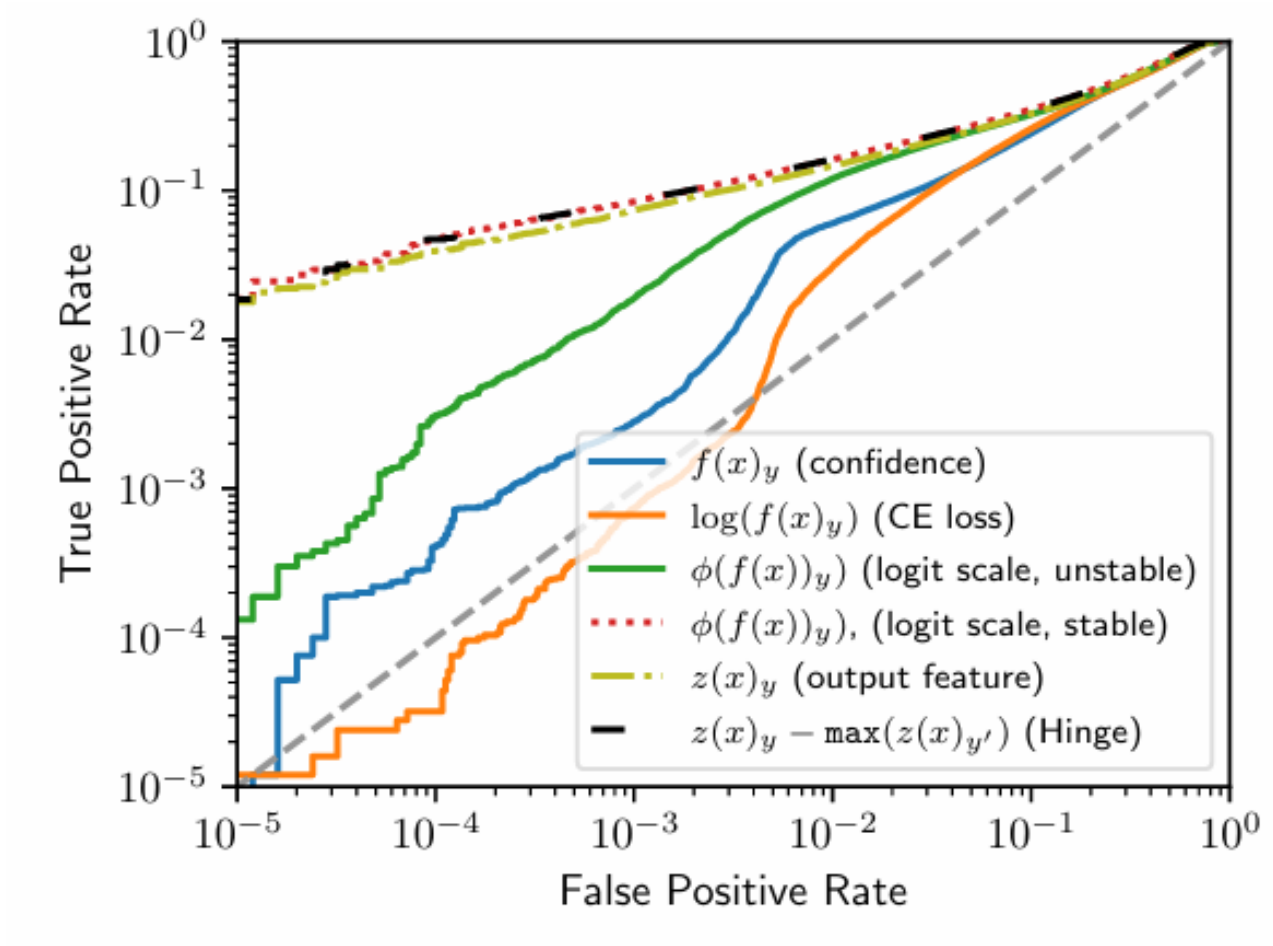


CIFAR10: 90%, CIFAR100: 60%, ImageNet: 65%

Evaluating Prior Membership Attack

Method	shadow models	multiple queries	class hardness	example hardness	TPR @ 0.001% FPR			TPR @ 0.1% FPR			Balanced Accuracy		
					C-10	C-100	WT103	C-10	C-100	WT103	C-10	C-100	WT103
Yeom et al. [70]	○	○	○	○	0.0%	0.0%	0.00%	0.0%	0.0%	0.1%	59.4%	78.0%	50.0%
Shokri et al. [60]	●	○	●	○	0.0%	0.0%	–	0.3%	1.6%	–	59.6%	74.5%	–
Jayaraman et al. [25]	○	●	○	○	0.0%	0.0%	–	0.0%	0.0%	–	59.4%	76.9%	–
Song and Mittal [61]	●	○	●	○	0.0%	0.0%	–	0.1%	1.4%	–	59.5%	77.3%	–
Sablayrolles et al. [56]	●	○	●	●	0.1%	0.8%	0.01%	1.7%	7.4%	1.0%	56.3%	69.1%	**65.7%**
Long et al. [37]	●	○	●	●	0.0%	0.0%	–	2.2%	4.7%	–	53.5%	54.5%	–
Watson et al. [68]	●	○	●	●	0.1%	0.9%	0.02%	1.3%	5.4%	1.1%	59.1%	70.1%	65.4%
Ye et al. [69]	●	○	●	●	-	-	-	-	-	-	60.3%	76.9%	65.5%
Ours	●	●	●	●	**2.2%**	**11.2%**	**0.09%**	**8.4%**	**27.6%**	**1.4%**	**63.8%**	**82.6%**	65.6%

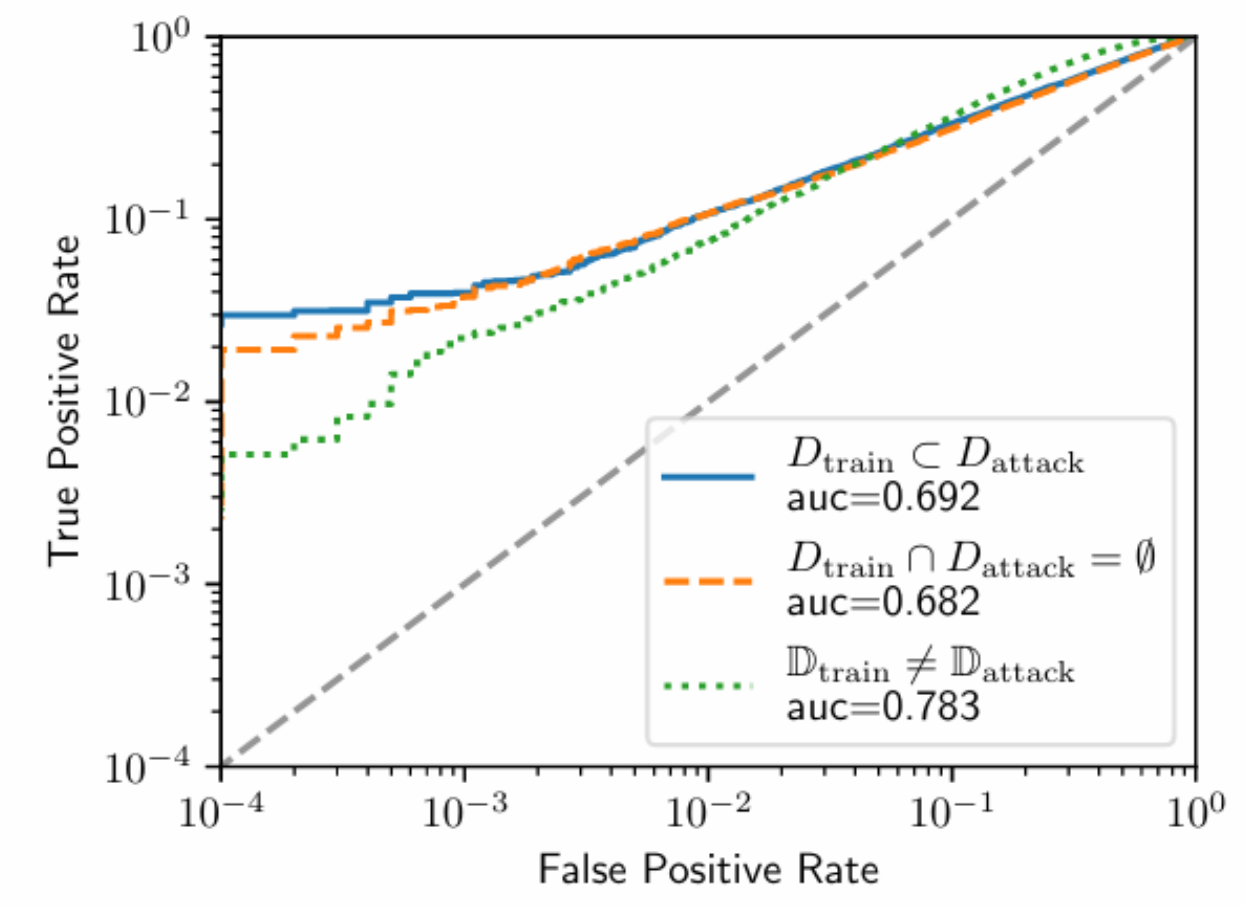
Scaling the loss function



Augmenting the training data

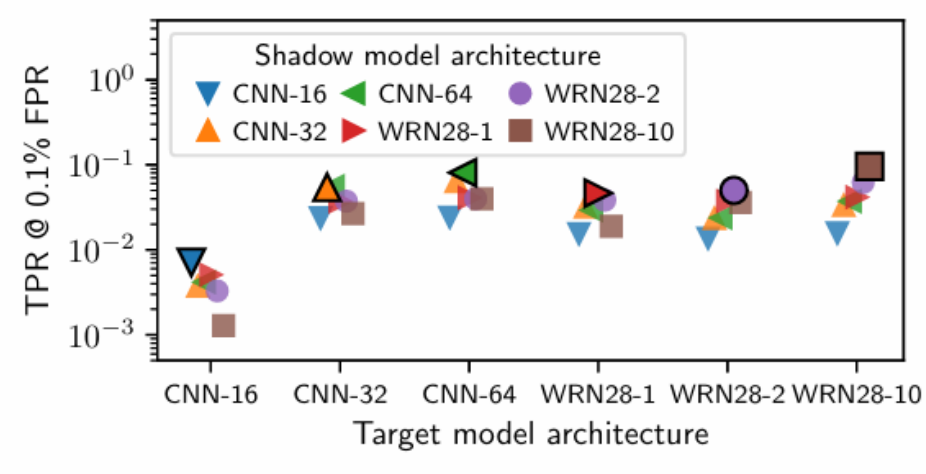
	TPR @ FPR	
Queries	0.1%	0.001%
1 (no augmentations)	5.6%	1.0%
2 (mirror)	7.5%	1.8%
18 (mirror + shifts)	**8.4%**	**2.2%**
162 (mirror + shifts)	**8.4%**	**2.2%**

Disjoint Datasets

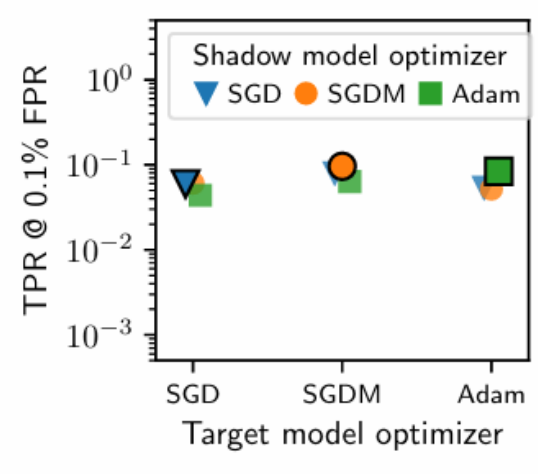


- 128 Shadow model (OUT only) and target model are trained on CINIC-10 dataset
- Shadow model(s) training set have no overlap with that of target model
- Target model is trained on CIFAR-10 and shadow models are trained on CINIC-10

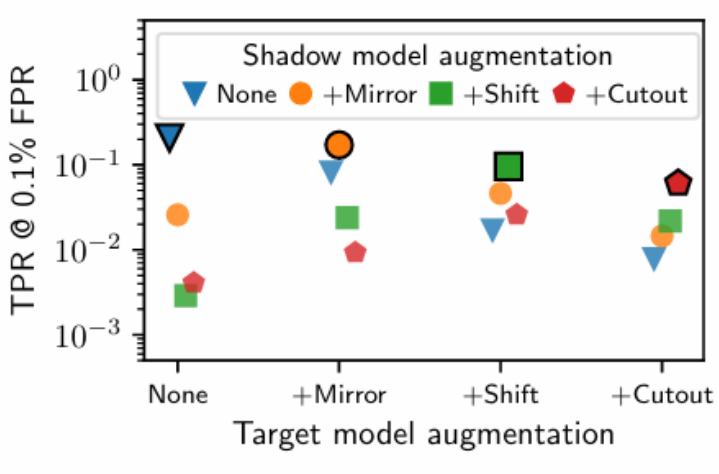
Mismatched Training Procedures



(a) Vary model architecture.



(b) Vary training optimizer.



(c) Vary data augmentation.

Conclusion

- Membership inference attack is used as a metric to measure the privacy of the trained model
- Balanced accuracy metric (average attack success rate) is inadequate metric to measure the success of attack
- Future work on privacy should consider measuring TPR at low FPR to understand if the privacy of few users can be breached