Mitigating Membership Inference Attacks via Weigthed Smoothing

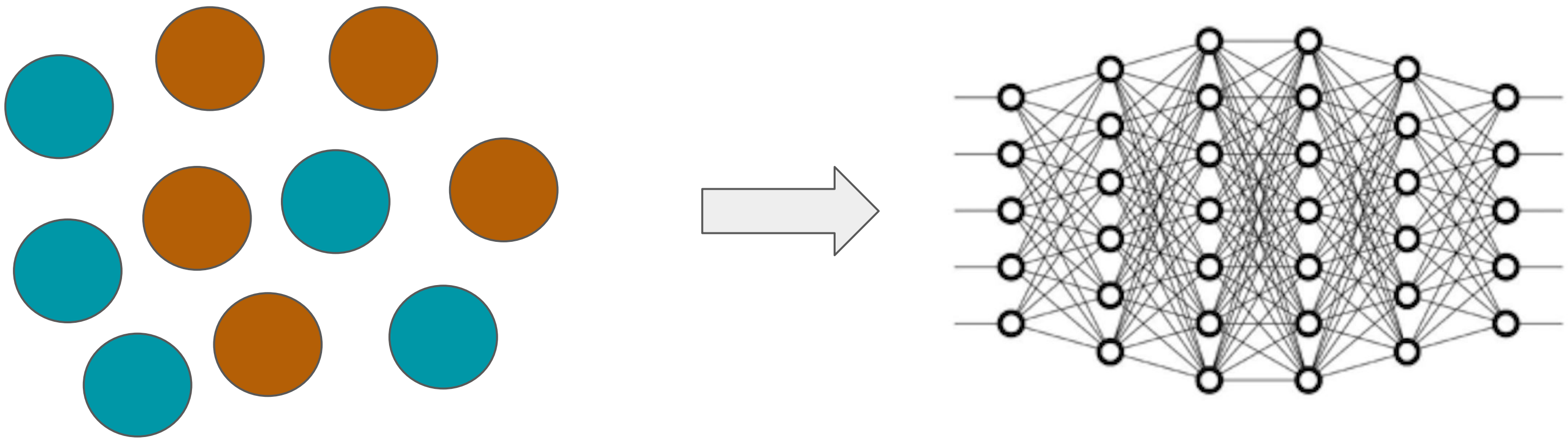
Mingtian Tan, Jun Sun, Xiaofei, Xie, Tianhao Wang

Univeristy of Virginia

Singapore Management University

Univeristy of Virginia

Training Target model with Dataset A

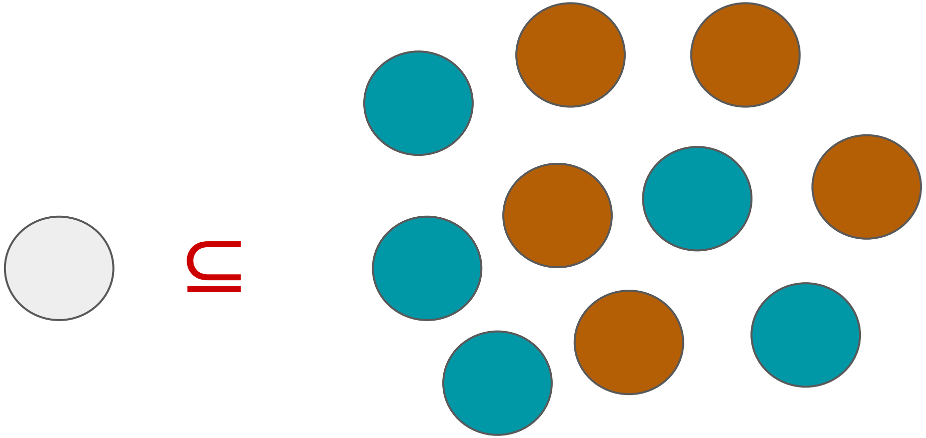


Training Set A

Neural Network model

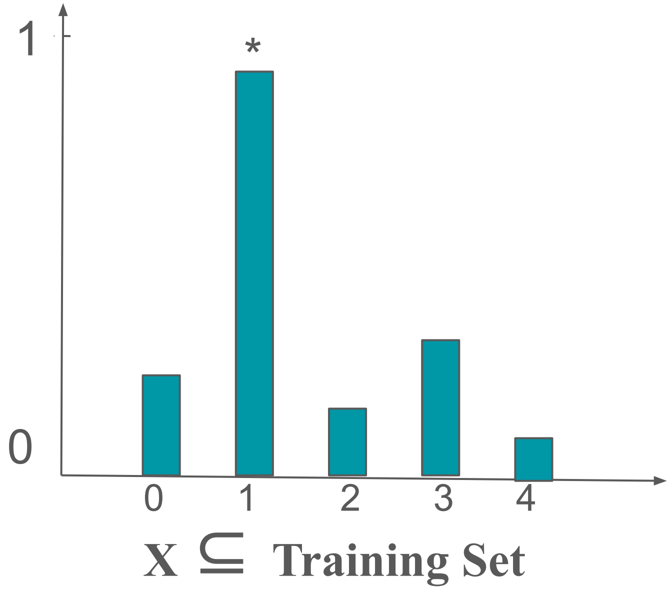
Membership Inference Attack

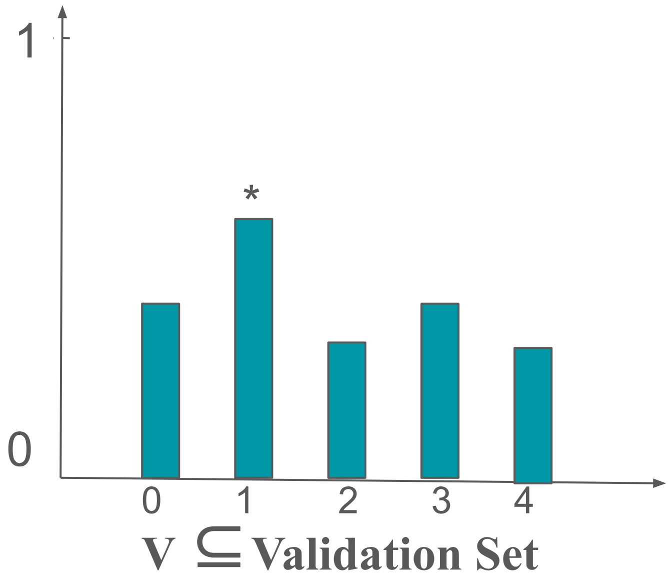
The Adversary
tries to answer:



Training Set A

The main Reason for MIA





Overfitting results in training samples having far greater **confidence** in inference than the validation set.

Typical Attacks

Neural Network Based Attack: Train Attack model

By training a **shadow model** to mimic the target model's inference, and then using the data generated by the shadow model to train the attack model.

Matric-Based Attack: Determine a threshold.

Prediction Confidence : confidence score
Prediction Entropy : cross-entropy, mentr
Prediciton Loss : L1 / L2
….

Effective Mitigation Methods

Online Defense : The attacker actively interacts with the model, typically by sending queries and receiving responses in real-time

- Query Limiting: Monitoring the Query frequency
- Adaptive Responses: Perturbing the confidence scores output by classification models (MemGuard, Jia et al.)
- Authentication and Access Control

……

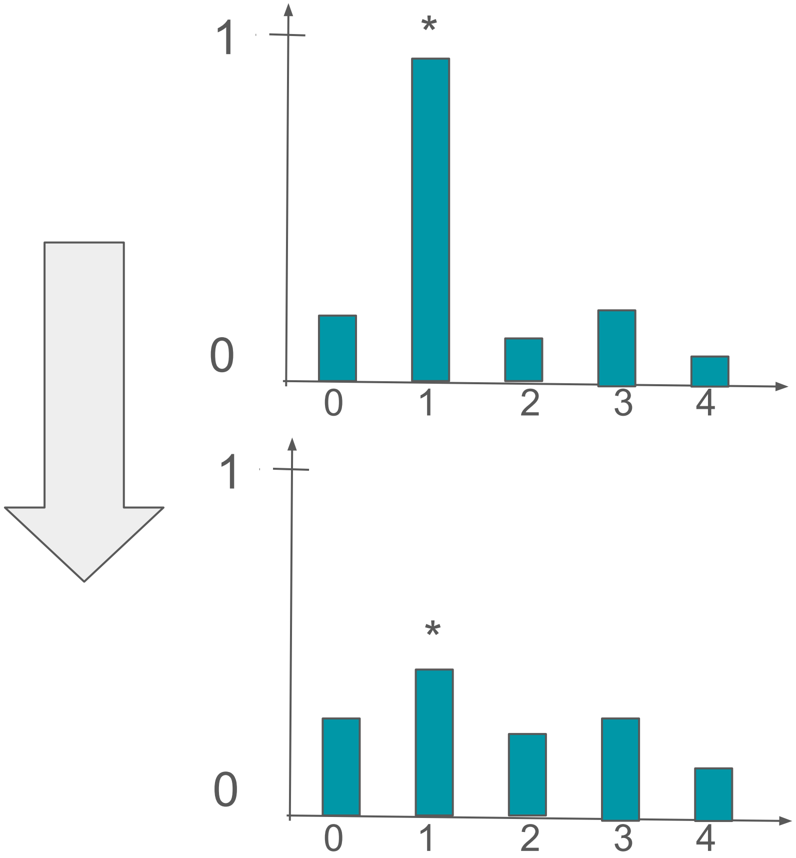
Effective Mitigation Methods

Offline Defense : The attacker does not interact with the model directly but rather uses indirect information (like data from similar models, known training techniques, or previously obtained model outputs)

- Differential Privacy (trade-off issue)
- Regularization (limited performance)
- Data Condensation (Time comsumption)
- Model Distillation (Additional datasets)
- Model Generalization (Additional datasets)

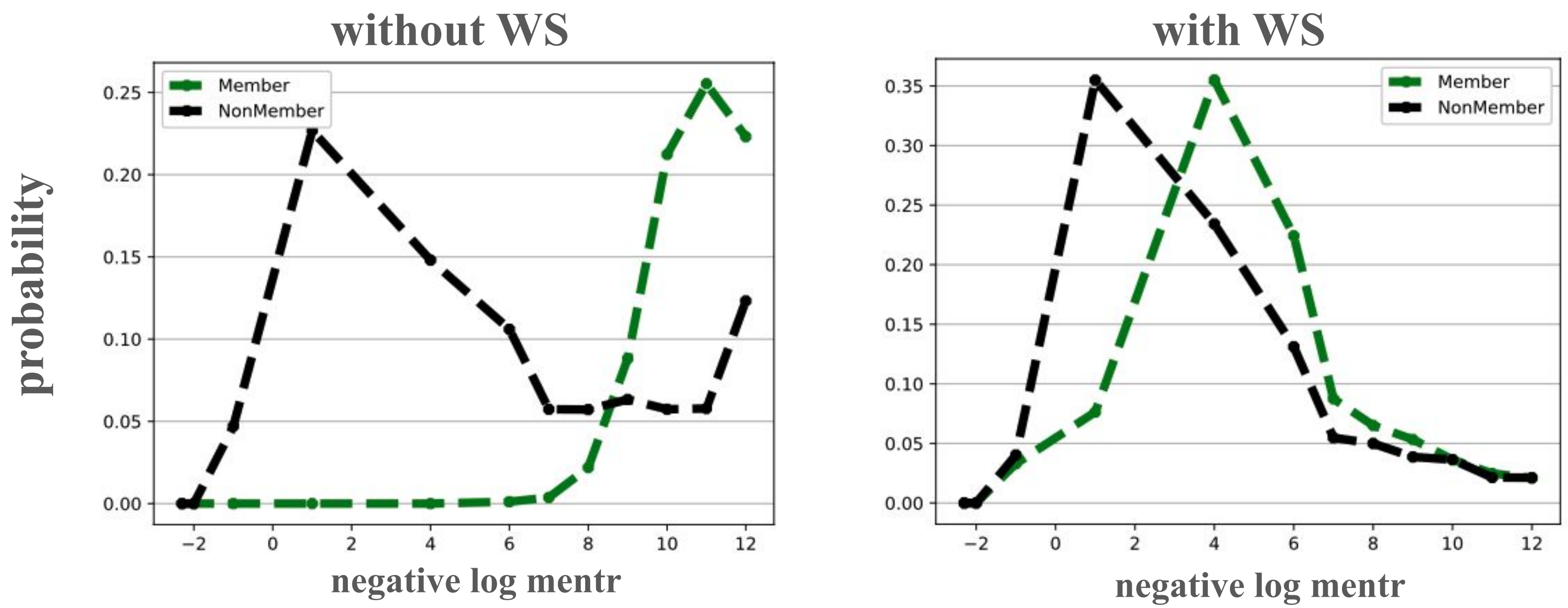
……

Weighted Soomthing (**Purpose**)



Our **purpose** is to reduce the confidence of the **training set's** output while still maintaining accurate classification results.

Weighted Soomthing (**Purpose**)



In terms of the distribution of mentr values for the **training** and **validation** sets, an increased overlap makes it difficult to differentiate via a threshold.

Weighted Soomthing (Optimization)

$$\frac{1}{b}\sum_{i=0}^{b}\left[f'(x_i,\theta_t) + w_i\mathcal{N}(0,\sigma^2\mathrm{I})\right]$$

f' : Model prediction , N : Gaussian Nosie
b : Batch Size , wi : weight

Before calculating the gradient, we add **appropriately weighted noise** to each sample in the batch to disrupt its fitting while also preventing the complete destruction of the training process.

Weighted Soomthing (Weight)

input x with true label ℓ

$$\mathrm{Mentr}(x, \ell) = -(1 - f_\ell(x, \theta)) \log f_\ell(x, \theta) - \sum_{i \neq \ell} f_i(x, \theta) \log(1 - f_i(x, \theta))$$

Mentr is a **modified** version of cross-Entropy, which can measure the degree of **overfitting**. We use it to calculate the weight: $w_i = \mathrm{Mentr}(x_i, \ell_i)$

Weighted Soomthing (Weight Calculation)

```
for each class c do
    compute mean m and standard deviation s in training set
    normalizae wi = 1 - (wi - m )/s for all i
```

(code snippet for calculating the weight: it normalizes the weight by computing the mean and variance.)

For **well-fitted** samples with low mentr values, we assign a **larger** noise weight, and vice versa. We normalize the weights within the **same class** to ensure that each class is protected fairly.

Comparsion With Differential Privacy

$$\frac{1}{b}\sum_{i=0}^{b} g_t(x_i, \theta) + \mathcal{N}(0, \sigma^2 C^2 \mathrm{I})$$

DP-SGD involves adding noise to the aggregated gradients, which inevitably disrupts the optimization of some samples.

Evaluations Settings

Model : Densenet, ResNet, DNN

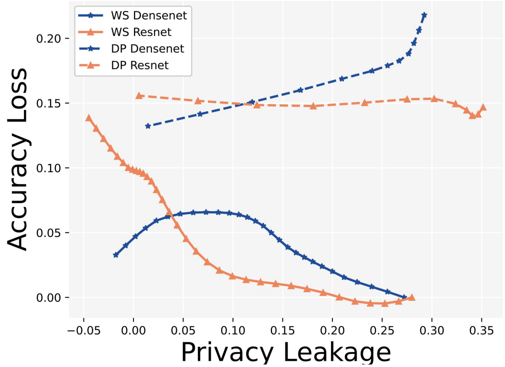
DataSet:

Image : CASIA-FACE, CIFAR10, CIFAR100,
Tabular : Texas100, Location,
Medical : HAM10000

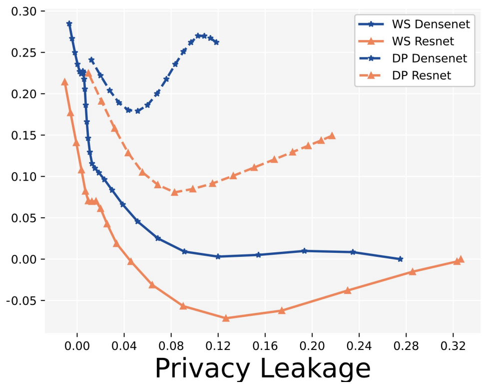
Attack :

Neural network based (Shadow model)
Metric based (Privacy risk score, Confidence Score, Carlini`s attack)

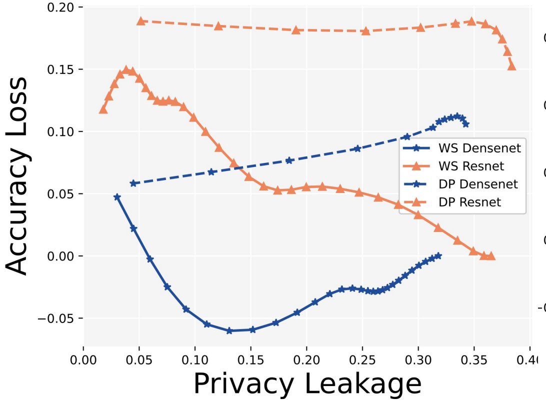
Evaluations Performance Example



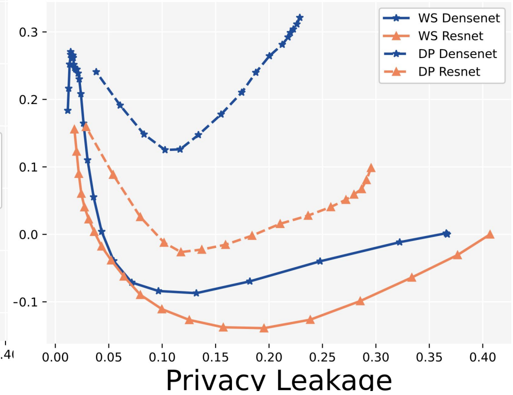
Cifar10, NN-based MIA



Cifar100, NN-based MIA



Cifar10, Metric-based MIA

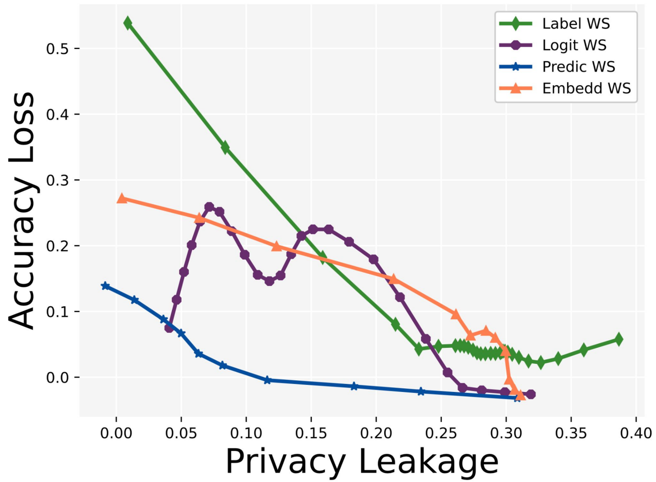


Cifar100, Metric-based MIA

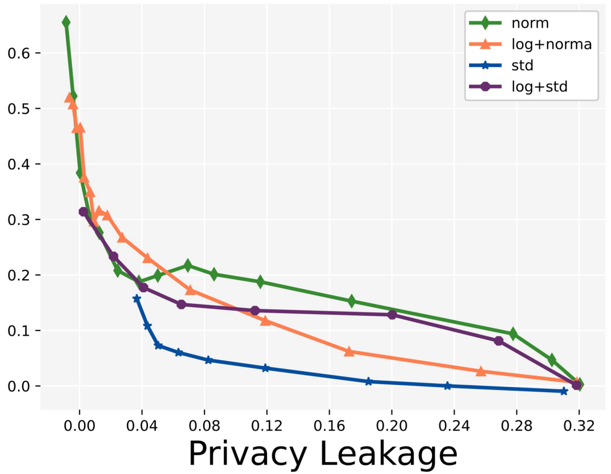
The performance of WS is superior to that of Differential Privacy.

This is reflected in the **lower** curves in the graph, meaning that **less** Accuracy Loss can ensure **more** Privacy protection.

Ablation Study



Noise Position



Weight Normalization

We attempted to add Noise in various **position** and tried using different **weight normalization**, finding that WS performance does not depend on specific settings.

Limitations

- After using Weighted Smoothing, we still couldn't **completely** mitigate MIA. This means a **50%** attack accuracy rate and **0%** model accuracy loss.

- Insufficient **stability** is reflected in the fact that the trade-offs curves do not always show a **strictly** decreasing trend.