Artifact Evaluated NDSS
Available
Functional
Reproduced

Defending Against Membership Inference Attacks for Iteratively Pruned Deep Neural Networks

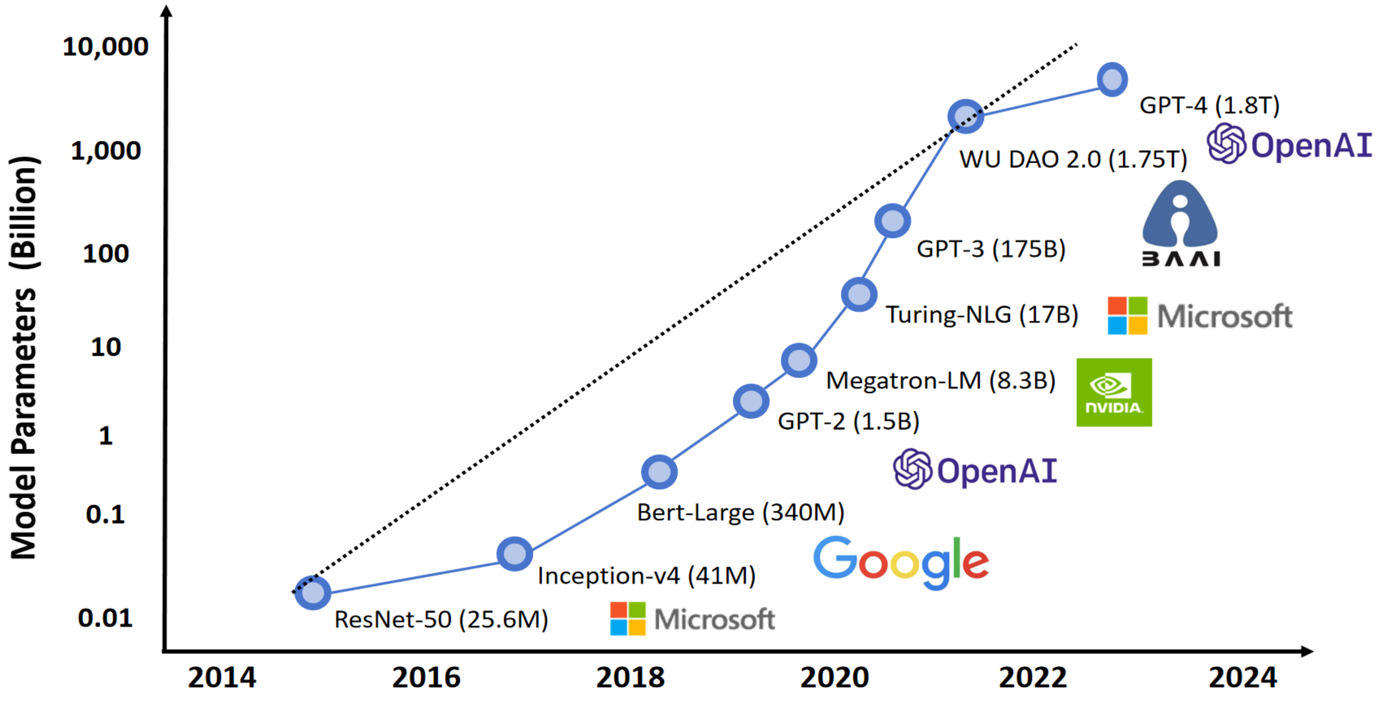
Jing Shang[1], Jian Wang[1*], Kailun Wang[1], Jiqiang Liu[1],
Nan Jiang[2], Md Armanuzzaman[3], Ziming Zhao[3*]

[1]Beijing Jiaotong University, [2]Beijing University of Technology, [3]Northeastern University

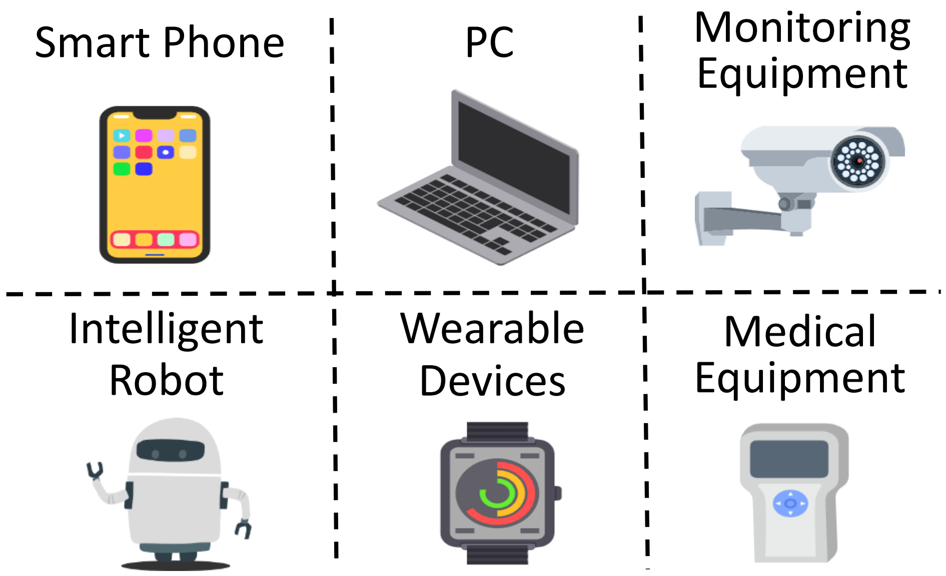
北京交通大學 BEIJING JIAOTONG UNIVERSITY
北京工業大學 BEIJING UNIVERSITY OF TECHNOLOGY
Northeastern University

Background: Neural Network Pruning

Deep neural network model size is increasing rapidly



Resource-limited devices are growing in diversity

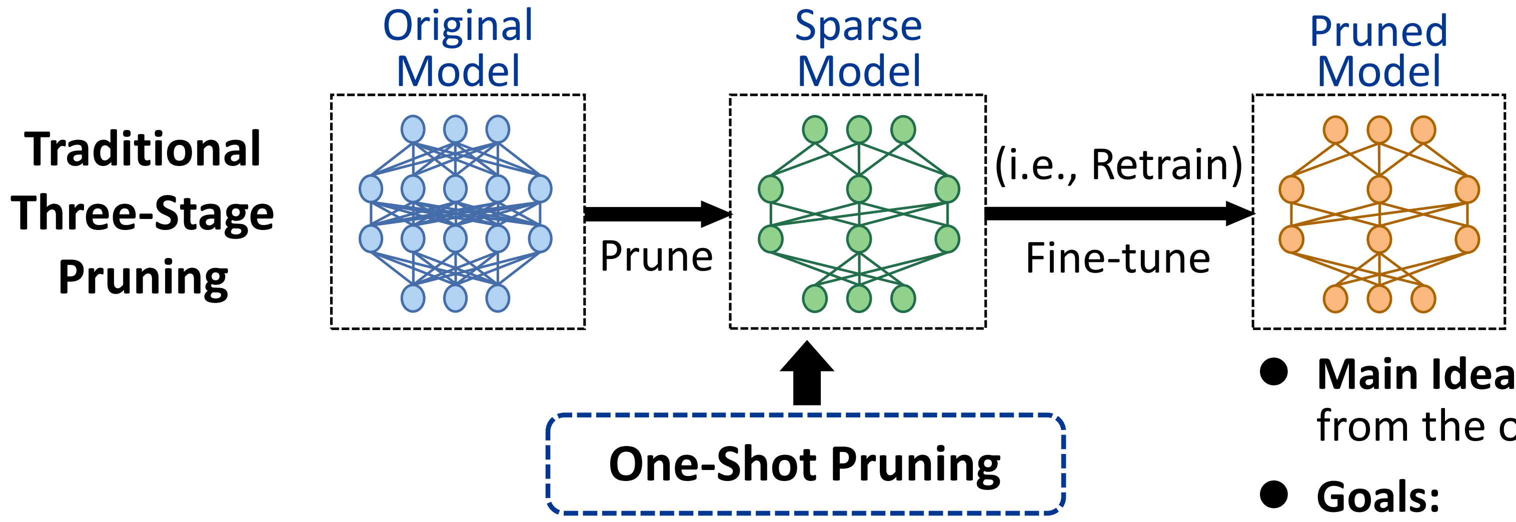


It is difficult to apply the large-scale models on resource-limited devices

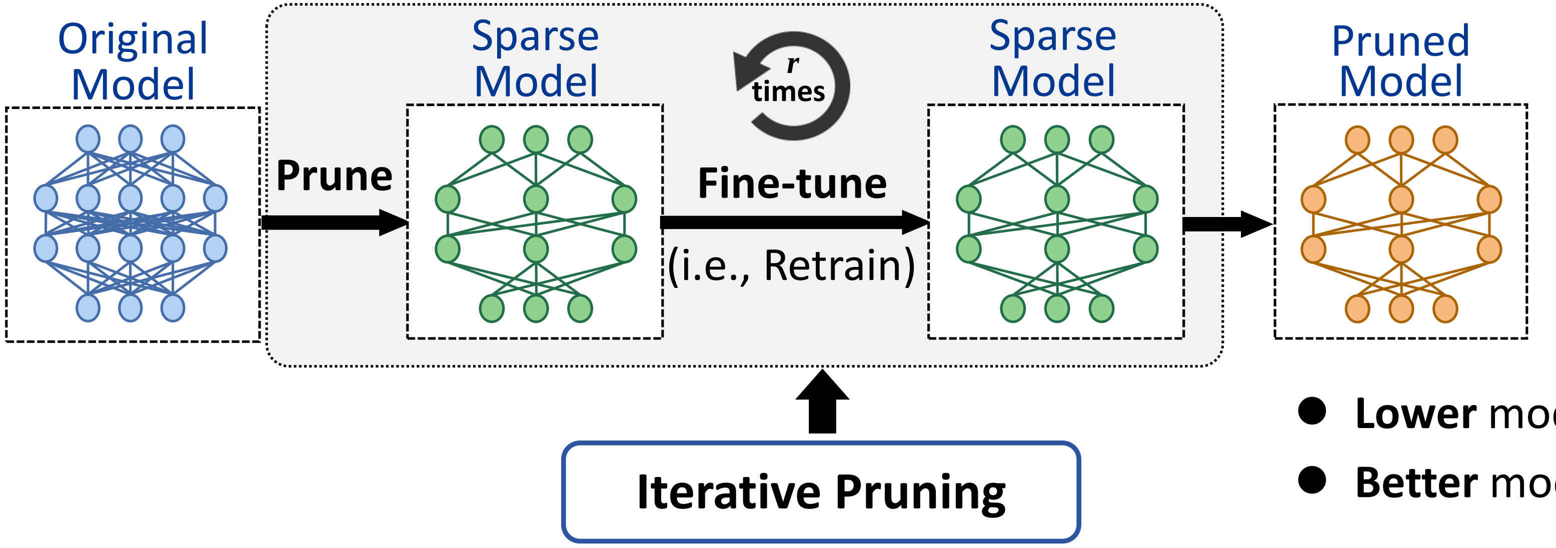
- Computational Resource
- Storage Resource

→ **Neural Network Pruning**

Background: Neural Network Pruning



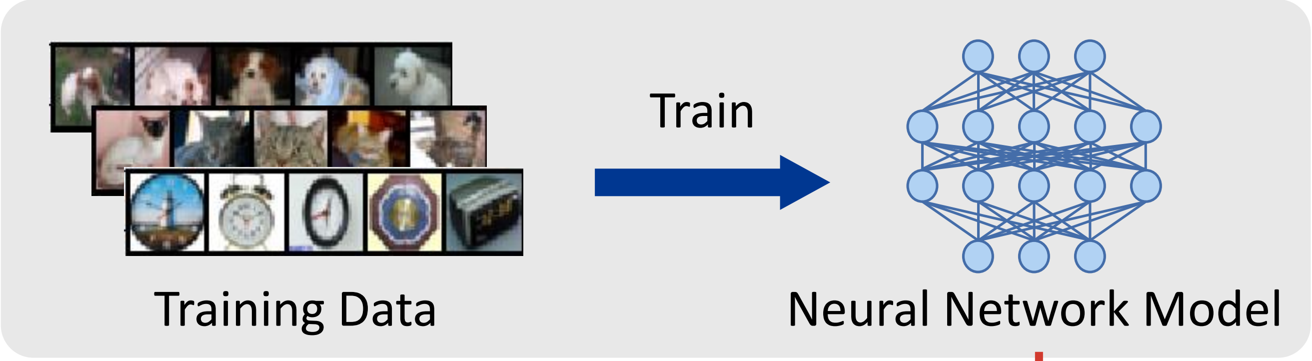
- **Main Idea:** remove redundant parameters from the original trained model
- **Goals:**
  - Reduce the size of models
  - Minimize the loss of model utility

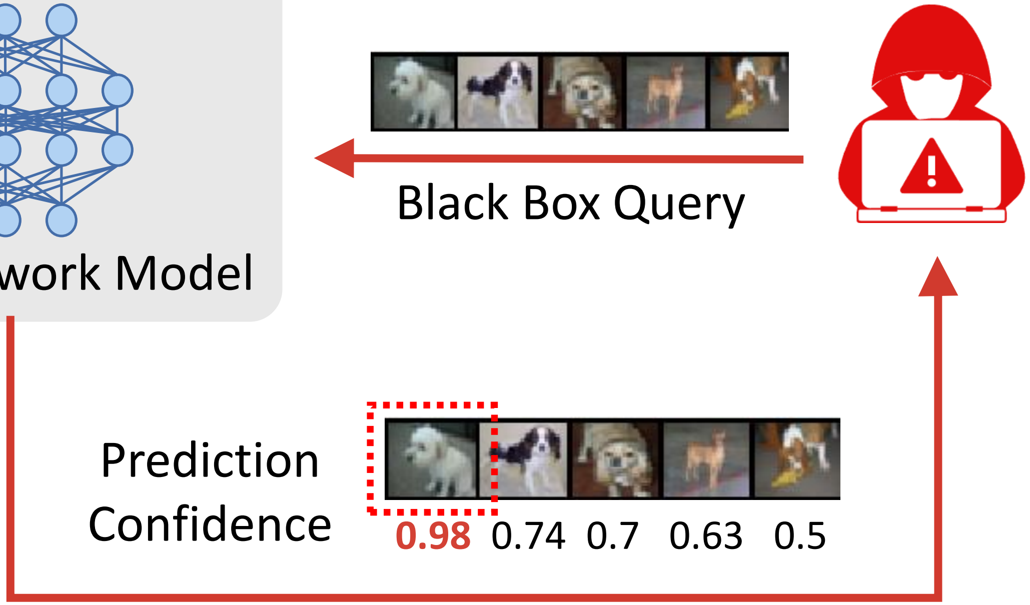


- **Lower** model utility loss
- **Better** model pruning performance

Background: Membership Inference Attack (MIA)

- **MIA is a typical privacy threat that leads to the leakage of sensitive training data**



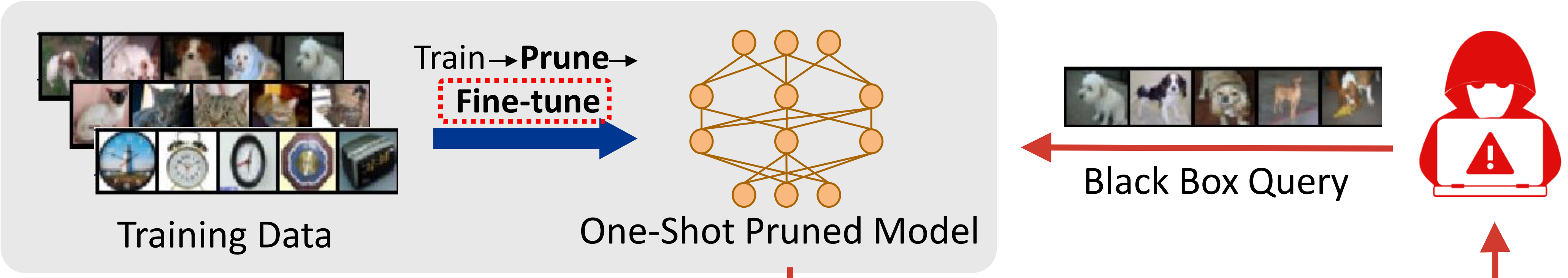


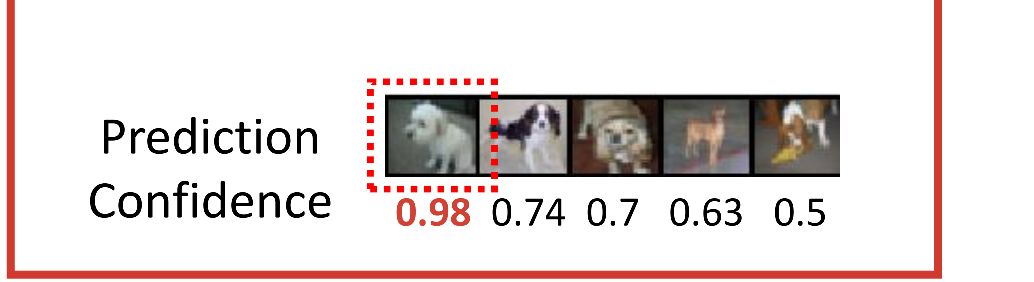
- Neural networks tend to **memorize training data** details excessively
- It is simple for models to differentiate between member and non-member samples

Was the data sample used for training?

Background: MIA in One-Shot Pruned Models

- **MIA is a typical privacy threat that leads to the leakage of sensitive training data**





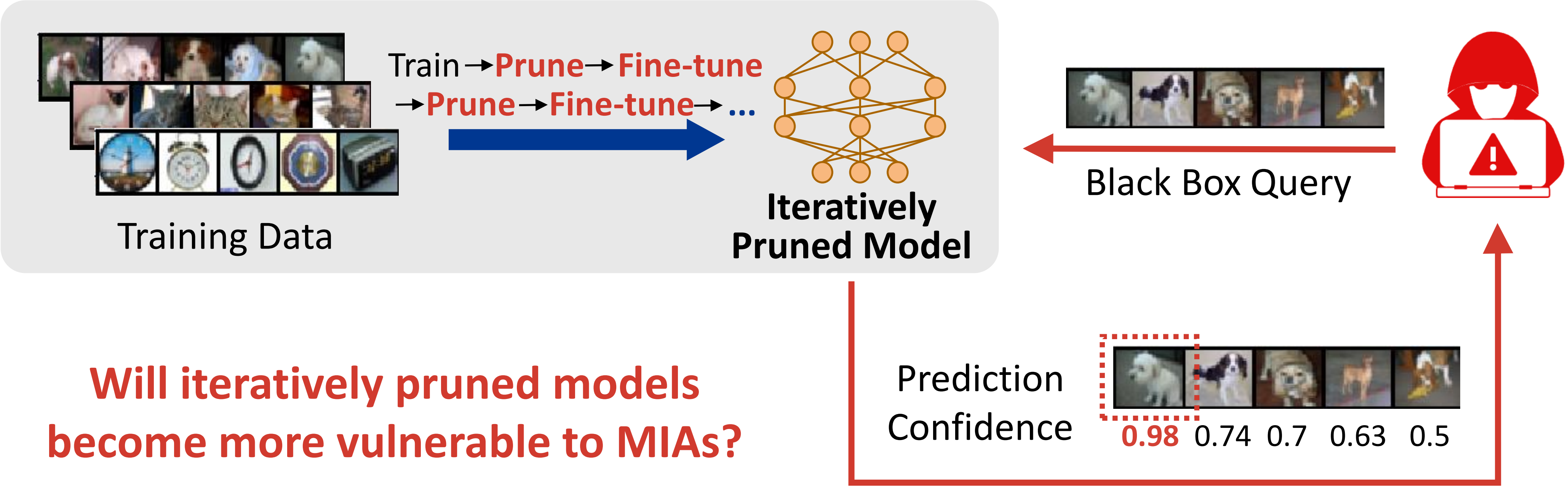
Was the data sample used for training?

- **Fine-tuning reuses training data** and **increases memorization** of training samples
- The attack accuracy of MIA in the one-shot pruned model is higher than in the original model

(Yuan et al., 2022)

Motivation

- MIA is a typical privacy threat that leads to the leakage of sensitive training data

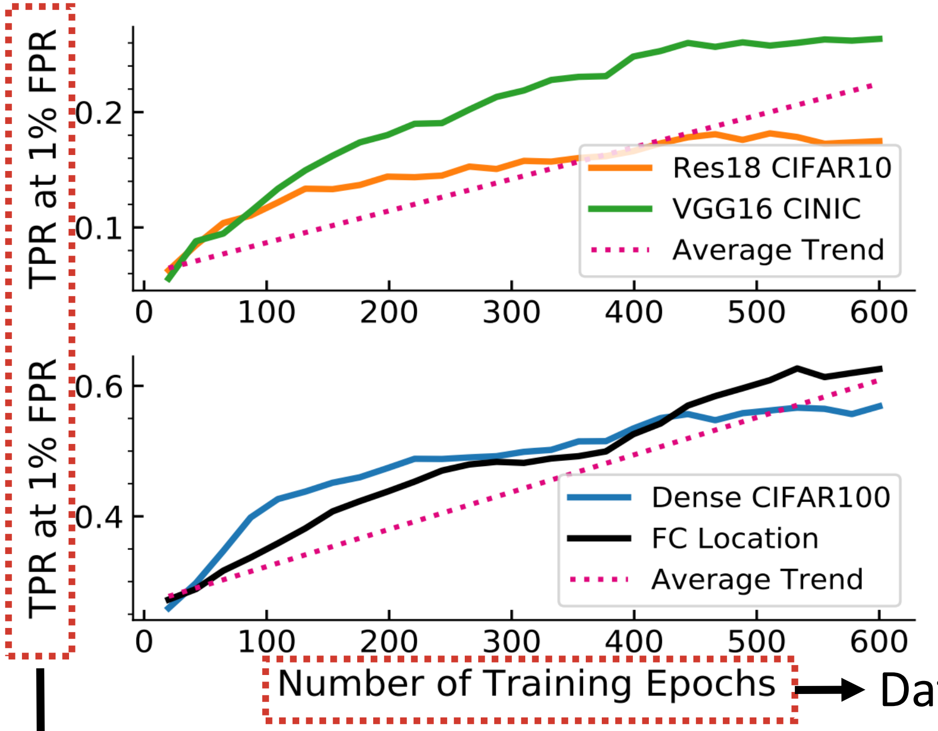


Will iteratively pruned models become more vulnerable to MIAs?

Was the data sample used for training?

Motivation

Reusing training data amplifies model memorization

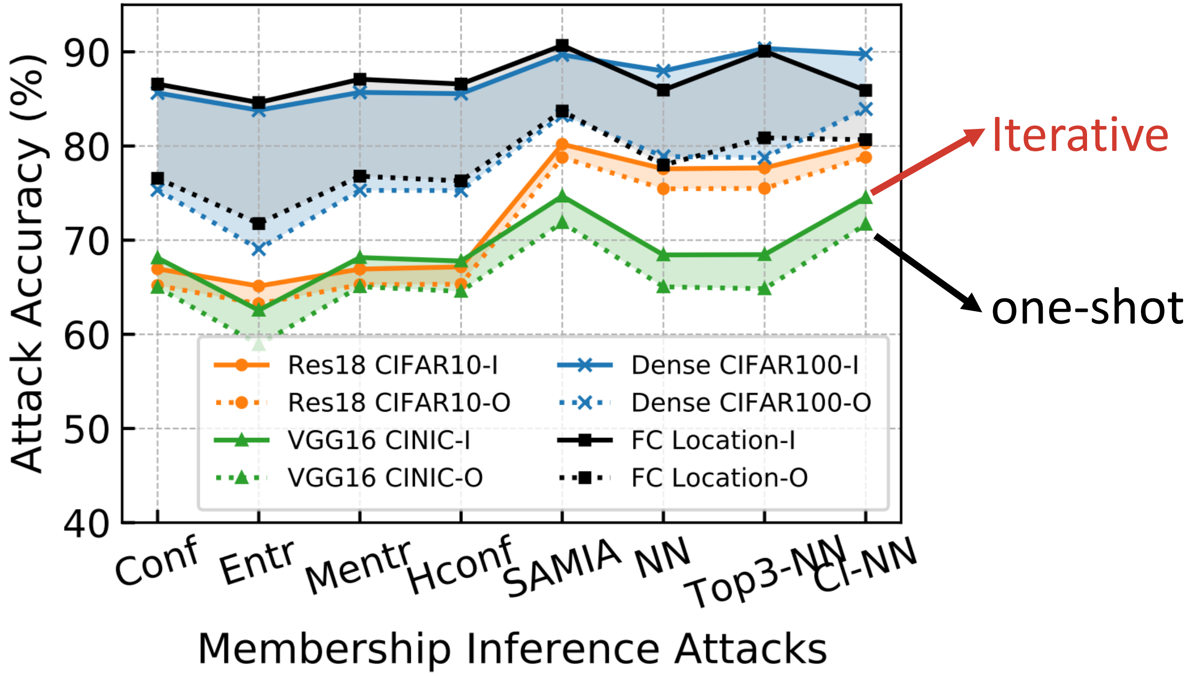


Data Reuse

The TPR of the recent MIA—*LiRA* is positively correlated with the model's memorization

(Carlini et al., 2022; Li et al, 2024)

MIA accuracy is higher in iteratively pruned models than in one-shot pruned models



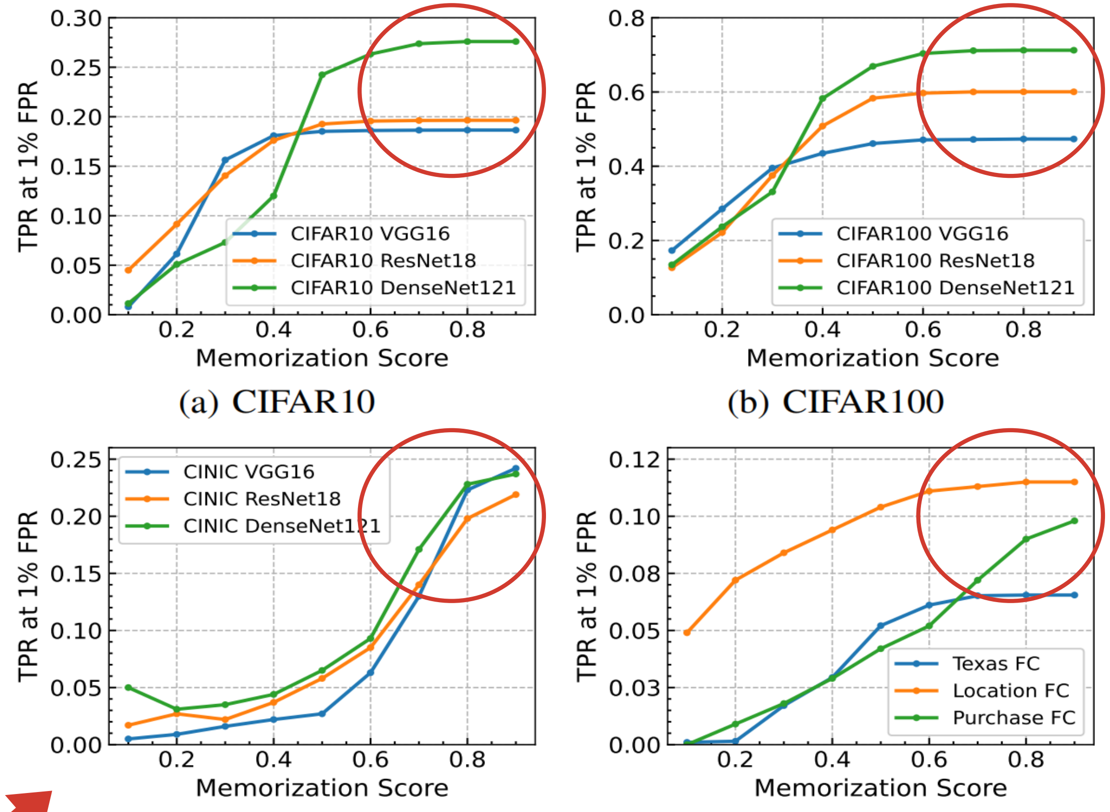
More data reuse enhances memorization and presents greater privacy risks to iteratively pruned models

Motivation

- **Memorization Score** (Feldman et al., 2020)
  - Measure the degree to which the model memorizes the training data

$$\mathrm{mem}(\mathcal{A}, D, (\boldsymbol{x}, y)) = \Pr_{f_{\boldsymbol{\theta}} \leftarrow \mathcal{A}(D)}[f_{\boldsymbol{\theta}}(\boldsymbol{x}) = y] - \Pr_{f_{\boldsymbol{\theta}} \leftarrow \mathcal{A}(D \setminus (\boldsymbol{x}, y))}[f_{\boldsymbol{\theta}}(\boldsymbol{x}) = y]$$

  - Models memorize training samples to varying degrees
  - **Data with higher memorization scores are more prone to be memorized**
- **Inherently easy-to-memorize training data is more vulnerable to serious privacy threats**



(a) CIFAR10 (b) CIFAR100

(c) CINIC (d) Tabular Datasets

Reuse easy-to-memorize data intensifies memorization and presents greater privacy risks to iteratively pruned models

Design Rationale

Two factors for increased memorization in iterative pruning:

- Reuse of training samples
- Inherently easy-to-memorize characteristics of some samples

Defend against MIAs in iteratively pruned models by weakening memorization

Scenario 1

- **Impact of data reuse:** using the entire training set in each epoch increases model memorization

Scenario 2

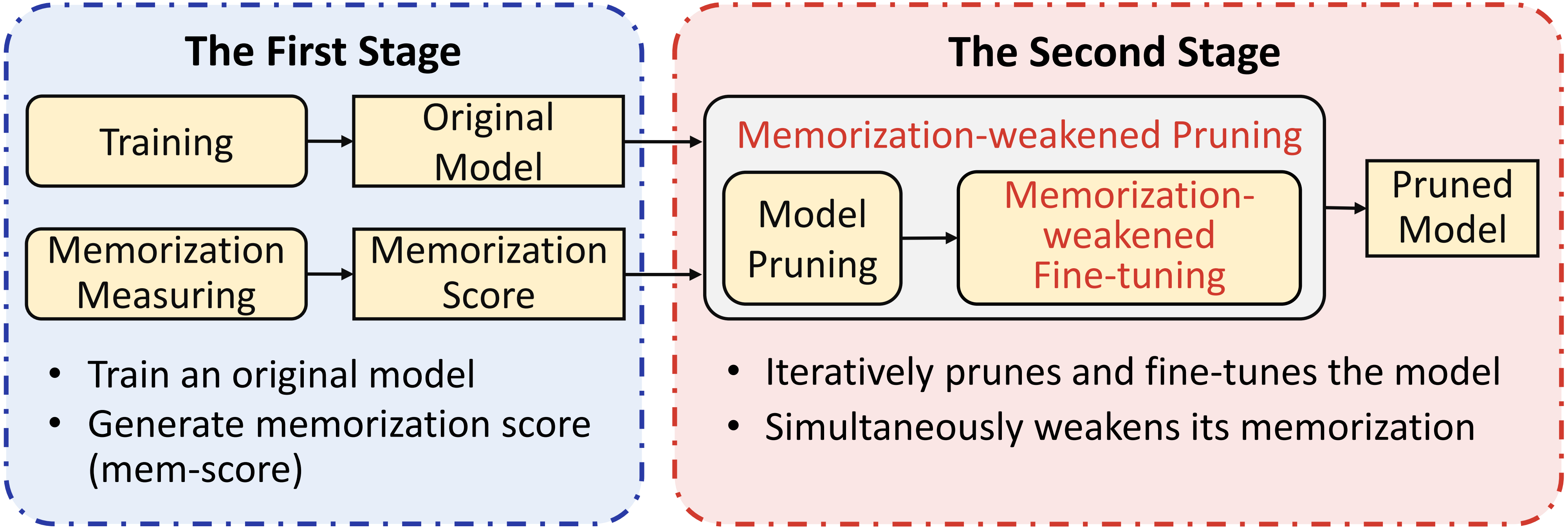
- **Impact of easy-to-memorize data:** the model retains stronger memory of easy-to-memorize samples

Scenario 3

- **Combined impact of data reuse and easy-to-memorize data:** reusing the entire dataset while retaining a deeper memory of easy-to-memorize data amplifies overall memorization and privacy risks

Our Defense: WeMem Framework

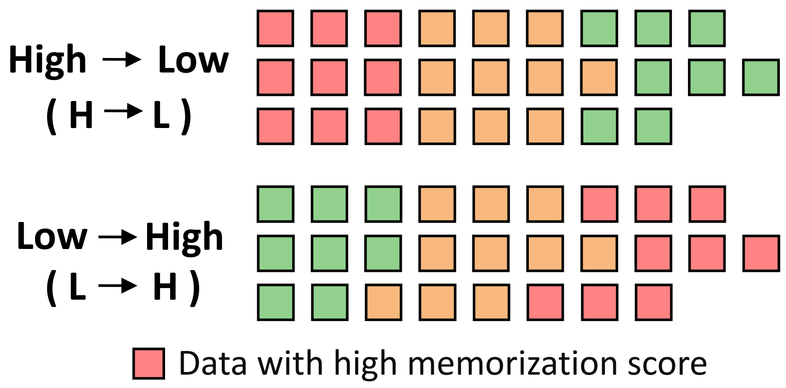
WeMem (Weaken Memorization) Defense Framework



Our Defense: Memorization-weakened Fine-tuning

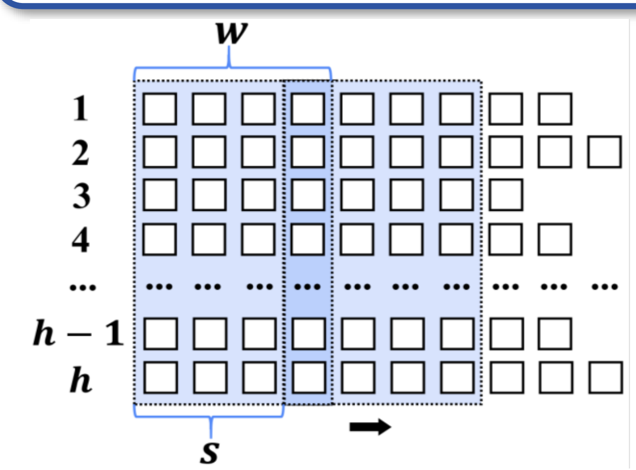
Three Memorization Weakening Primitives

Memorization-score-based Data Ranking



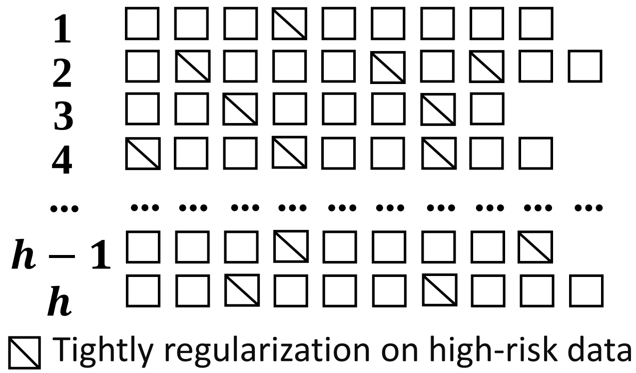
- High mem-score data samples are the primary target of WeMem's defenses
- Ranks data samples within each class based on their mem-scores
- The basic primitive used by all defense methods

Sliding-window-based Data Sampling



- Control the amount of training data in each epoch
- ***h*** --> Number of Data Classes
- ***w*** --> Window Width
- ***s*** --> Sliding Step Size
- Each sampling by a window provides data for one training epoch

Adaptive Regularization

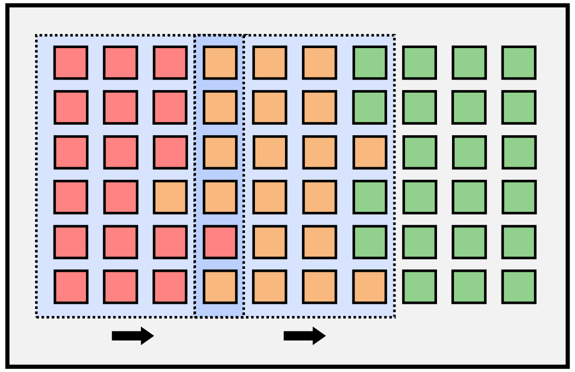


- Use L2 parameter regularization to constrain with different intensities on high- and low-risk data
- Adaptive to the privacy risk of the training data

Our Defense: Memorization-weakened Fine-tuning

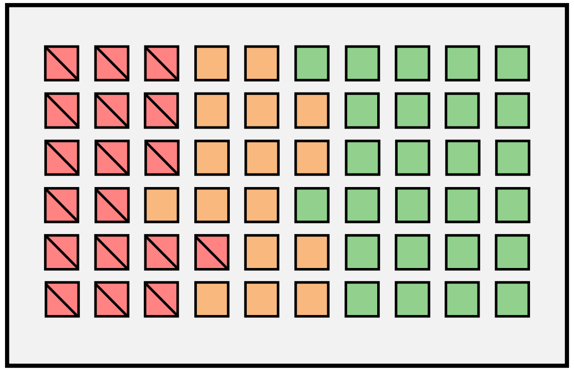
Memorization Weakening Methods for Three methods

Ranking-based Sliding Window (RSW)



- Adjust how training data is utilized **without modifying the training algorithm**
- Reduce data reuse and let high mem-score samples appear in fewer epochs

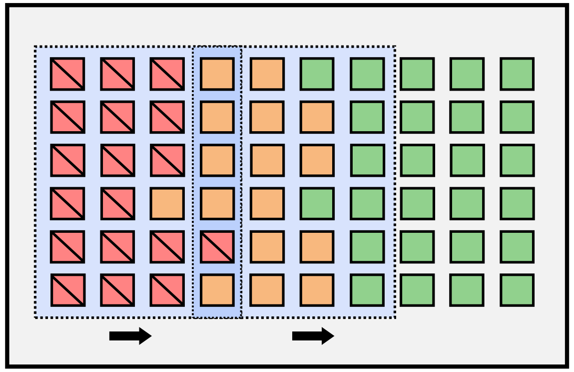
Risky Memory Regularization (RMR)



- Identify samples with risky memory (i.e., high mem-score) by a mem-score threshold τ
- Use L2 regularization to tightly constrain the model's learning capacity on them

$$\mathcal{L}_{\mathrm{RMR}} = \mathcal{L}(f_{\boldsymbol{\theta}}(\boldsymbol{x}), y) + \begin{cases} \frac{1}{2}\overset{\text{Large}}{\boxed{\lambda_r}}\|\boldsymbol{\theta}\|_2^2, & \mathrm{mem}(\boldsymbol{x}) \geq \tau \\ \frac{1}{2}\underset{\text{Small}}{\boxed{\lambda_g}}\|\boldsymbol{\theta}\|_2^2, & \mathrm{mem}(\boldsymbol{x}) < \tau \end{cases}$$

Sliding Window and Memory Regularization (SWMR)



- Combine the RSW and RMR to weaken memorization further
- In fine-tuning, the threshold τ identifies high-risk data in each window, and L2 regularization imposes a strict constraint on the training of these data

Evaluation: Setup

General Settings

- **6 Datasets**
  - CIFAR10, CIFAR100, CINIC, Texas, Location, Purchase
- **4 Deep Neural Networks**
  - Image datasets: ResNet18, VGG16, DefenseNet121
  - Tabular datasets: Fully Connected Neural Network
- **3 Pruning Rates (Proportion of Weights Removed)**
  - 50%, 60% (mainly used), 70%
- **10 Adaptive Membership Inference Attacks**
  - 4 metric-based attacks; 6 classifier-based attacks
- **5 Existing MIA Defenses**
  - Base (early stopping and L2), PPB (Yuan et al., 2022), ADV (Nasr et al., 2018), DPSGD (Abadi et al., 2016) , RelaxLoss (Chen et al., 2022)
- **3 Pruning Approaches with 5 Iterations**
  - L1 unstructured pruning; L1 structured pruning; L2 structured pruning

Our Defense Settings

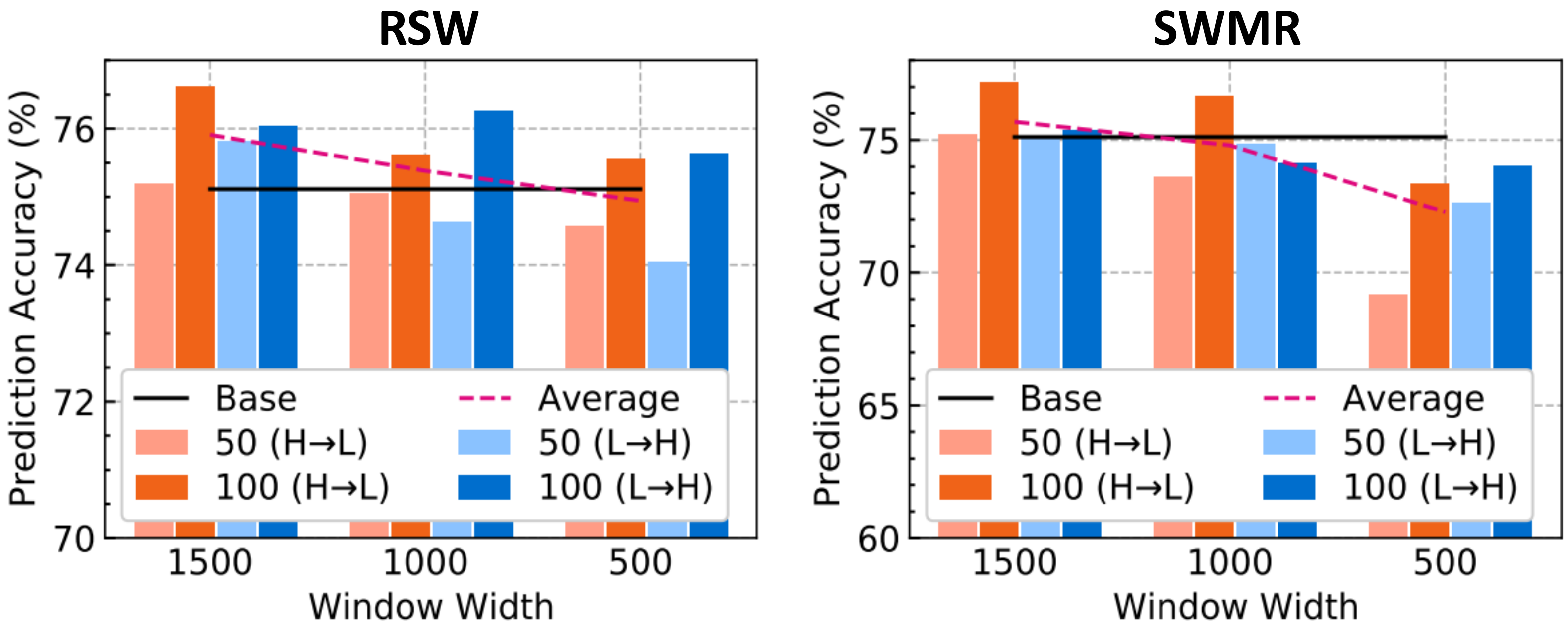
- **Sliding Windows and Mem-score Threshold Settings**

Data	Height (h)	Width (w)	Step Size (s)	Model	Threshold
CIFAR10	10	{1500, 1000, 500}	{50, 100}	All three DNNs	$\tau = 0.5$
CIFAR100	100	{150, 100, 50}	{5, 10}	All three DNNs	$\tau = 0.6$
CINIC	10	{2700, 1800, 900}	{100, 200}	ResNet18, VGG16 DenseNet121	$\tau = 0.7$ $\tau = 0.65$
Texas	100	{160, 110, 55}	{5, 10}	FC	$\tau = 0.6$
Location	30	{40, 30, 15}	{1, 3}	FC	$\tau = 0.6$
Purchase	100	{474, 316, 158}	{25, 35}	FC	$\tau = 0.75$

- **L2 Regularization Coefficients Settings**
  - $\lambda_g = 0.0005$
  - $\lambda_r \in \{0.01, 0.1, 1\}$

Evaluation: Key Results

Prediction accuracy of the pruned models using two data rankings
(CIFAR10, ResNet18)



- As window width decreases, model prediction accuracy declines
- SWMR's prediction accuracy is often lower than under RSW with identical settings

Evaluation: Key Results

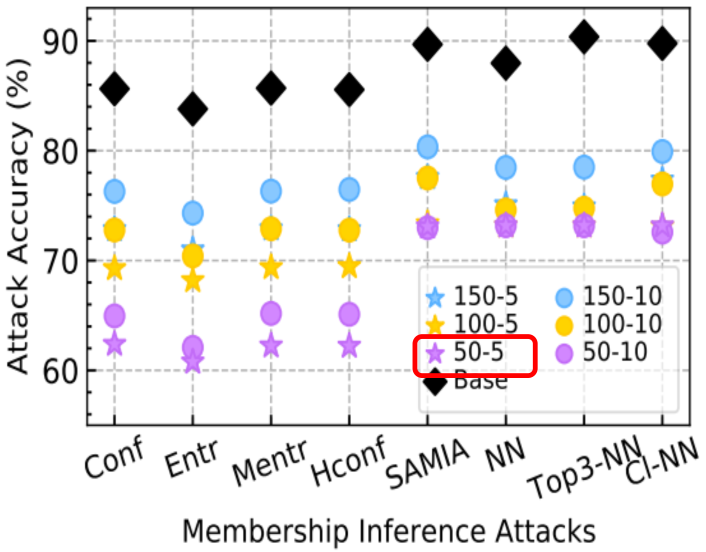
Under RMR defense with λg = 0.0005 and λr ∈ {0.01, 0.1, 1}, the test and attack accuracy on different pruned models

Data&Model	λ_r	Test Acc (%)	Adaptive Attack Accuracy (%)							
			Conf	Entr	Mentr	Hconf	SAMIA	NN	Top3-NN	Cl-NN
CIFAR10 DenseNet121	Base	80.01	63.91	62.05	63.96	64.33	78.10	75.85	76.08	78.44
	0.01	78.96	60.69	58.43	60.67	60.78	76.19	73.50	73.41	76.22
	0.1	**77.81**	**54.60**	**53.06**	**54.78**	**54.84**	**73.07**	**72.89**	**73.17**	**73.13**
	1	69.83	52.14	50.97	51.99	51.93	72.79	73.27	72.04	73.03
CIFAR100 ResNet18	Base	42.44	91.91	91.02	92.10	92.09	94.39	93.98	94.84	94.36
	0.01	41.03	90.03	88.68	90.18	90.24	93.17	92.78	93.02	93.58
	0.1	**37.46**	**60.12**	**54.69**	**60.07**	**59.93**	**73.30**	**73.29**	**72.45**	**72.91**
	1	10.13	50.88	50.07	50.88	51.21	71.32	72.05	71.67	72.37

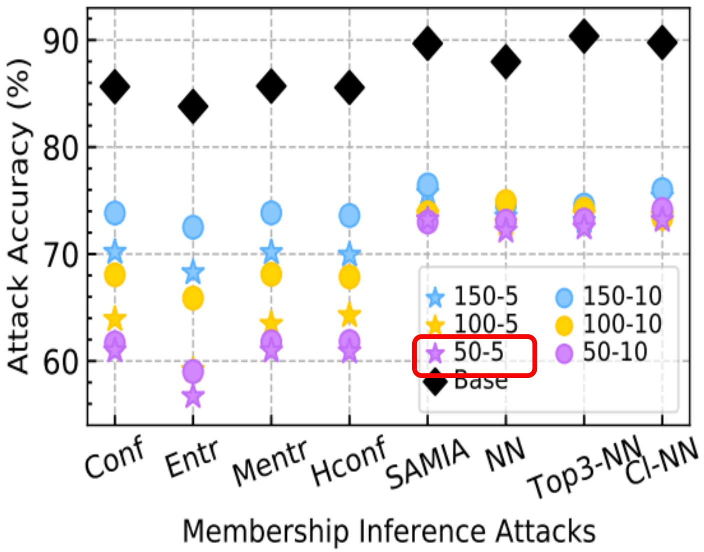
RMR achieves the best privacy-utility tradeoff when **λr = 0.1**

Evaluation: Key Results

Defense effectiveness of the pruned models using two data rankings (CIFAR100, DenseNet121)



RSW

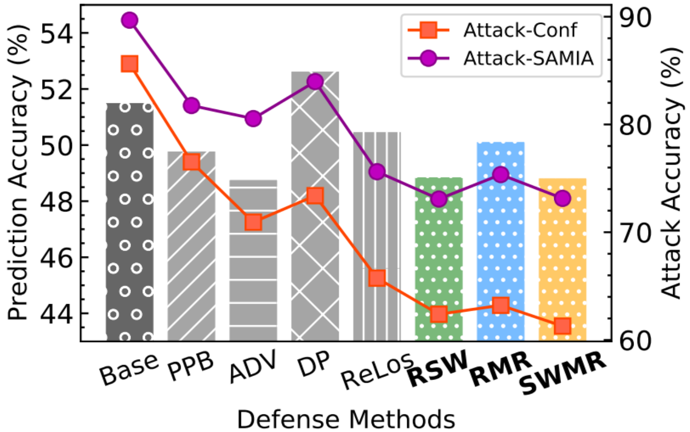


SWMR

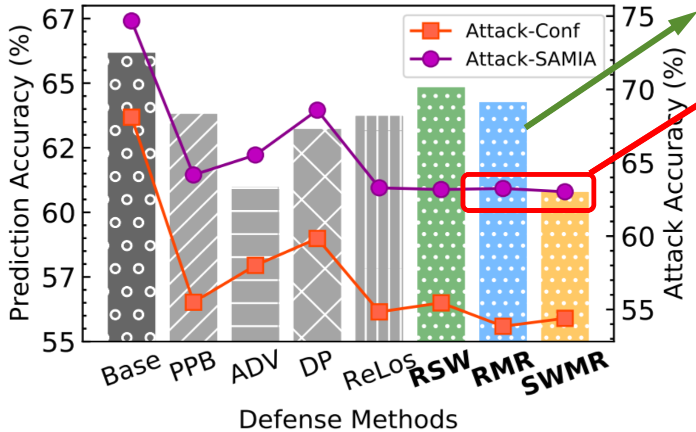
- A sliding window with a **small width and small step size** significantly weakens memorization, achieving the **best defense**
- **SWMR** provides better defense compared to RSW under identical settings

Evaluation: Key Results

Performance Comparison with Existing Defenses



CIFAR100, DenseNet121



CINIC, VGG16

Prediction Accuracy

Attack Accuracy

- **WeMem** achieves **high prediction accuracy** while **reducing attack accuracy more** than other defense methods

Time Cost Comparision in Iterative Pruning

Data&Model	Base	RSW	RMR	SWMR	PPB	ADV	RelaxLoss	DP
CIFAR10 VGG16	630s	**269s**	468s	332s	571s	275s	434s	7h
CIFAR100 ResNet18	458s	**174s**	611s	259s	643s	226s	532s	9h
CINIC DenseNet121	1616s	**463s**	2404s	1498s	1759s	495s	1696s	50h
Location FC	93s	**68s**	98s	95s	99s	195s	88s	231s

- **Sliding window** sampling reduces the amount of training data in each epoch, **speeding up the iterative fine-tuning** process

Summary

- **Data reuse** and the **easy-to-memorize characteristics** of some data are important factors that increase memorization during **iterative pruning**, leading to greater privacy risks
- Considered two factors' separate and combined impacts across **three scenarios** that make iteratively pruned models more vulnerable to MIAs
- Proposed **WeMem**, defending MIAs in iterative pruning by **weakening memorization**
- Designed **three defense primitives** and proposed **methods tailored to each scenario** that effectively weaken memorization
- WeMem provides effective defenses **against ten adaptive MIAs** and **outperforms five existing defenses** in terms of privacy-utility tradeoff and defense time cost

Thank you!

Jian wang

Beijing Jiaotong University

wangjian@bjtu.edu.cn

Source Code

北京交通大學 BEIJING JIAOTONG UNIVERSITY

北京工業大學 BEIJING UNIVERSITY OF TECHNOLOGY

Northeastern University