CompLeak: Deep Learning Model Compression Exacerbates Privacy Leakage

Na Li[*,1], Yansong Gao[*,2], Hongsheng Hu[3], Boyu Kuang[†,1], Anmin Fu[†,1]

1: Nanjing University of Science and Technology, China
2: Southeast University, China
3: Shanghai Jiao Tong University, China





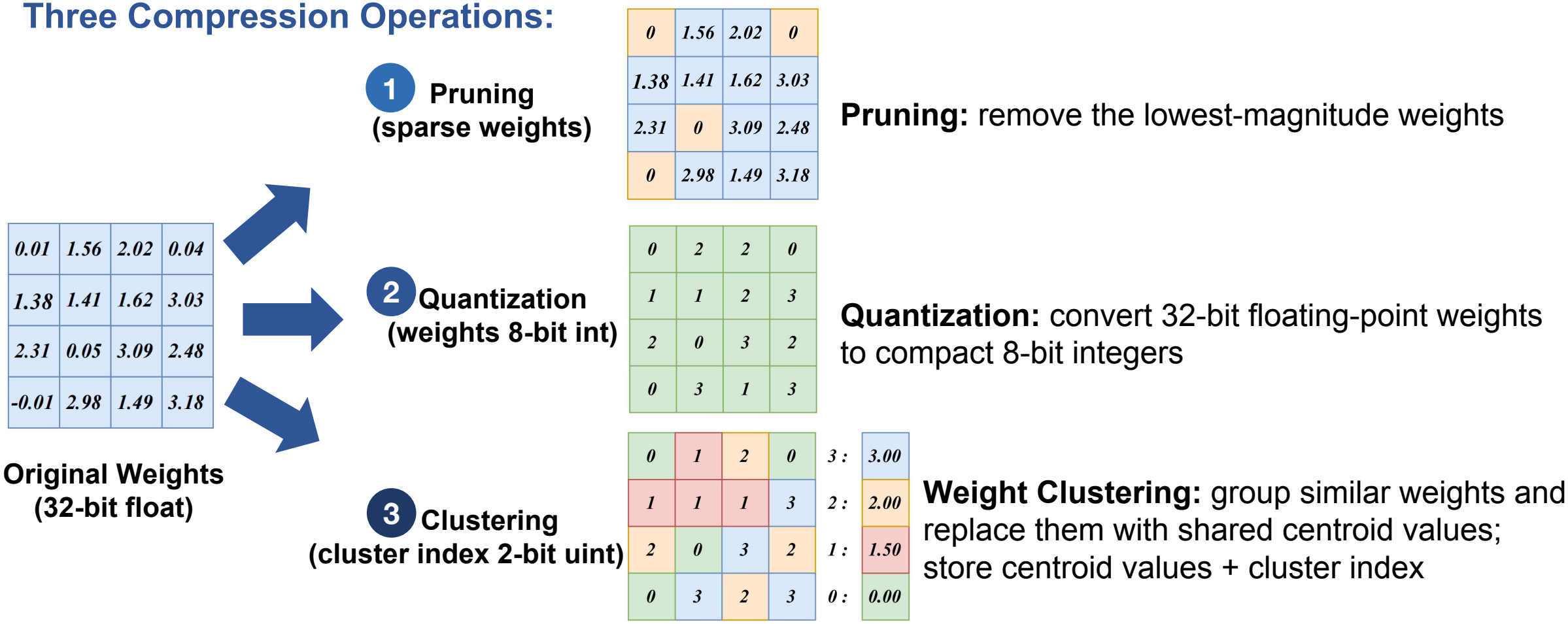


USENIX Security 2026

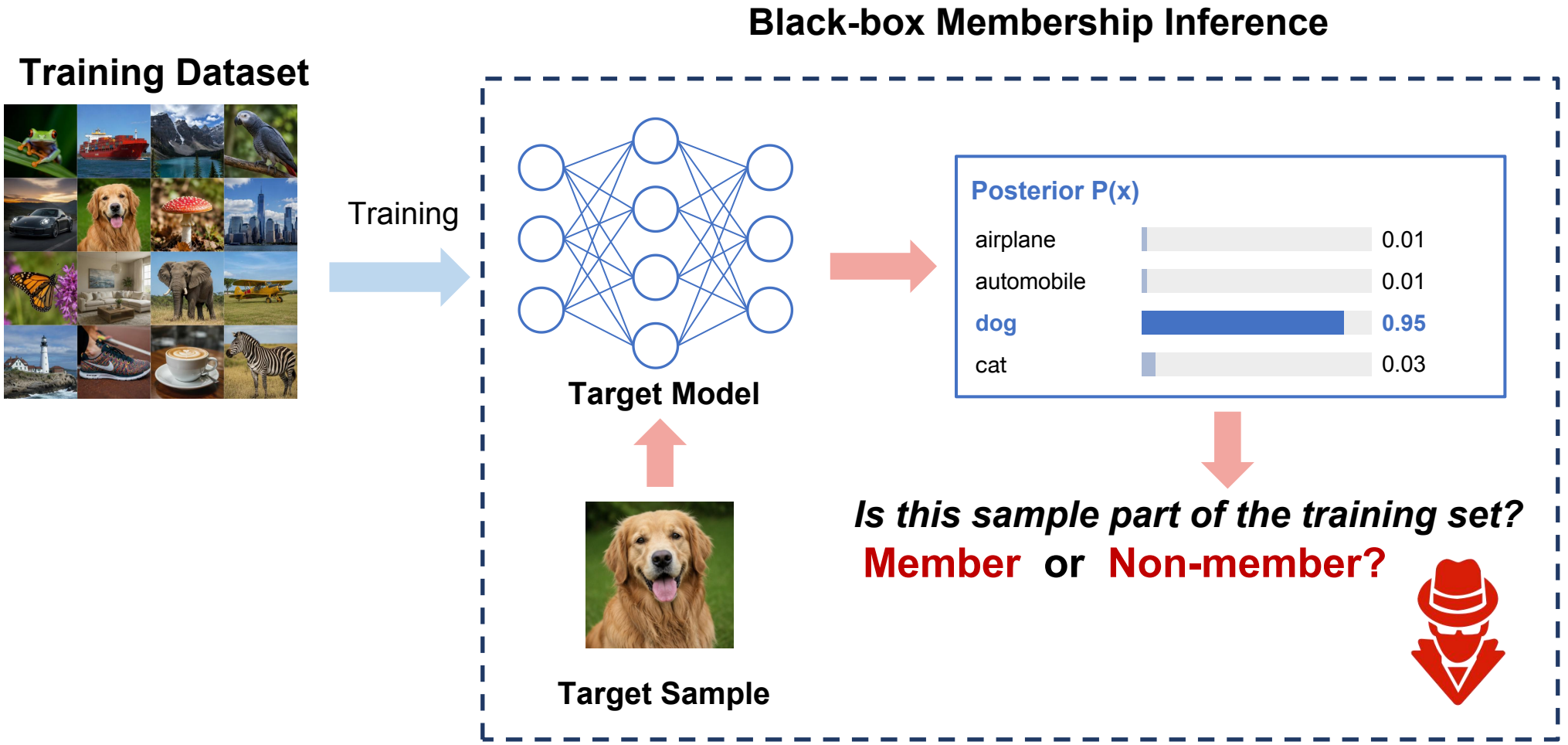
Model Compression

- The rapid growth of DL models makes ***deployment on resource-constrained devices challenging***
- Model compression mitigates this by ***reducing parameters, computation, and memory usage***

Three Compression Operations:



Membership Inference Attack (MIA)



The adversary wants to know **whether a specific sample was in the model's training set** and only observes the model output, i.e., the **full posterior probability vector**.

Privacy Risk of Compression Is Overlooked

- **Overlooked 1:** Many compression methods require fine-tuning on the original data, which may alter memorization and raise privacy risks.
- **Overlooked 2:** Multiple compressed variants of the same model are publicly available in practice, yet their combined privacy impact remains largely unexplored.

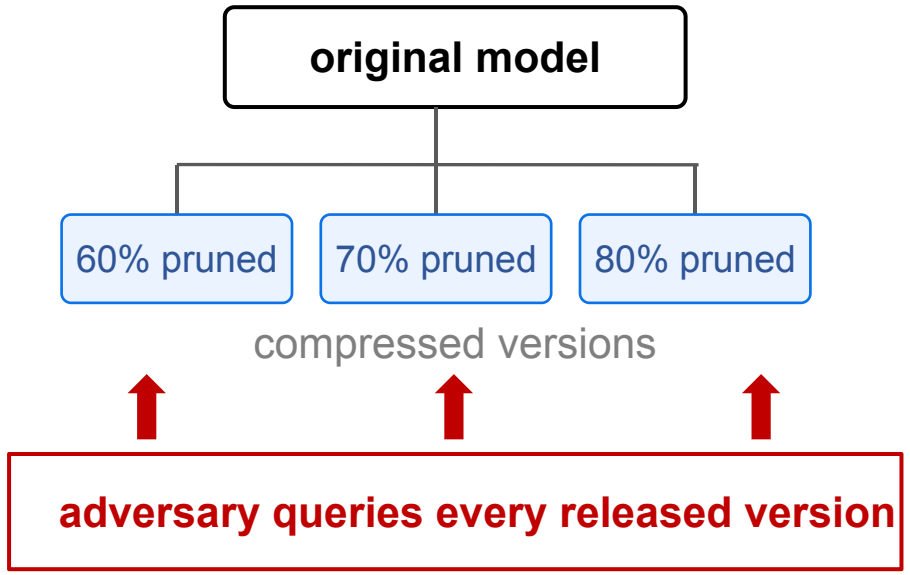
RESEARCH QUESTION



Does compression exacerbate privacy leakage as access to multiple compressed models grows — and if so, by how much?

Threat Model

Model providers release several compressed versions of one model, e.g., on HuggingFace.



Adversary Knowledge

- **Black-box access** — posterior outputs only, from the original model and its compressed versions
- Knows the **architecture** and the **compression configuration**, which providers publish

Adversary Capability

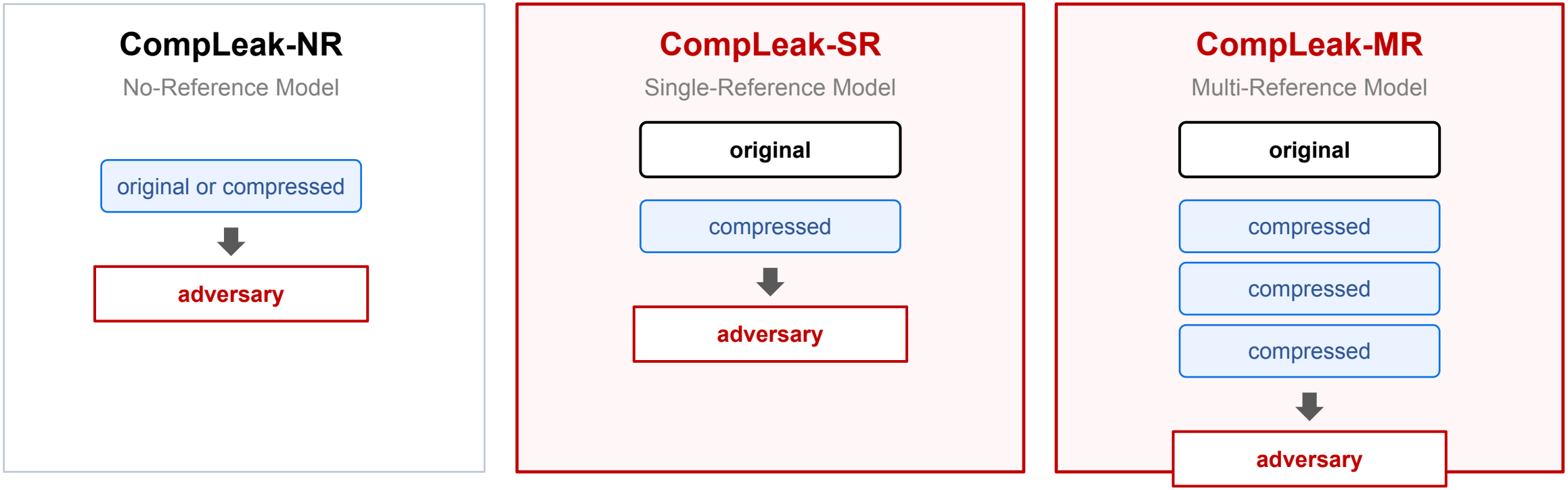
- Holds a **shadow dataset** from the victim's distribution, with no overlap sample
- Trains **shadow original and shadow compressed models** to mimic the victim's behavior

Adversary Goal

- Given a target sample, inference whether it was in the victim model's **training set**

Our Framework: CompLeak

- Through the lens of membership inference as a privacy auditing tool, we confirm: **model compression exacerbates privacy leakage**
- **Three variants**, by how many models the adversary can reach



Six existing MIAs on one model in isolation. **Baseline** per compressed version.

Pair original model with a compressed model **to capture compression-induced shifts**

Pair original model with multiple compressed models **to assess whether multiple versions further exacerbate the leakage**

CompLeak-NR

Six existing MIA methods audit the original model and per compressed version **at different compression degree**

Training-based MIA

[1] Reza Shokri, et al. "Membership Inference Attacks against Machine Learning Models." In *IEEE S&P*, 2017.

[2] Milad Nasr, Reza Shokri, and Amir Houmansadr. "Machine Learning with Membership Privacy using Adversarial Regularization." In *ACM CCS*, 2018.

[3] Xiaoyong Yuan and Lan Zhang. "Membership Inference Attacks and Defenses in Neural Network Pruning." In *USENIX Security*, 2022.

Metric-based MIA

[4] Samuel Yeom, et al. "Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting." In *IEEE CSF*, 2018.

[5] Liwei Song and Prateek Mittal. "Systematic Evaluation of Privacy Risks of Machine Learning Models." In *USENIX Security*, 2021.

State-of-the-art MIA

[6] Sajjad Zarifzadeh, Philippe Liu, and Reza Shokri. "Low-cost High-power Membership Inference Attacks." In *ICML*, 2024.



Purpose: quantify how much each compressed model (across compression methods and levels) leaks on its own — the baseline the other two variants build on.

CompLeak-NR (Experimental Results)

- **Datasets & Architectures**

Dataset	Architecture
CIFAR-10	ResNet18
CIFAR-100	ResNet50
Mini-ImageNet	VGG16
Tiny-ImageNet	MobileNetV2

- **Compression Settings**

Pruning: 60%/70%/80%/90% sparsity

Quantization: int-8

Clustering: 4/8/16 clusters

- **Metrics**

Balanced Accuracy

AUC

TPR @ 0.1% FPR

- **What We Observe**

Highly-compressed models are **HARDER** to attack

Why: Aggressive compression → Less capacity → Less overfitting

Example - Attack performance on varying pruned rate (VGG16+Mini-ImageNet)

Attack	TPR @ 0.1% FPR (%)					Balanced Accuracy (%)					AUC (%)				
Method	original	60%	70%	80%	90%	original	60%	70%	80%	90%	original	60%	70%	80%	90%
CompLeak$_{NR}$ [60] (RF)	1.6	1.6	1.5	1.4	1.1	59.3	59.4	59.1	58.6	57.8	63.2	63.1	62.7	62.1	60.7
CompLeak$_{NR}$ [51] (RF)	1.5	1.4	1.4	1.4	1.0	59.2	59.3	59.2	58.7	57.5	63.3	63.3	62.8	62.3	60.5
CompLeak$_{NR}$ [73]	0.1	0.0	0.0	0.0	0.0	56.7	57.0	56.2	54.2	52.9	54.8	54.9	54.2	52.9	50.4
CompLeak$_{NR}$ [63]	0.1	0.1	0.1	0.1	0.1	59.6	59.8	59.1	58.7	57.3	58.7	58.8	58.2	57.1	54.7
CompLeak$_{NR}$ [74]	0.9	0.7	0.7	0.8	0.7	61.9	61.7	61.4	60.5	59.7	66.0	66.4	64.7	64.8	63.3
CompLeak$_{NR}$ (RMIA) [75]	**3.4**	3.4	3.1	2.6	2.3	**66.9**	67.4	66.8	66.3	64.6	**73.7**	73.8	73.3	72.5	70.1

Example - Attack performance on varying number of clusters (VGG16+Mini-ImageNet)

Attack	TPR @ 0.1% FPR (%)				Balanced Accuracy (%)				AUC (%)			
Method	original	16	8	4	original	16	8	4	original	16	8	4
CompLeak$_{NR}$ [60] (RF)	1.6	1.4	1.2	1.4	59.3	59.1	58.5	57.9	63.2	63.2	62.4	61.7
CompLeak$_{NR}$ [51] (RF)	1.5	1.3	1.2	1.3	59.2	58.6	57.9	57.4	63.3	62.8	62.2	61.6
CompLeak$_{NR}$ [73]	0.1	0.1	0.1	0.0	56.7	56.5	56.0	54.4	54.8	54.8	54.7	53.7
CompLeak$_{NR}$ [63]	0.1	0.1	0.1	0.1	59.6	60.1	59.6	59.3	58.7	58.9	58.6	58.3
CompLeak$_{NR}$ [74]	0.9	1.3	0.6	0.7	61.9	61.7	61.5	61.2	66.0	64.3	64.1	62.9
CompLeak$_{NR}$ (RMIA) [75]	**3.4**	3.2	3.1	3.1	**66.9**	66.0	65.5	65.4	**73.7**	72.5	72.6	71.5

CompLeak-SR

- **Key Insight: Compression Shifts Members ≠ Non-Members**
  - Compression changes the posterior distribution **very differently for members and non-members** (Figure 1). e.g., on a 40%-pruned MobileNetV2:
    Members shift more — lost capacity cannot preserve their fine-grained features
    Non-members barely shift — generalization is stable
  - So we pair one compressed model with the original as a **reference** to capture that change.

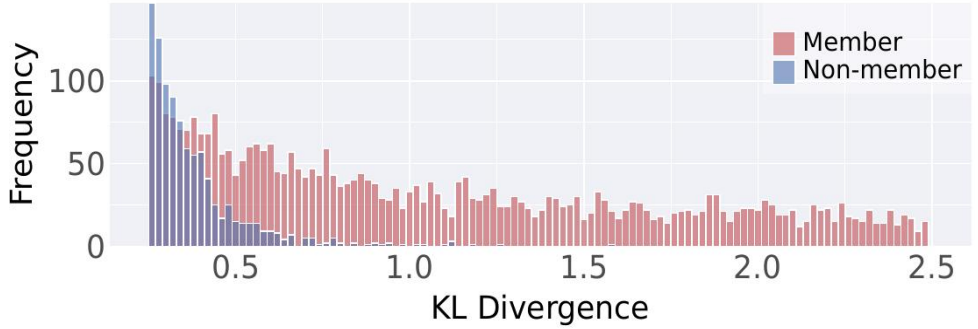


Figure 1: KL divergence between the original and 40%-pruned MobileNetV2 posteriors on the same samples.

- **Attack Steps**
  1. **Shadow Model Training**
  2. **Posterior Generation**
  3. **Meta-data Construction**
  4. **Attack Meta-classifier Training**
  5. **Membership Inference**

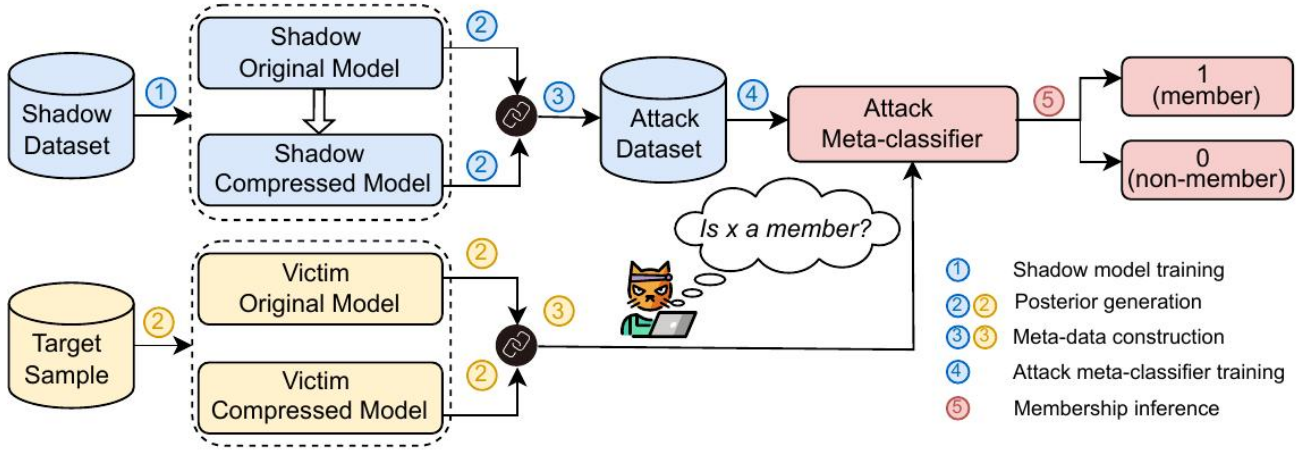


Figure 2: attack pipeline of CompLeak-SR.

CompLeak-SR (Experimental Results)

Table 2: Attack performance of different attacks on varying pruned rate (VGG16+Mini-ImageNet).

Attack	TPR @ 0.1% FPR (%)					Balanced Accuracy (%)					AUC (%)				
Method	original	60%	70%	80%	90%	original	60%	70%	80%	90%	original	60%	70%	80%	90%
CompLeak$_{NR}$ [60] (LR)	0.0	0.0	0.0	0.0	0.0	48.3	48.9	47.1	48.5	50.0	47.3	47.9	45.8	48.0	48.0
CompLeak$_{NR}$ [51] (LR)	0.0	0.0	0.0	0.0	0.0	48.3	50.8	49.6	50.0	51.6	50.1	50.7	50.2	51.0	51.4
CompLeak$_{NR}$ [60] (RF)	1.6	1.6	1.5	1.4	1.1	59.3	59.4	59.1	58.6	57.8	63.2	63.1	62.7	62.1	60.7
CompLeak$_{NR}$ [51] (RF)	1.5	1.4	1.4	1.4	1.0	59.2	59.3	59.2	58.7	57.5	63.3	63.3	62.8	62.3	60.5
CompLeak$_{NR}$ [73]	0.1	0.0	0.0	0.0	0.0	56.7	57.0	56.2	54.2	52.9	54.8	54.9	54.2	52.9	50.4
CompLeak$_{NR}$ [63]	0.1	0.1	0.1	0.1	0.1	59.6	59.8	59.1	58.7	57.3	58.7	58.8	58.2	57.1	54.7
CompLeak$_{NR}$ [74]	0.9	0.7	0.7	0.8	0.7	61.9	61.7	61.4	60.5	59.7	66.0	66.4	64.7	64.8	63.3
CompLeak$_{NR}$ (RMIA) [75]	**3.4**	3.4	3.1	2.6	2.3	**66.9**	67.4	66.8	66.3	64.6	**73.7**	73.8	73.3	72.5	70.1
CompLeak$_{SR}$ 1 (LR)	-	11.3	11.1	10.6	8.7	-	59.5	59.7	59.3	59.4	-	67.1	67.6	66.5	66.8
CompLeak$_{SR}$ 2 (LR)	-	8.8	8.6	8.5	8.0	-	59.7	59.9	59.9	60.3	-	68.4	68.5	68.8	69.4
CompLeak$_{SR}$ 1 (RF)	-	**39.0**	**42.1**	**34.5**	**25.5**	-	**84.1**	**84.2**	**83.4**	79.8	-	93.0	92.8	91.8	88.0
CompLeak$_{SR}$ 2 (RF)	-	36.4	41.7	34.2	24.2	-	83.9	83.9	83.2	**79.9**	-	**93.2**	**93.0**	**92.1**	**88.6**

Table 3: Attack performance of different attacks on original and quantized model (VGG16+Mini-ImageNet).

Attack	TPR @ 0.1% FPR (%)		Balanced Accuracy (%)		AUC (%)	
Method	original	int-8	original	int-8	original	int-8
CompLeak$_{NR}$ [60] (LR)	0.0	0.0	48.3	48.3	47.3	47.4
CompLeak$_{NR}$ [51] (LR)	0.0	0.0	48.3	51.1	50.1	50.5
CompLeak$_{NR}$ [60] (RF)	1.6	1.3	59.3	59.4	63.2	63.4
CompLeak$_{NR}$ [51] (RF)	1.5	1.2	59.2	59.3	63.3	63.5
CompLeak$_{NR}$ [73]	0.1	0.1	56.7	58.2	54.8	55.3
CompLeak$_{NR}$ [63]	0.1	0.1	59.6	59.6	58.7	59.1
CompLeak$_{NR}$ [74]	0.9	0.5	61.9	61.1	66.0	64.5
CompLeak$_{NR}$ (RMIA) [75]	**3.4**	3.2	**66.9**	66.4	**73.7**	72.9
CompLeak$_{SR}$ 1 (LR)	-	23.1	-	60.8	-	71.1
CompLeak$_{SR}$ 2 (LR)	-	10.0	-	59.6	-	70.8
CompLeak$_{SR}$ 1 (RF)	-	**81.0**	-	**91.1**	-	**98.3**
CompLeak$_{SR}$ 2 (RF)	-	80.7	-	90.3	-	**98.3**

Table 4: Attack performance of different attacks on varying number of clusters (VGG16+Mini-ImageNet).

Attack	TPR @ 0.1% FPR (%)				Balanced Accuracy (%)				AUC (%)			
Method	original	16	8	4	original	16	8	4	original	16	8	4
CompLeak$_{NR}$ [60] (LR)	0.0	0.0	0.0	0.0	48.3	49.3	47.4	51.2	47.3	47.8	48.0	48.7
CompLeak$_{NR}$ [51] (LR)	0.0	0.0	0.0	0.0	48.3	50.6	49.7	50.8	50.1	50.8	50.8	50.4
CompLeak$_{NR}$ [60] (RF)	1.6	1.4	1.2	1.4	59.3	59.1	58.5	57.9	63.2	63.2	62.4	61.7
CompLeak$_{NR}$ [51] (RF)	1.5	1.3	1.2	1.3	59.2	58.6	57.9	57.4	63.3	62.8	62.2	61.6
CompLeak$_{NR}$ [73]	0.1	0.1	0.1	0.0	56.7	56.5	56.0	54.4	54.8	54.8	54.7	53.7
CompLeak$_{NR}$ [63]	0.1	0.1	0.1	0.1	59.6	60.1	59.6	59.3	58.7	58.9	58.6	58.3
CompLeak$_{NR}$ [74]	0.9	1.3	0.6	0.7	61.9	61.7	61.5	61.2	66.0	64.3	64.1	62.9
CompLeak$_{NR}$ (RMIA) [75]	**3.4**	3.2	3.1	3.1	**66.9**	66.0	65.5	65.4	**73.7**	72.5	72.6	71.5
CompLeak$_{SR}$ 1 (LR)	-	16.4	11.3	8.1	-	60.4	60.1	59.3	-	69.6	68.3	65.0
CompLeak$_{SR}$ 2 (LR)	-	10.0	8.1	6.2	-	59.9	59.9	60.2	-	70.8	68.4	67.3
CompLeak$_{SR}$ 1 (RF)	-	**67.2**	**47.3**	**28.5**	-	**93.2**	89.7	84.1	-	**98.7**	96.8	92.4
CompLeak$_{SR}$ 2 (RF)	-	66.3	46.8	**28.5**	-	**93.2**	**90.1**	**84.5**	-	**98.7**	**97.0**	**92.8**

Regardless of the compression type and level, CompLeak-SR **consistently outperforms** CompLeak-NR, indicating that **compression indeed leads to additional privacy leakage**.

CompLeak-MR

- **Design Rationale: Different Versions Leak Differently**
  - ➢ We find **two membership signals**:
    - **CompLeak-SR Meta-classifiers' Posterior:** compression levels → different SR meta-classifiers → divergent inferences on the same sample
    - **Loss** — member loss climbs with sparsity, non-member loss fluctuates

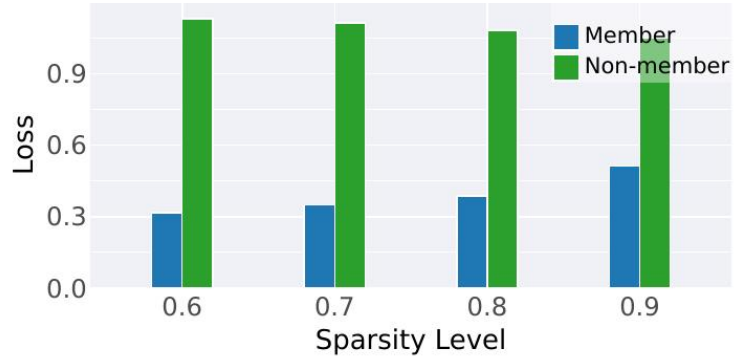


Figure 3: average loss of members vs non-members across sparsity levels.

  - ➢ So we **stack both signals across every released version** to accumulate the leakage a single pair cannot reveal.

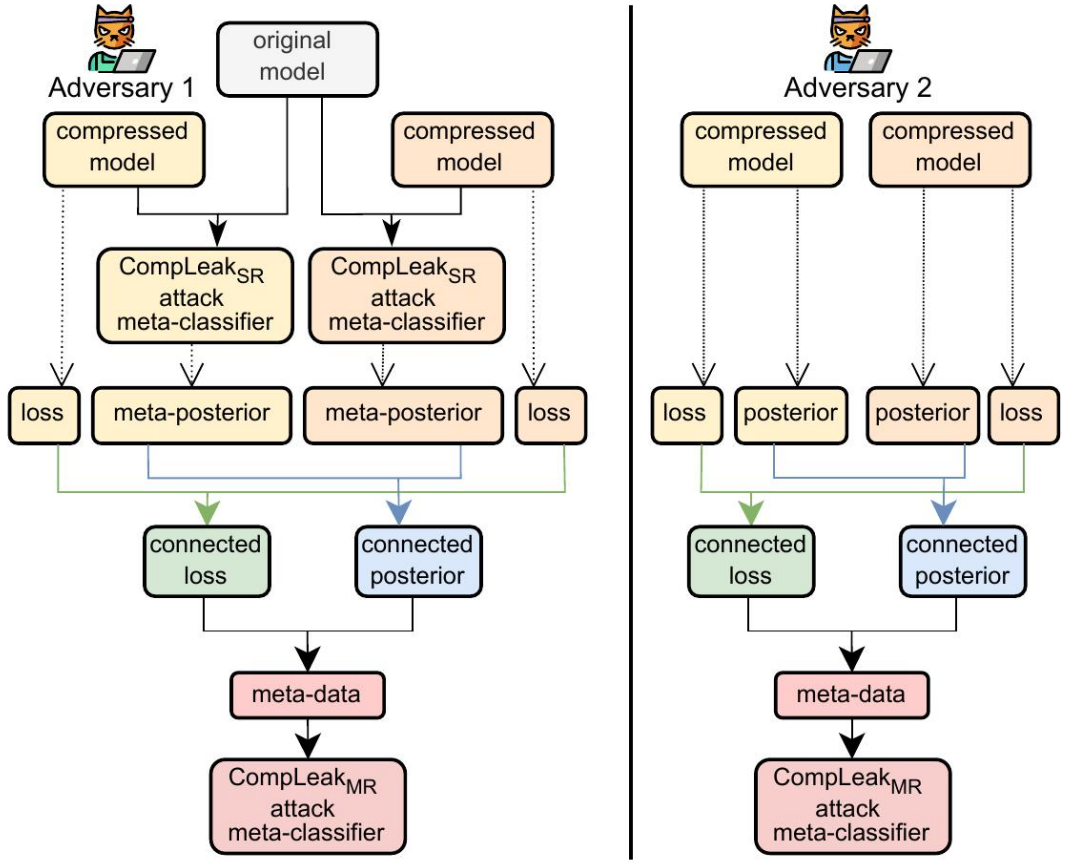


Figure 4: general attack pipeline of CompLeak-MR, using two compressed models as an example.

- **Attack Steps**

1. Shadow Model Training → 2. Posterior Concatenation → 3. Loss Concatenation → 4. Attack Meta-classifier Training → 5. Membership Inference

CompLeak-MR (Experimental Results)

Table 6: Average performance of $\texttt{CompLeak}_{\text{MR}}$ on Mini-ImageNet under Adversary 1 (five repetitions).

Operation	TPR @ 0.1% FPR (%)	Balanced Accuracy (%)	AUC (%)
Pruning	61.3 (19.2% ↑)	94.8 (10.6% ↑)	98.9 (5.7% ↑)
Clustering	72.7 (5.5% ↑)	95.1 (1.9% ↑)	99.0 (0.3% ↑)
Three Operations	95.7 (14.7% ↑)	98.8 (5.6% ↑)	99.9 (1.2% ↑)

In parentheses is the improvement over the optimal result when $\texttt{CompLeak}_{\text{SR}}$ attacking a single compressed model in the respective compression scenario.

Four pruned models with L = {0.6, 0.7, 0.8, 0.9}.
Three clustered models with N = {4, 8, 16}.

Table 7: Average performance of $\texttt{CompLeak}_{\text{MR}}$ on Mini-ImageNet under Adversary 2 (five repetitions).

Operation	TPR @ 0.1% FPR (%)	Balanced Accuracy (%)	AUC (%)
Pruning	2.6 (1.8% ↑)	71.3 (9.6% ↑)	77.6 (11.2% ↑)
Clustering	1.1 (0.2% ↓)	65.6 (3.9% ↑)	70.6 (6.3% ↑)
Three Operations	8.8 (7.5% ↑)	71.4 (9.7% ↑)	80.6 (14.2% ↑)

In parentheses is the improvement over the optimal result when SAMIA [74] attacking a single compressed model in the respective compression scenario.

From the CompLeak-MR results, **multiple compressed models further amplify privacy leakage**, regardless of whether the original model is available.

Summary

- We propose CompLeak, the first systematic privacy-risk evaluation framework examining pruning, quantization and weight clustering through the lens of membership inference.
- CompLeak-NR — employs existing MIA, relying solely on information from the underlying model, to comprehensively assess the privacy leakage per compressed model
- CompLeak-SR — for the single-compression scenario, unveiling that regardless of the compression degree, the compression operation indeed jeopardizes privacy
- CompLeak-MR — aggregates leaks from multiple compressed models, further amplifying the privacy leakage caused by compression

Thank you!

Q & A