To Trust or Not To Trust Prediction Scores for Membership Inference Attacks


Dominik Hintersdorf *, 1


Lukas Struppek *, 1

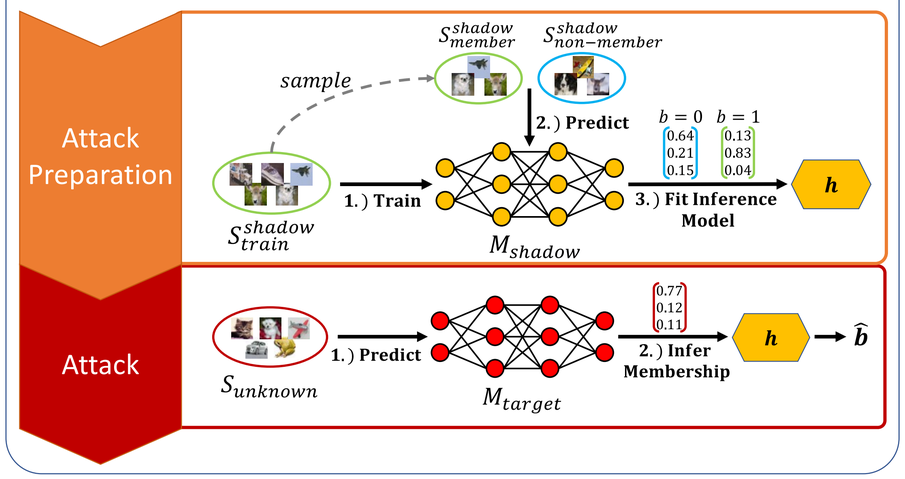

Kristian Kersting 1, 2

* Equal Contribution [1] Technical University of Darmstadt, Germany
[2] Centre for Cognitive Science, TU Darmstadt and [3]Hessian Center for AI (hessian.AI), Germany

What are Membership Inference Attacks (MIAs)?

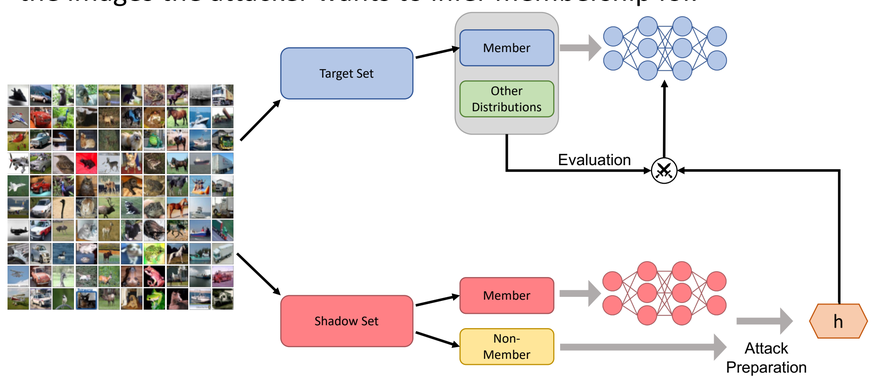
Given **datapoint x** and **model M** trained on **dataset D**, the attacker tries to answer the following question:

Was x part of the training dataset D?



Experimental Setup

Samples from other distributions with similar content might be in the images the attacker wants to infer membership for.



Evaluated Score-based Attacks (Salem et al., '19)

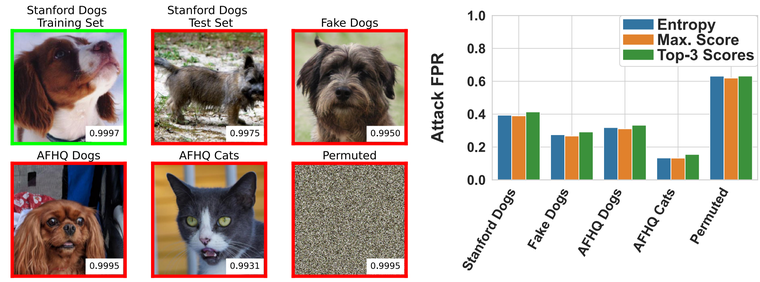
Entropy Attack	Max. Pred. Score Attack	Top-3 Scores Attack
0.64	0.64	0.64
0.09	0.09	0.09
0.15	0.15	0.15
0.12	0.12	0.12

Evaluated Models

Training Set	Architecture
Stanford Dogs (Khosla et al., '11)	ResNet-50 (Results on this Poster!)
CIFAR-10 (Krizhevsky, '09)	ResNet-18, Simple CNN, EfficientNetB0

1) Score-based MIAs Are Not Robust!

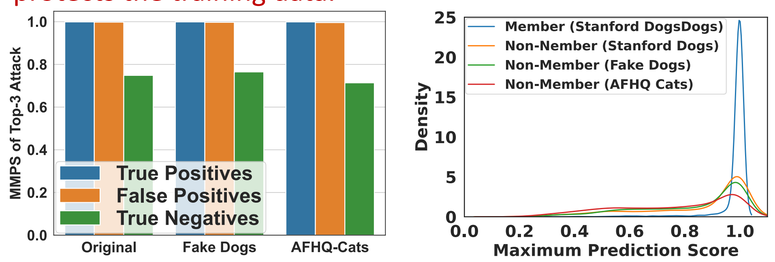
Results: MIAs have high false-positive rates on samples from the same and different data distributions.



2) High Prediction Scores, Lower Privacy Risks

Results:

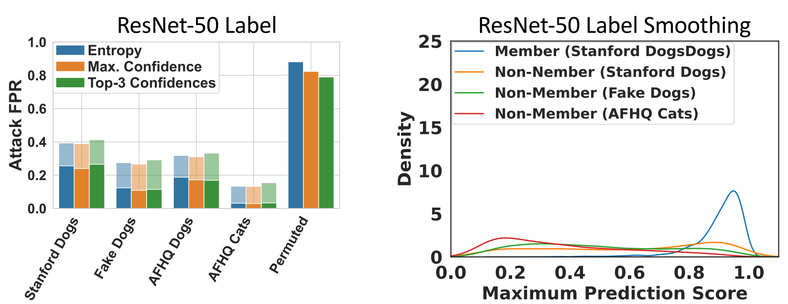
- The Mean Maximum Prediction Scores (MMPS) indicate that score-based MIAs rely primarily on the max. prediction score.
- Overconfidence causes high false-positive rates which implicitly protects the training data.



3) Mitigating Overconfidence Increases Privacy Risks

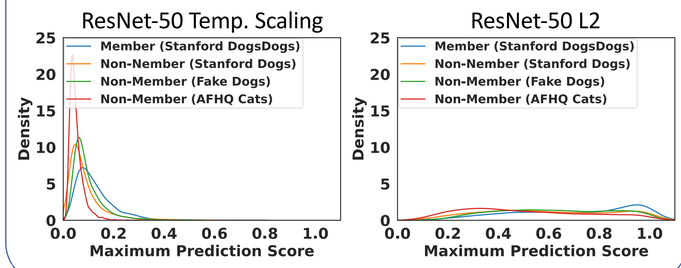
Results:

- Label Smoothing (Müller et al., '19) and LLLA (Kristiadi et al., '20) reduce the FPR of MIA attacks
- Lower FPR ➔ Higher Privacy Risks



4) Tradeoff between Calibration and Defenses

Results: Defenses have contrary effects to calibration. Calibration separates the distributions while defenses align them.



Conclusion

- MIAs have high false-positive rates
- Overconfidence causes high false-positive rates
- Calibration increases privacy risks
- Defenses are contrary to calibration

Contact

Dominik Hintersdorf
Technical University of Darmstadt
dominik.hintersdorf@cs.tu-darmstadt.de
@D0miH

Lukas Struppek
Technical University of Darmstadt
lukas.struppek@cs.tu-darmstadt.de
@LukasStruppek

Code

https://github.com/ml-research/To-Trust-or-Not-To-Trust-Prediction-Scores-for-Membership-Inference-Attacks