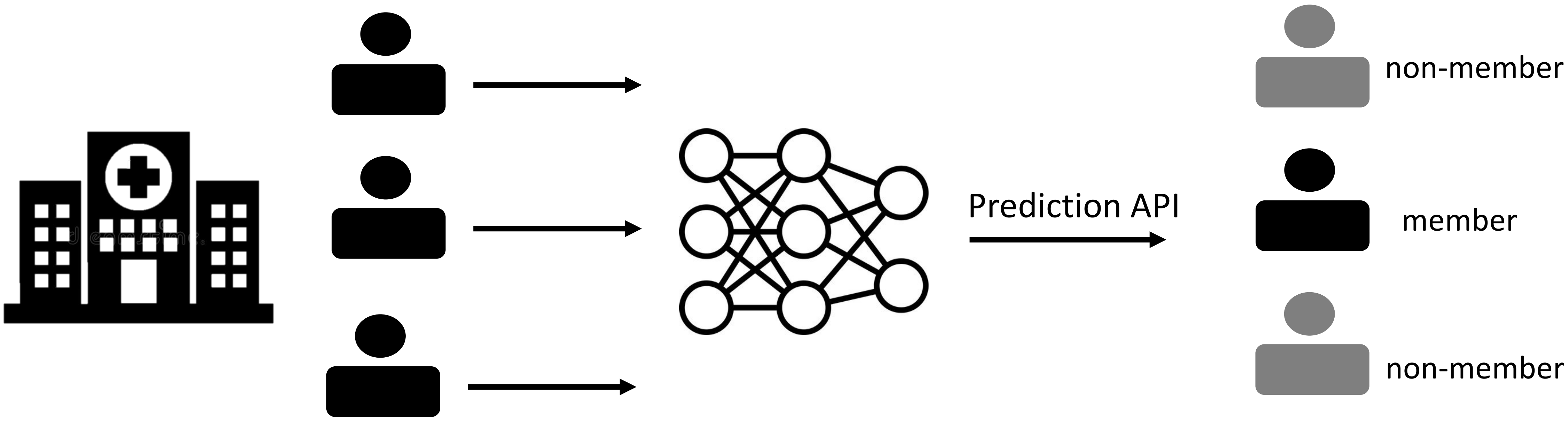
Overconfidence is a Dangerous Thing:
Mitigating Membership Inference Attacks by Enforcing Less Confident Prediction

Zitao Chen, Karthik Pattabiraman

THE UNIVERSITY
OF BRITISH COLUMBIA

Artifact Evaluated
NDSS SYMPOSIUM
Available
Functional
Reproduced

Membership Inference Attacks (MIAs)



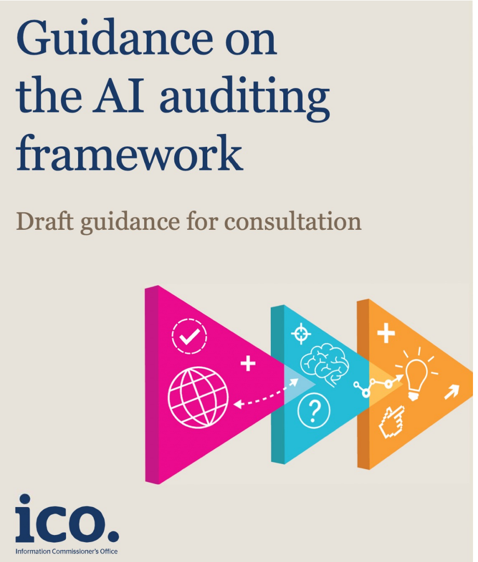
Does the sensitive training set contain a target record?

MIAs as a privacy threat

MIAs as a privacy threat

Confidentiality violation

MIAs as a privacy threat



Guidance on the AI auditing framework

Draft guidance for consultation

ico.

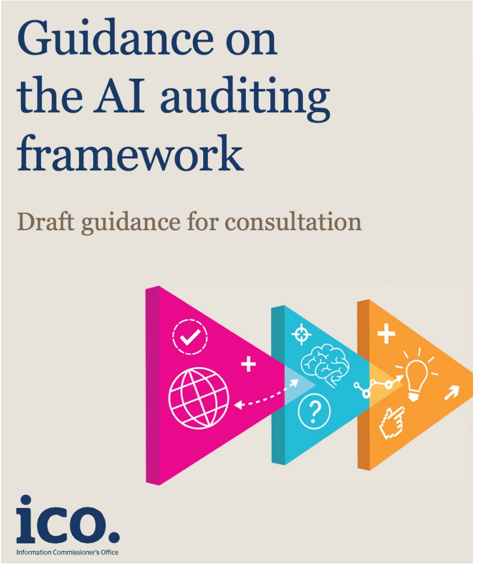
Information Commissioner's Office

Confidentiality violation

Auditing purpose

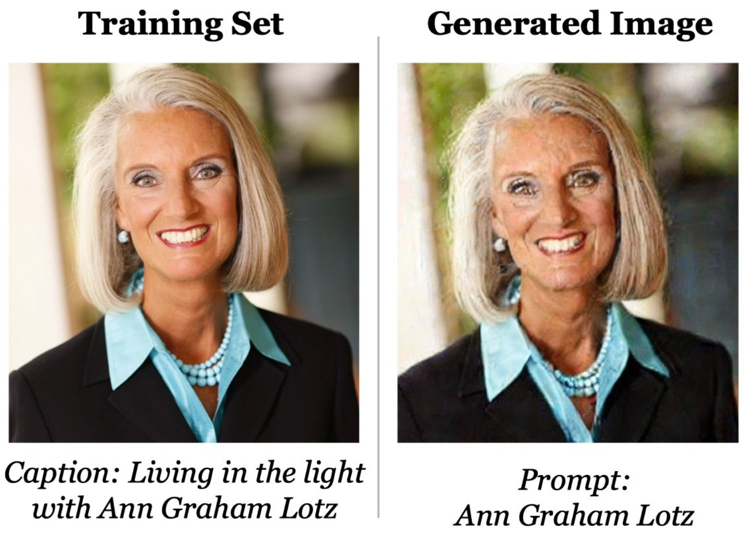
https://ico.org.uk/media/about-the-ico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf

MIAs as a privacy threat


Guidance on the AI auditing framework

Draft guidance for consultation

ico.



Confidentiality violation

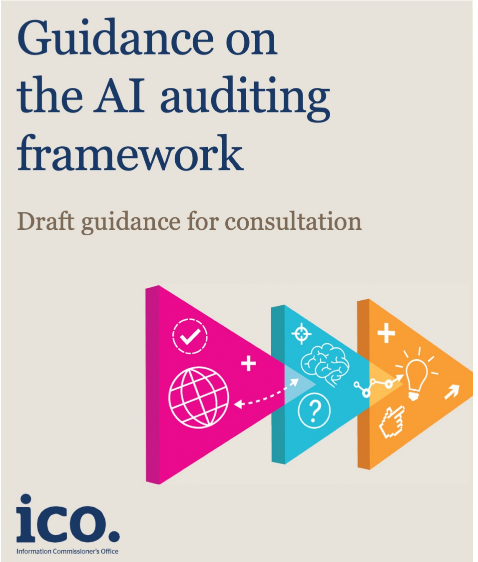
Auditing purpose

Stepping stone for more powerful attack

https://ico.org.uk/media/about-the-ico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf

Carlini et al., Extracting Training Data from Diffusion Models, USENIX’23

MIAs as a privacy threat



Training Set | **Generated Image**

Caption: Living in the light with Ann Graham Lotz

Prompt: Ann Graham Lotz

Confidentiality violation

Auditing purpose

Stepping stone for more powerful attack

We need effective defense against MIAs!

https://ico.org.uk/media/about-the-ico/consultations/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf

Carlini et al., Extracting Training Data from Diffusion Models, USENIX’23

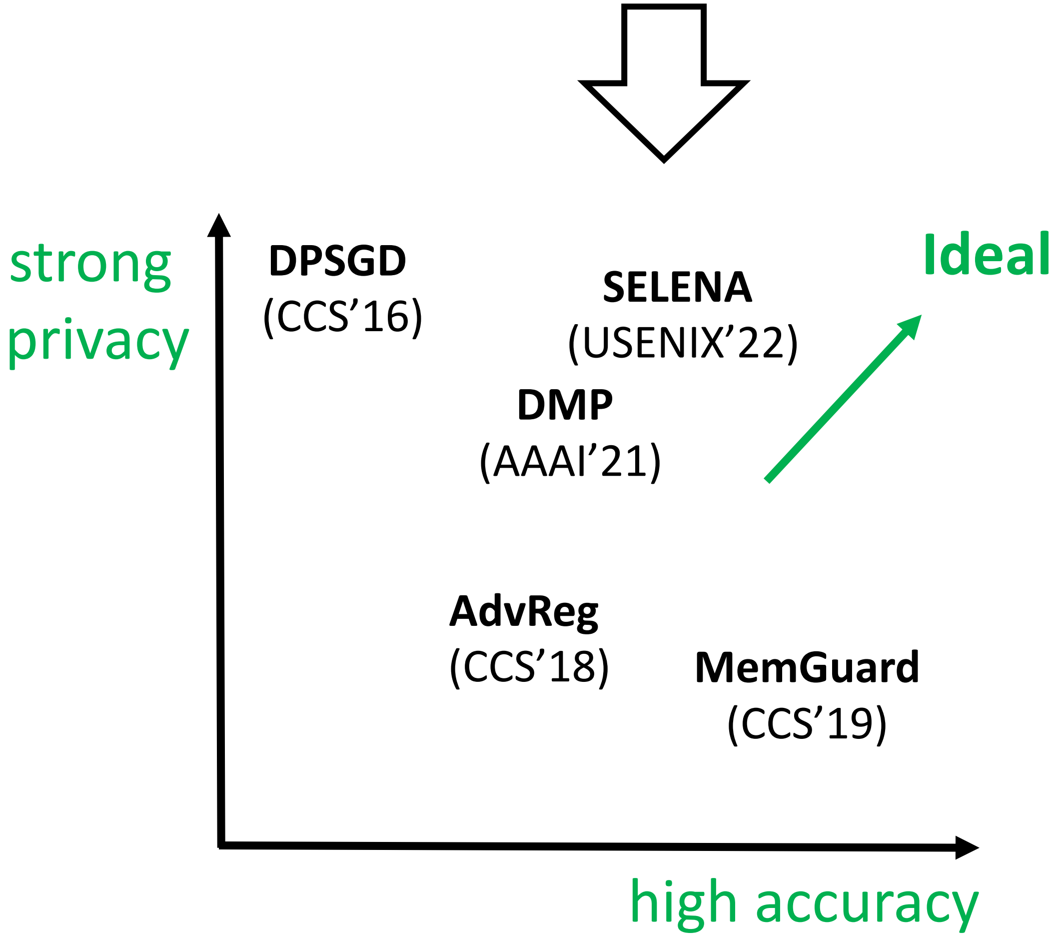
Existing defenses

Existing defenses

Poor privacy-utility trade off or requiring additional data

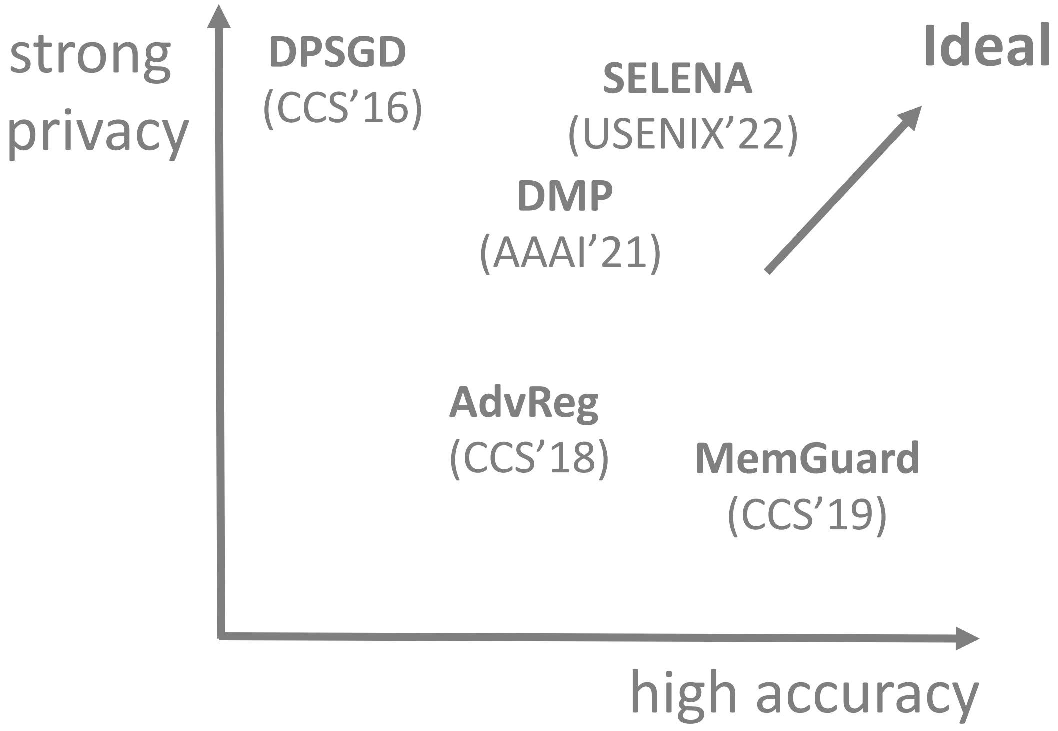
Existing defenses

Poor privacy-utility trade off **or requiring additional data**



Existing defenses

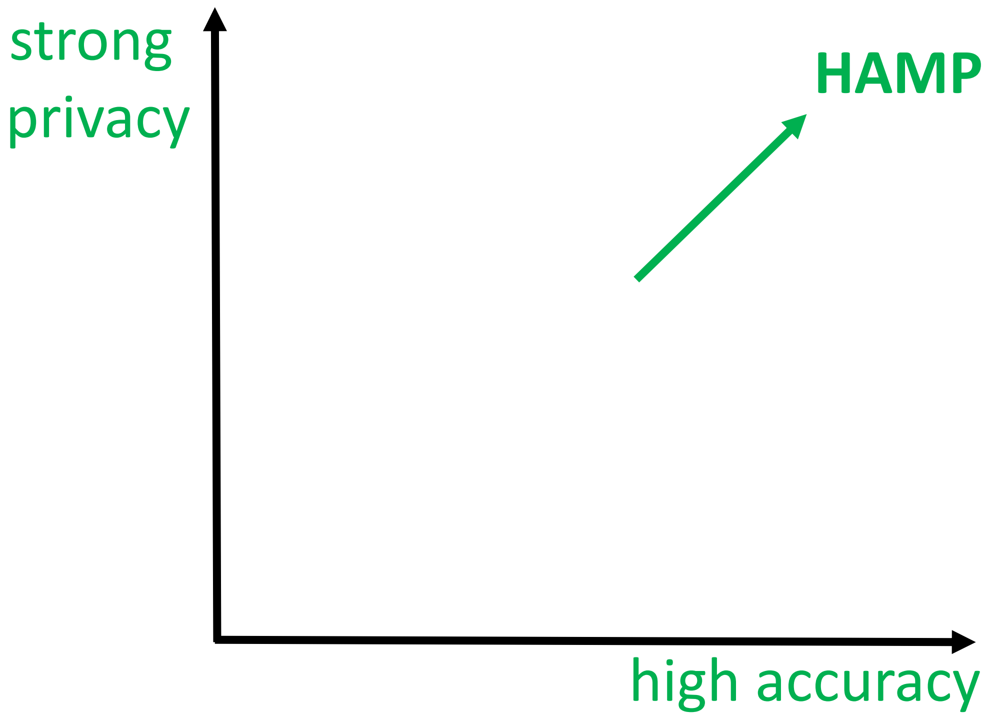
Poor privacy-utility trade off or requiring additional data



DMP (AAAI'21) **MMD** (CODASPY'21) **PATE** (ICLR'17)

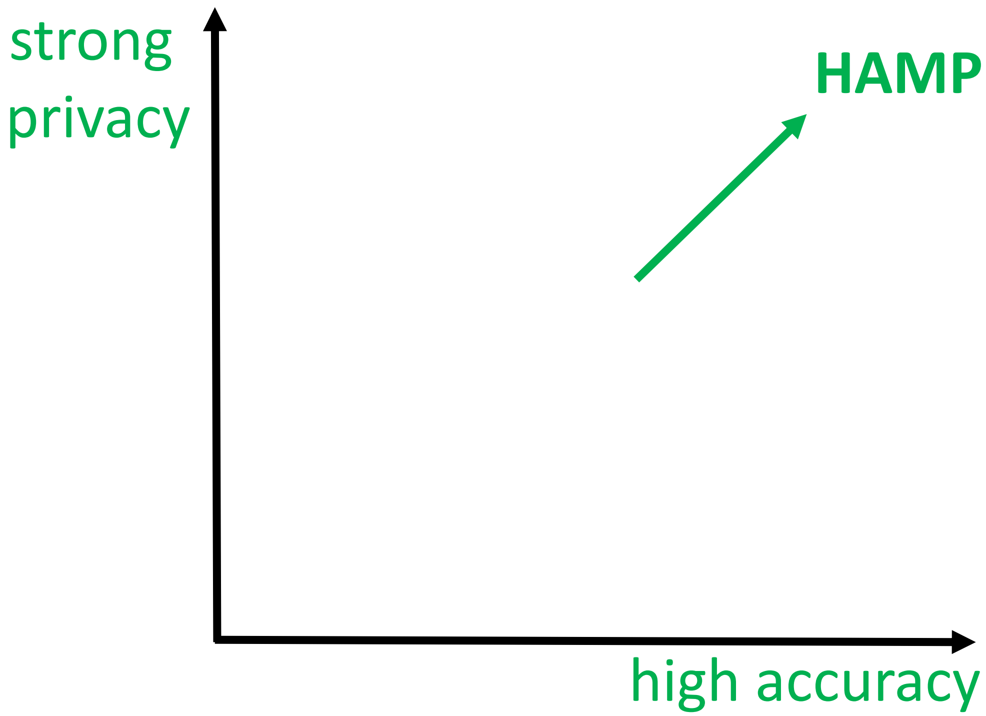
Our Work: HAMP

High Accuracy and Membership Privacy without additional data



Our Work: HAMP

A new way to combine soft label training, training regularization and output modification for privacy-preserving training!



Threat model

Adversary

- **Knowledge**:
  - Black-box adversary.
  - Half members and non-members.
  - Full defense knowledge.

- **Goal**: Membership inference

Threat model

Adversary

- **Knowledge**:
  - Black-box adversary.
  - Half members and non-members.
  - Full defense knowledge.
- **Goal**: Membership inference

Defender

- **Knowledge**:
  - The private dataset only.
- **Goal**: Model with high accuracy & membership privacy

Existing attacks

Diverse strategies

Existing attacks

Diverse strategies

- Scaled-logit loss
- Prediction entropy
- Adv robustness
- ...

Existing attacks

Diverse strategies

- Scaled-logit loss
- Prediction entropy
- Adv robustness
- ...

Common exploitation

Existing attacks

Diverse strategies

- Scaled-logit loss
- Prediction entropy
- Adv robustness
- …

Common exploitation

Exploit ML model's overconfident prediction on training samples

Existing attacks

Diverse strategies

Scaled-logit loss

Prediction entropy

Adv robustness

...

Common exploitation

Exploit ML model's overconfident prediction on training samples

Overconfident Prediction manifest as

Existing attacks

Diverse strategies

- Scaled-logit loss
- Prediction entropy
- Adv robustness
- …

Common exploitation

Exploit ML model's overconfident prediction on training samples

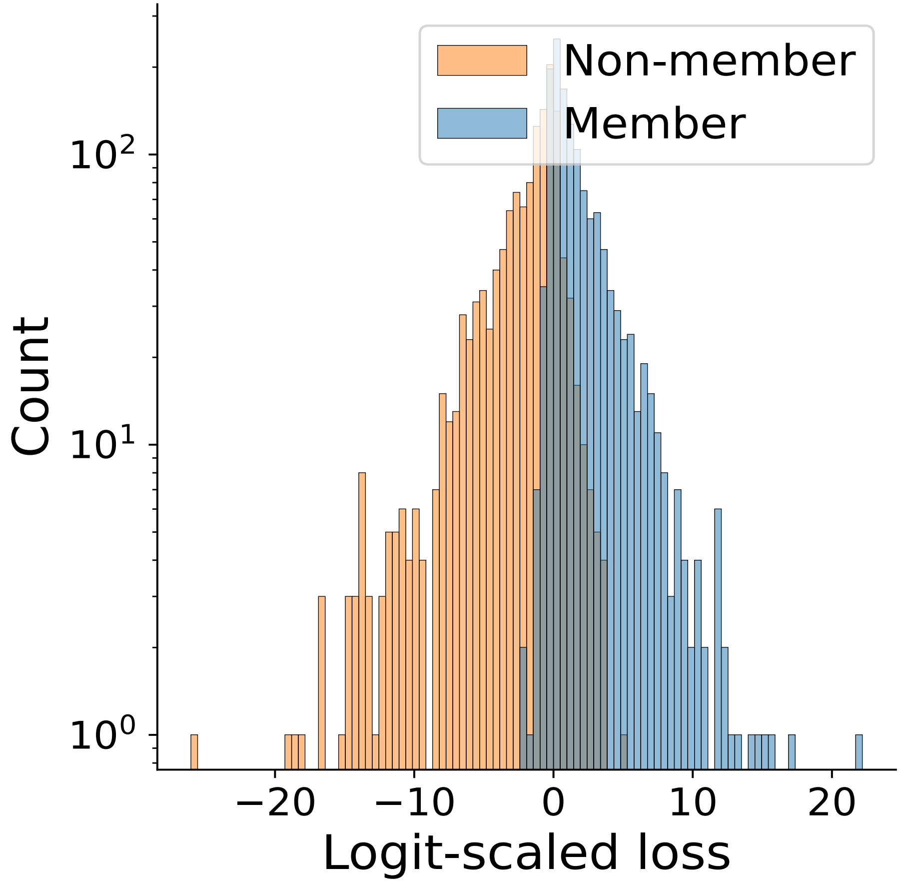
Overconfident Prediction manifest as

- High logit-scaled loss;
- Low prediction entropy;
- High robustnesss to adv perturbations;
- …

Example: Overconfident prediction via logit-scaled loss

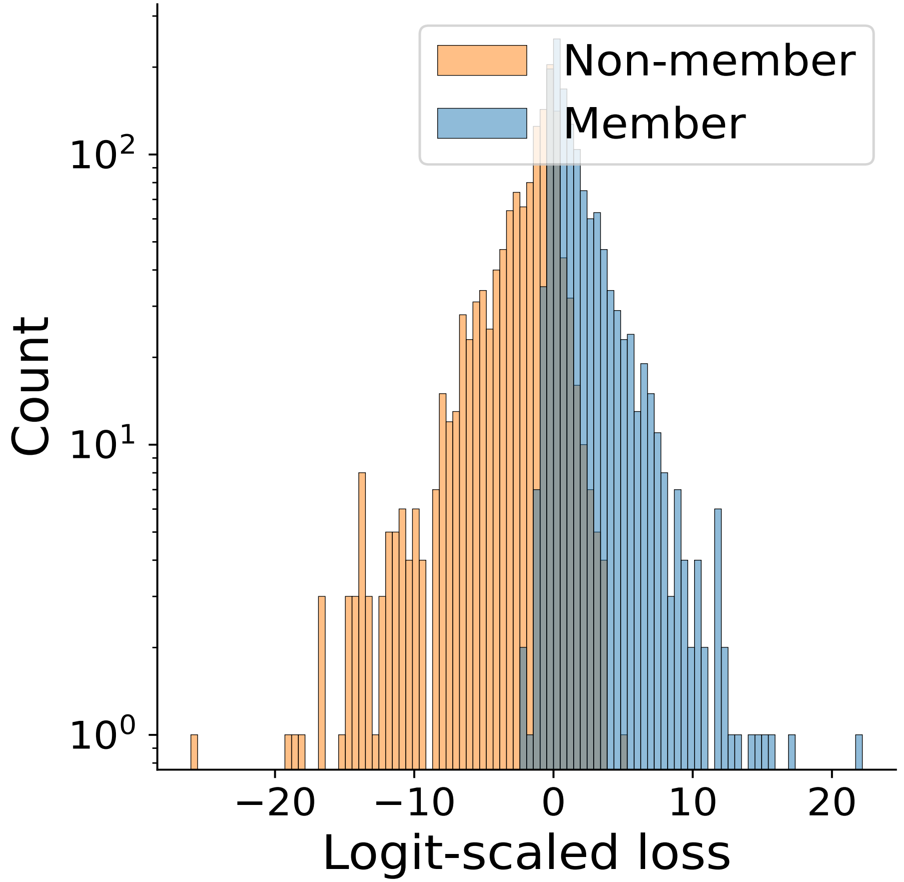
Carlini et al., Membership Inference Attacks From First Principles, S&P’22

Example: Overconfident prediction via logit-scaled loss



Carlini et al., Membership Inference Attacks From First Principles, S&P'22

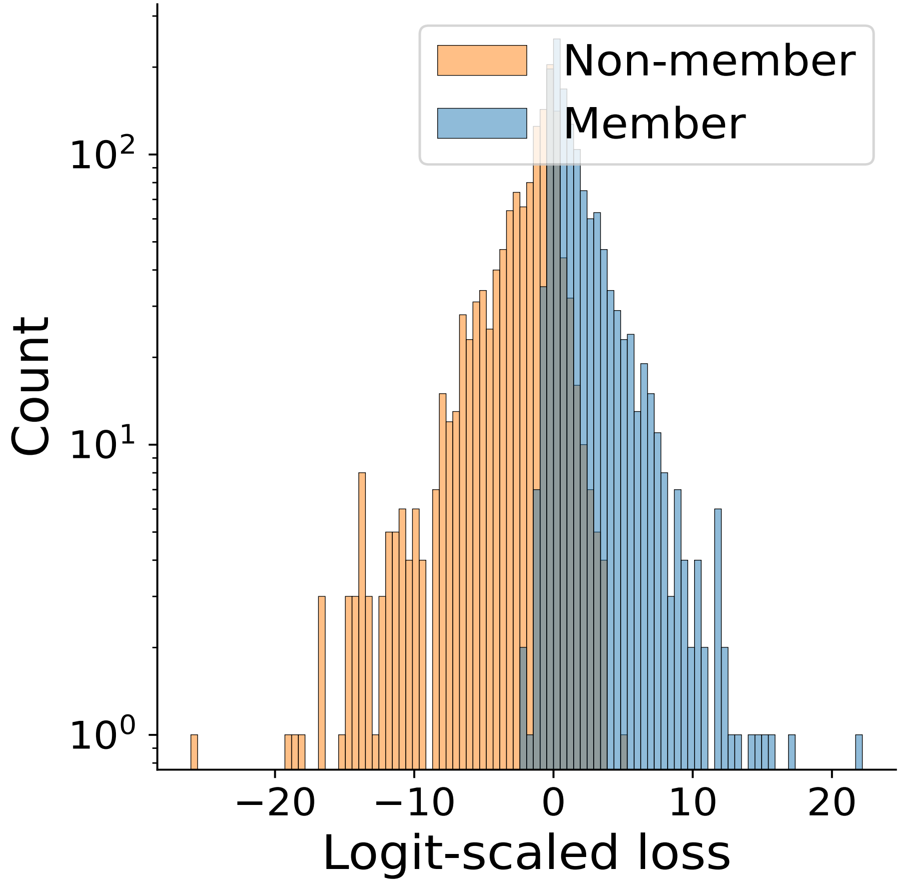
Example: Overconfident prediction via logit-scaled loss



Member samples with high scaled loss

Carlini et al., Membership Inference Attacks From First Principles, S&P'22

Example: Overconfident prediction via logit-scaled loss



Member samples with high scaled loss

Due to overly high prediction confidence

Carlini et al., Membership Inference Attacks From First Principles, S&P'22

Defense principle

MIAs exploit ML model's overconfident prediction on training samples

Mitigating ML model's overconfident prediction on training samples without jeopardizing model accuracy

What leads to overconfident prediction?

What leads to overconfident prediction?

- Training with one-hot hard label.

What leads to overconfident prediction?

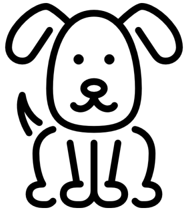
- Training with one-hot hard label.

Training

What leads to overconfident prediction?

- ❑ Training with one-hot hard label.

Training

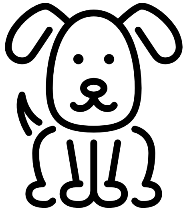


Dog	Cat
1	0

What leads to overconfident prediction?

- ❑ Training with one-hot hard label.

Training



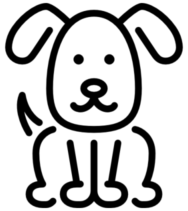
Dog	Cat
1	0

Testing

What leads to overconfident prediction?

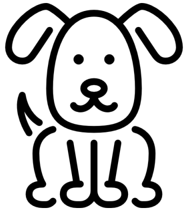
- Training with one-hot hard label.

Training



Dog	Cat
1	0

Testing

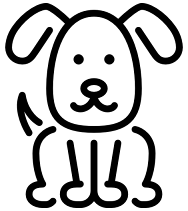


Dog	Cat
0.95	0.05

What leads to overconfident prediction?

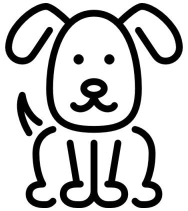
- Training with one-hot hard label.

Training



Dog	Cat
1	0

Testing

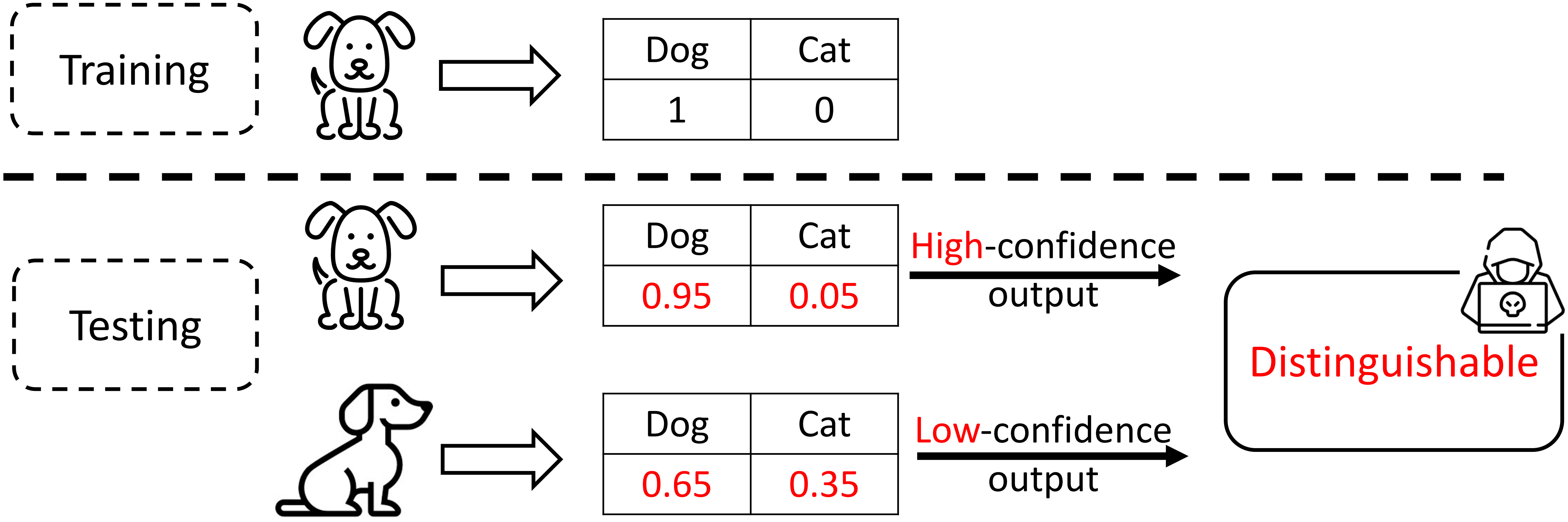


Dog	Cat
0.95	0.05

Dog	Cat
0.65	0.35

What leads to overconfident prediction?

- Training with one-hot hard label.



HAMP

Training-time defense

Testing-time defense

HAMP

Training-time defense

Testing-time defense

Produce high-utility models with strong membership privacy

HAMP

Training-time defense

Produce high-utility models with strong membership privacy

Testing-time defense

Gain higher privacy without degrading accuracy

HAMP - Training-time defense

High-entropy soft labels

HAMP - Training-time defense

High-entropy soft labels

Original hard label

Dog	Cat
1	0

High-entropy soft label

Dog	Cat
0.7	0.3

HAMP - Training-time defense

High-entropy soft labels

	Dog	Cat
Original hard label	1	0
High-entropy soft label	0.7	0.3

Explicitly enforce the model to make less confident prediction

HAMP - Training-time defense

High-entropy soft labels

Entropy-based regularization

Original hard label

Dog	Cat
1	0

High-entropy soft label

Dog	Cat
0.7	0.3

Explicitly enforce the model to make less confident prediction

HAMP - Training-time defense

High-entropy soft labels

Original hard label	Dog	Cat
	1	0

High-entropy soft label	Dog	Cat
	0.7	0.3

Explicitly enforce the model to make less confident prediction

Entropy-based regularization

Penalize low-entropy predictions

HAMP - Training-time defense

High-entropy soft labels

Original hard label

Dog	Cat
1	0

High-entropy soft label

Dog	Cat
0.7	0.3

Explicitly enforce the model to make less confident prediction

Entropy-based regularization

Penalize low-entropy predictions

Regularize the prediction confidence level

HAMP – Testing-time defense

HAMP – Testing-time defense

- ❑ Modify all output vectors → low confidence outputs.

HAMP – Testing-time defense

- ❑ Modify all output vectors → low confidence outputs.

Output vector

HAMP – Testing-time defense

- Modify all output vectors → low confidence outputs.

Output vector

Prediction label

(e.g., top-1 is dog)

HAMP – Testing-time defense

- Modify all output vectors → low confidence outputs.

Output vector

Prediction label
(e.g., top-1 is dog)

Confidence score
(e.g., top-1 with 90% confidence)

HAMP – Testing-time defense

- ❑ Modify all output vectors → low confidence outputs.

Output vector

Prediction label
(e.g., top-1 is dog)
→ Unchanged
(preserve accuracy)

Confidence score
(e.g., top-1 with 90% confidence)

HAMP – Testing-time defense

❑ Modify all output vectors → low confidence outputs.

Output vector

- **Prediction label** (e.g., top-1 is dog) → Unchanged (preserve accuracy)
- **Confidence score** (e.g., top-1 with 90% confidence) → Modified (improve privacy)

HAMP – Testing-time defense

- Modify all output vectors → low confidence outputs.
- How to obtain low confidence outputs?

HAMP – Testing-time defense

- Modify all output vectors → low confidence outputs.
- How to obtain low confidence outputs?
  - **Utilize random samples as (highly probable) non-members.**

HAMP – Testing-time defense

- ❑ Modify all output vectors → low confidence outputs.
- ❑ How to obtain low confidence outputs?
  - **Utilize random samples as (highly probable) non-members.**

Simple ✓
(optimization-free)

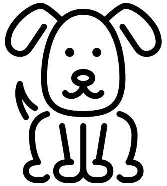
HAMP – Testing-time defense

- Modify all output vectors → low confidence outputs.
- How to obtain low confidence outputs?
  - **Utilize random samples as (highly probable) non-members.**

Simple ✓
(optimization-free)

Effective ✓
(improve membership privacy)

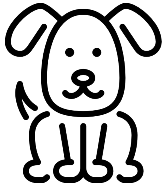
Output modification with random samples



Dog	Cat
0.85	0.15

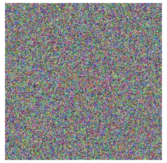
member sample

Output modification with random samples



Dog	Cat
0.85	0.15

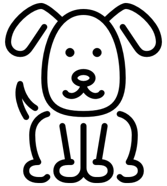
member sample



Dog	Cat
0.45	0.55

random sample

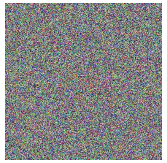
Output modification with random samples



Dog	Cat
0.85	0.15

Keep prediction label
(top-1 → dog)

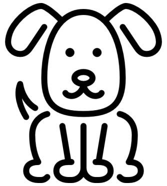
member sample



Dog	Cat
0.45	0.55

random sample

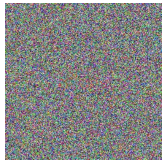
Output modification with random samples



Dog	Cat
0.85	0.15

Keep prediction label
(top-1 → dog)

member sample

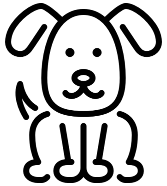


Dog	Cat
0.45	0.55

Keep prediction scores
(0.45, 0.55)

random sample

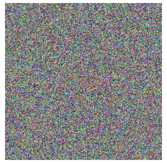
Output modification with random samples



Dog	Cat
0.85	0.15

member sample

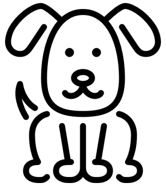
Keep prediction label
(top-1 → dog)



Dog	Cat
0.45	0.55

random sample

Keep prediction scores
(0.45, 0.55)



Dog	Cat
0.55	**0.45**

Low confidence output

Evaluation

Evaluation

5 datasets

Purchase100
Texas100
Location30
CIFAR10
CIFAR100

Evaluation

5 datasets

- Purchase100
- Texas100
- Location30
- CIFAR10
- CIFAR100

9 attacks

- NN-based
- Loss-based
- Entropy-based
- Modified-entropy-based
- Confidence-based
- **Likelihood-ratio attack (LiRA)**
- Correctness-based
- Boundary-based
- Augmentation-based

Evaluation

5 datasets

- Purchase100
- Texas100
- Location30
- CIFAR10
- CIFAR100

9 attacks

- NN-based
- Loss-based
- Entropy-based
- Modified-entropy-based
- Confidence-based
- **Likelihood-ratio attack (LiRA)**
- Correctness-based
- Boundary-based
- Augmentation-based

7 defenses

- AdvReg (CCS'18)
- MemGuard (CCS'19)
- DMP (AAAI'21)
- SELENA (USENIX'22)
- Early stopping (USENIX'21)
- Label Smoothing (CVPR'16)
- DPSGD (CCS'16)

Evaluation

5 datasets

- Purchase100
- Texas100
- Location30
- CIFAR10
- CIFAR100

9 attacks

- NN-based
- Loss-based
- Entropy-based
- Modified-entropy-based
- Confidence-based
- **Likelihood-ratio attack (LiRA)**
- Correctness-based
- Boundary-based
- Augmentation-based

7 defenses

- AdvReg (CCS'18)
- MemGuard (CCS'19)
- DMP (AAAI'21)
- SELENA (USENIX'22)
- Early stopping (USENIX'21)
- Label Smoothing (CVPR'16)
- DPSGD (CCS'16)

HAMP configuration

α for high-entropy soft labels

γ for regularization strength

Refer to the paper for details

Evaluation

5 datasets

- Purchase100
- Texas100
- Location30
- CIFAR10
- CIFAR100

9 attacks

- NN-based
- Loss-based
- Entropy-based
- Modified-entropy-based
- Confidence-based
- **Likelihood-ratio attack (LiRA)**
- Correctness-based
- Boundary-based
- Augmentation-based

7 defenses

- AdvReg (CCS'18)
- MemGuard (CCS'19)
- DMP (AAAI'21)
- SELENA (USENIX'22)
- Early stopping (USENIX'21)
- Label Smoothing (CVPR'16)
- DPSGD (CCS'16)

HAMP configuration

α for high-entropy soft labels

γ for regularization strength

Refer to the paper for details

2 metrics

- TPR @ 0.1% FPR
- TNR @ 0.1% FNR

Evaluation

Artifact Evaluated NDSS SYMPOSIUM
Available
Functional
Reproduced

5 datasets
- Purchase100
- Texas100
- Location30
- CIFAR10
- CIFAR100

9 attacks
- NN-based
- Loss-based
- Entropy-based
- Modified-entropy-based
- Confidence-based
- **Likelihood-ratio attack (LiRA)**
- Correctness-based
- Boundary-based
- Augmentation-based

7 defenses
- AdvReg (CCS'18)
- MemGuard (CCS'19)
- DMP (AAAI'21)
- SELENA (USENIX'22)
- Early stopping (USENIX'21)
- Label Smoothing (CVPR'16)
- DPSGD (CCS'16)

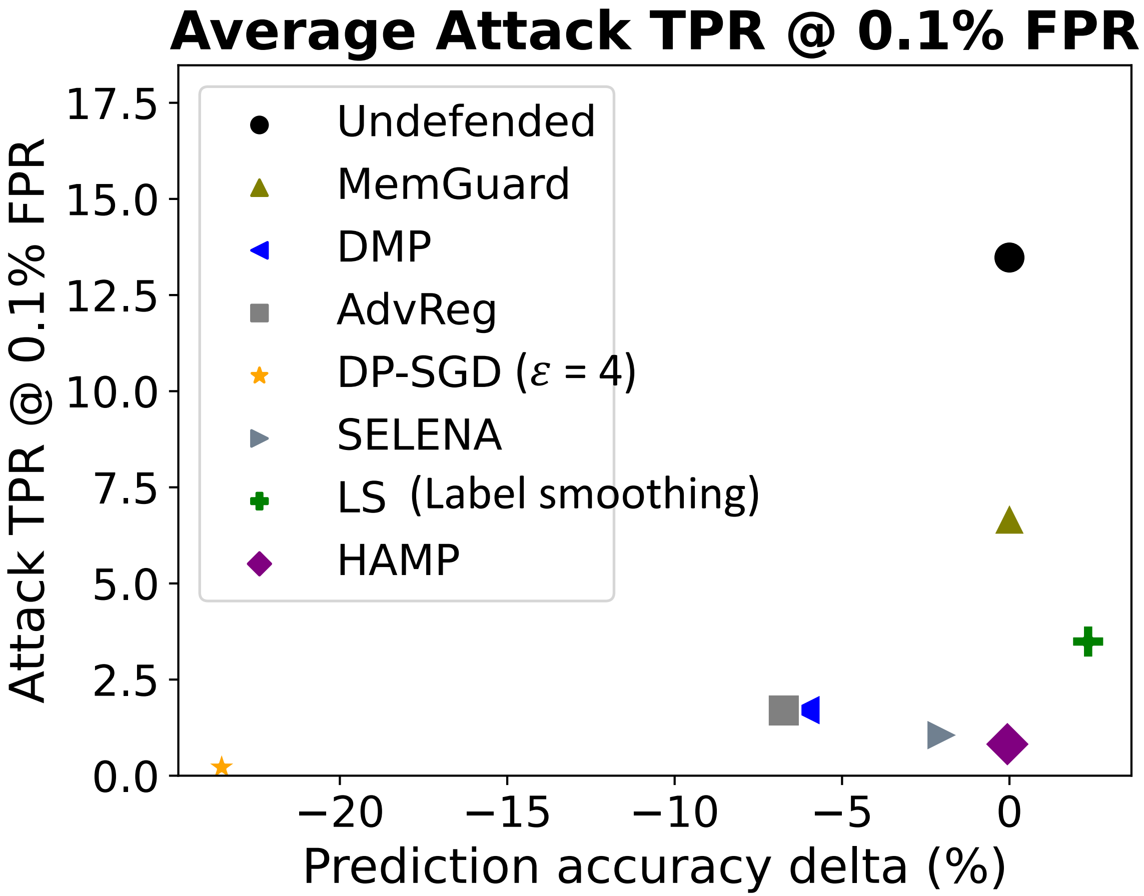
HAMP configuration
- **α for high-entropy soft labels**
- **γ for regularization strength**

Refer to the paper for details

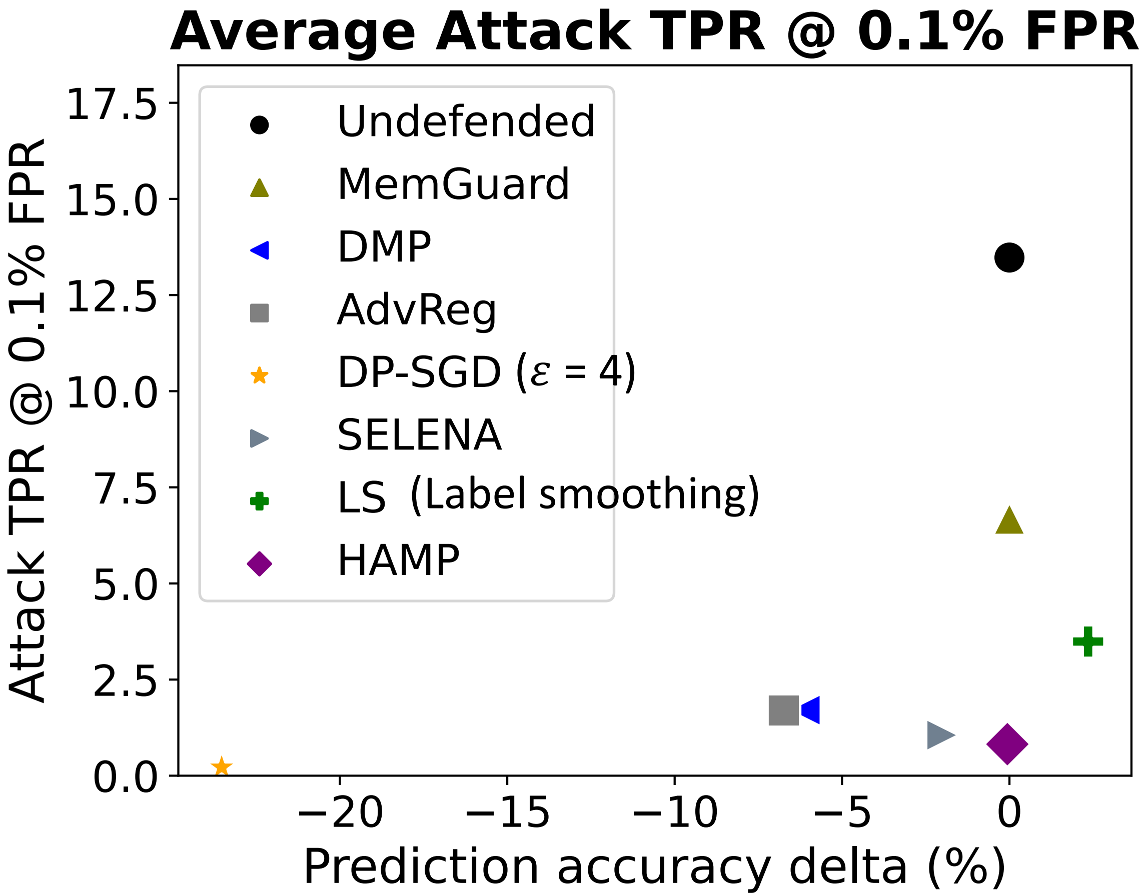
2 metrics
- TPR @ 0.1% FPR
- TNR @ 0.1% FNR

Artifact https://github.com/DependableSystemsLab/MIA_defense_HAMP

Key results

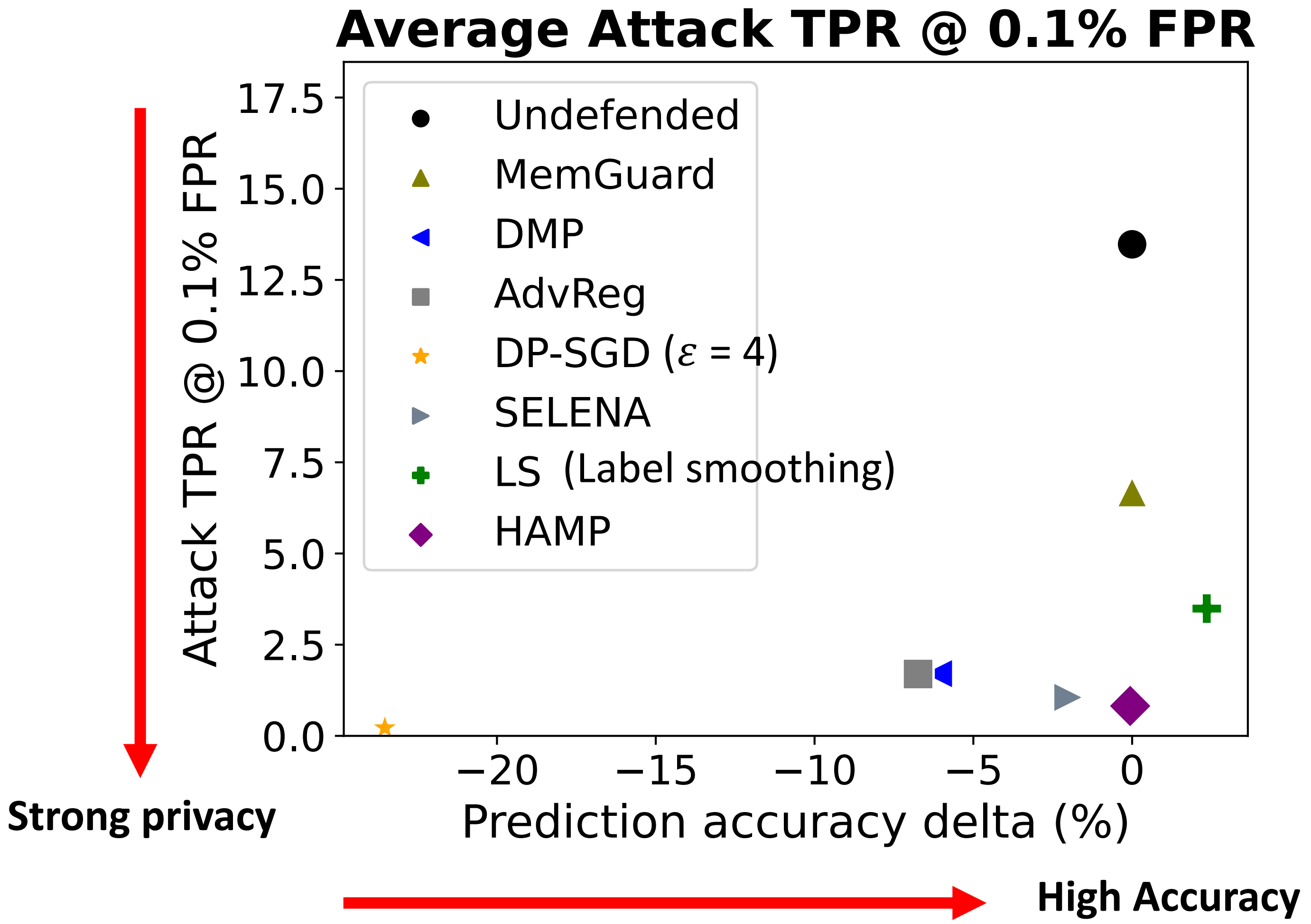

Average Attack TPR @ 0.1% FPR
Attack TPR @ 0.1% FPR
Prediction accuracy delta (%)
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
−20
−15
−10
−5
0
Undefended
MemGuard
DMP
AdvReg
DP-SGD (ε = 4)
SELENA
LS (Label smoothing)
HAMP

Key results

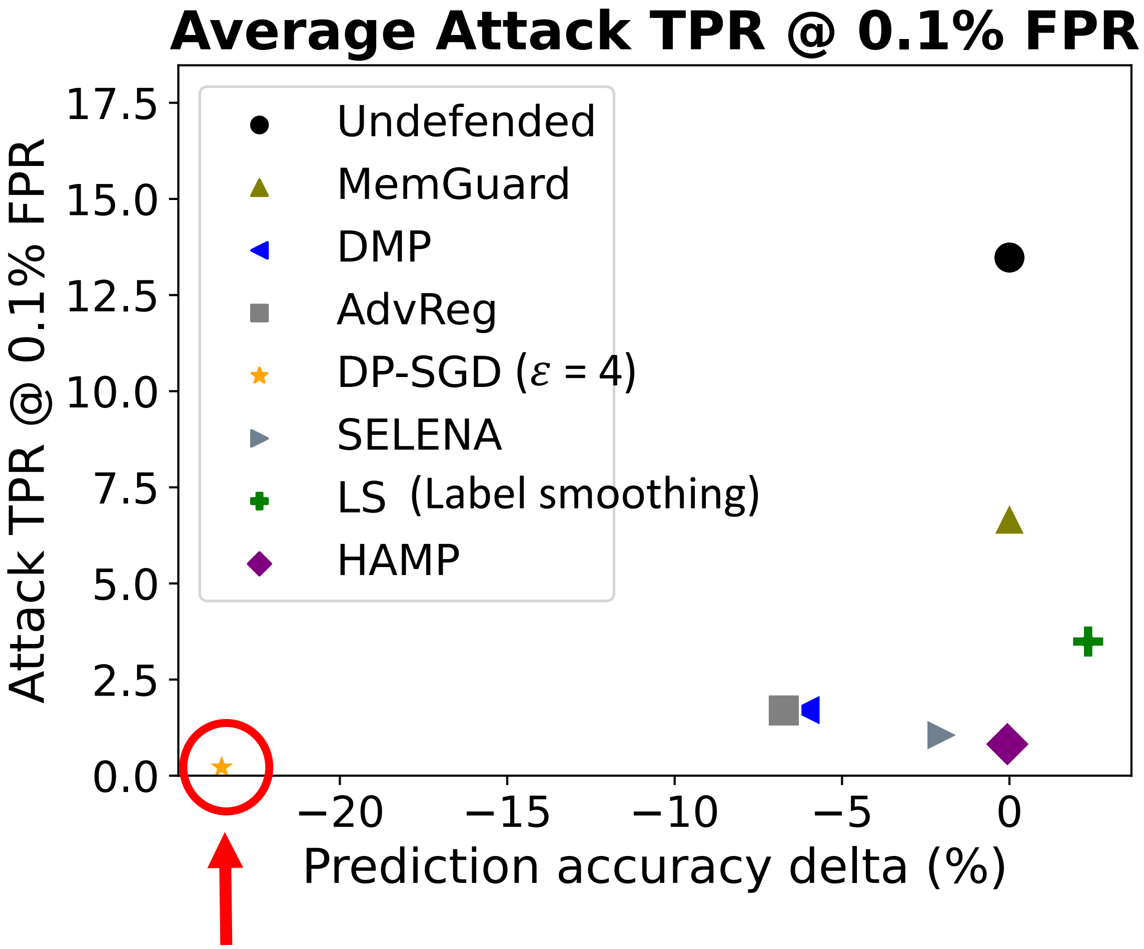


High Accuracy

Key results

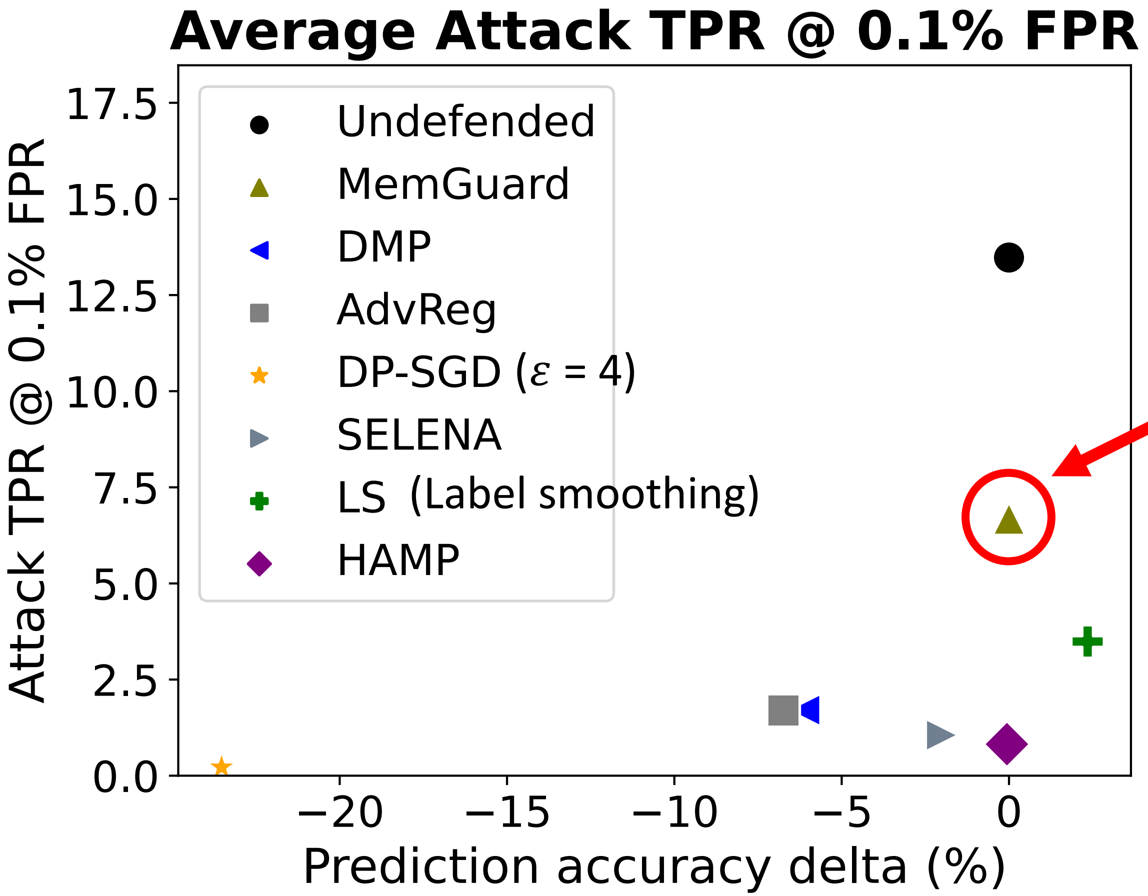

Average Attack TPR @ 0.1% FPR
Attack TPR @ 0.1% FPR
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
−20
−15
−10
−5
0
Prediction accuracy delta (%)
Undefended
MemGuard
DMP
AdvReg
DP-SGD (ε = 4)
SELENA
LS (Label smoothing)
HAMP
Strong privacy
High Accuracy

Key results



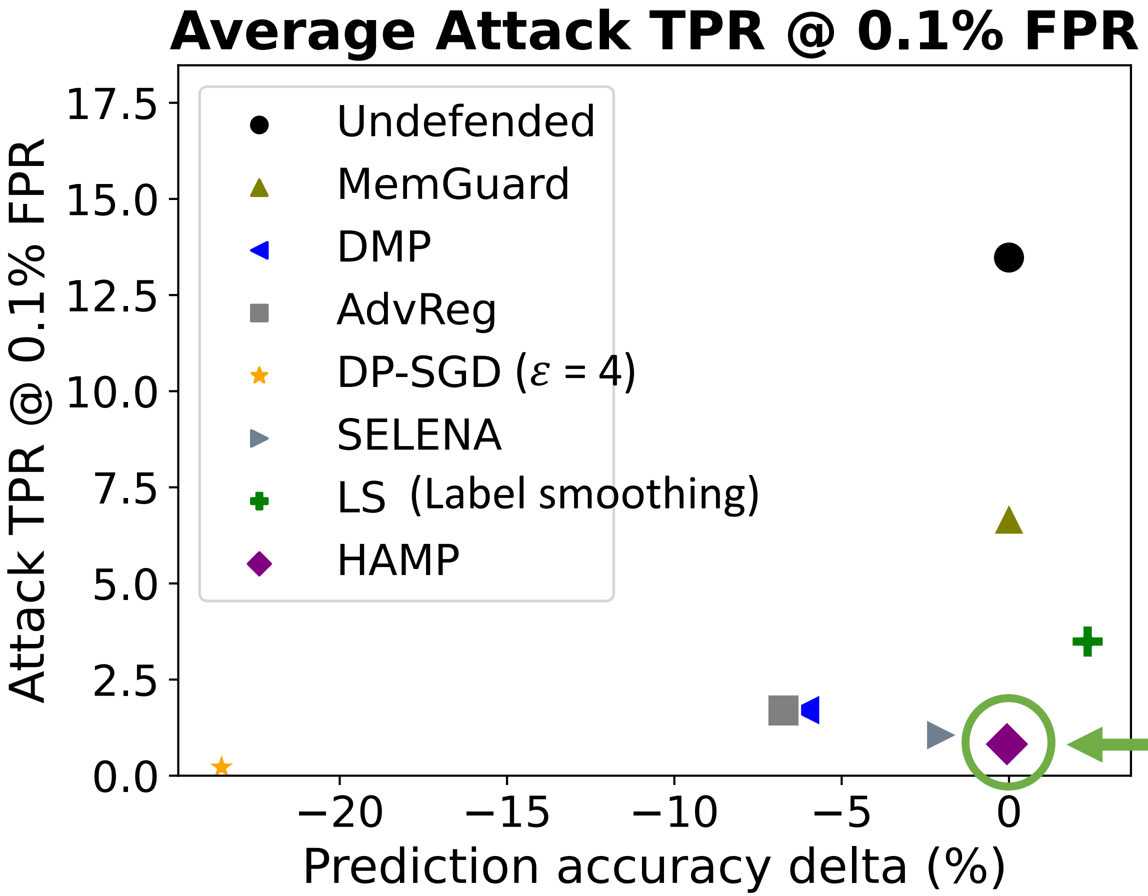
DPSGD: Strong privacy, but low accuracy

Key results



MemGuard: High accuracy, but poor privacy

Key results



HAMP

Strong privacy: attack TPR ↓ 94%

High accuracy: 0.46% accuracy drop

Summary

Artifact Evaluated
NDSS SYMPOSIUM
Available
Functional
Reproduced

❓ How to mitigate membership inference attacks with strong privacy protection and low accuracy drop?

Summary

Artifact Evaluated NDSS SYMPOSIUM
Available
Functional
Reproduced

How to mitigate membership inference attacks with strong privacy protection and low accuracy drop?

Mitigating ML model's overconfident prediction on training samples without jeopardizing model accuracy.

Summary

Artifact Evaluated
NDSS SYMPOSIUM
Available
Functional
Reproduced

? How to mitigate membership inference attacks with strong privacy protection and low accuracy drop?

Mitigating ML model's overconfident prediction on training samples without jeopardizing model accuracy.

HAMP: A new way to combine soft label training, training regularization and output modification for privacy-preserving training!

Paper

Code

zitaoc@ece.ubc.ca