Membership Inference Attacks against Adversarially Robust Deep Learning Models

Liwei Song[1], Reza Shokri[2], Prateek Mittal[1]

[1]Princeton University, [2]National University of Singapore

Deep Learning

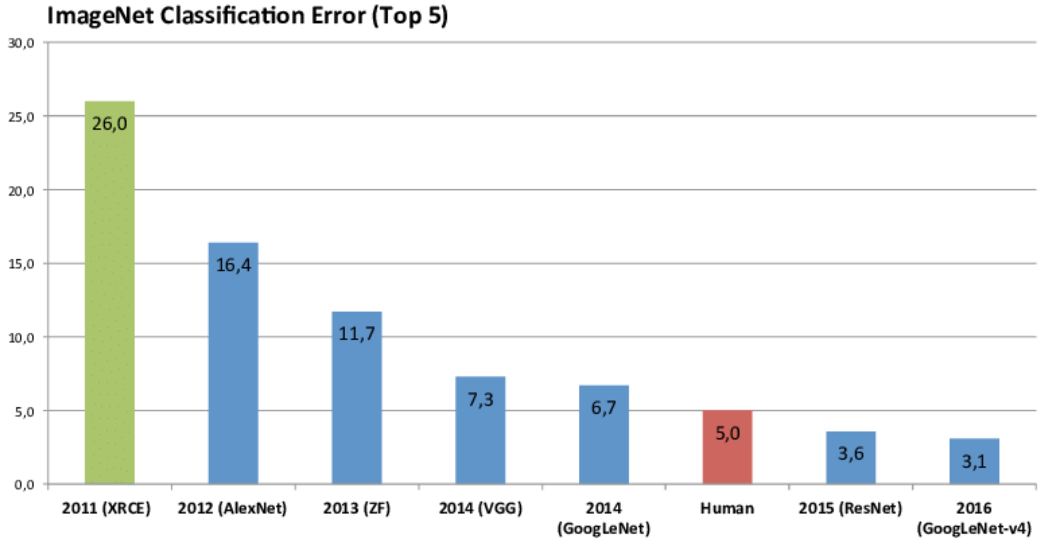


Image Classification

SQuAD1.1 Leaderboard

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) *Google AI Language* https://arxiv.org/abs/1810.04805	**87.433**	**93.160**
2 Sep 09, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.356	91.202
3 Jul 11, 2018	QANet (ensemble) *Google Brain & CMU*	84.454	90.490

Natural Language Processing

OpenAI
@OpenAI
Following

OpenAI Five is now the first AI to beat the world champions in an esports game. Here's what happened, and how we made our comeback since losing to pros in Aug 2018: openai.com/blog/how-to-tr ...

10:18 AM - 15 Apr 2019

Game Playing

Security Vulnerabilities of Deep Learning

- **Evasion Attacks** (Biggio et al., *ECML PKDD'13;* Goodfellow et al., *ICLR'15*; Carlini & Wagner, *S&P'17*)
  - Perturb inputs at the test time to induce model misclassifications.

- **Poisoning Attacks** (Biggio et al., *ICML'12*; Koh & Liang, *ICML'17;* Shafahi et al., *NeurIPS'18*)
  - Manipulate part of training data to compromise the trained models.

Privacy Vulnerabilities of Deep Learning

- **Membership Inference** (Shokri et al., *S&P'17*)
  - Infer whether an input was used to trained the model or not.
- **Property Inference** (Ganju et al., *CCS'18*)
  - Learn global property of training data.
- **Model Inversion** (Fredrikson et al., *CCS'15*)
  - Reconstruct training data from model predictions.
- **Malicious Training** (Song et al., *CCS'17*)
  - Modify the training algorithm to memorize sensitive information.

Defenses to Mitigate Security & Privacy Vulnerabilities

❑ Defenses against Security Vulnerabilities

- Madry et al., “Towards deep learning models resistant to adversarial attacks”, *ICLR’18*;
- Wong & Kolter, “Provable defenses against adversarial examples via the convex outer adversarial polytope”, *ICML’18*;
- Steinhardt et al., “Certified defense against data poisoning attacks”, *NeurIPS’17*;
- Jagielski et al., “Poisoning attacks and countermeasures for regression learning”, *S&P’18*.

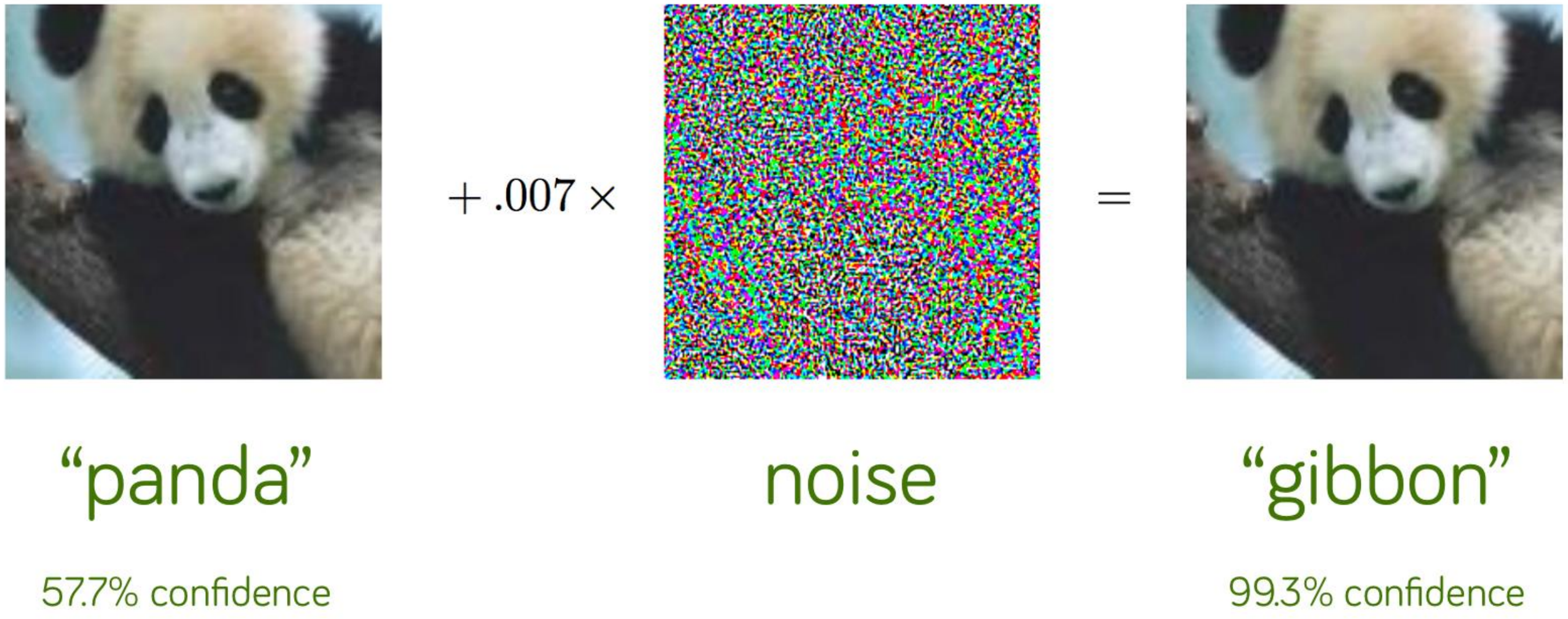
❑ Defenses against Privacy Vulnerabilities

- Nasr et al., “Machine learning with membership privacy using adversarial regularization”, *CCS’18*;
- Shokri & Shmatikov, “Privacy-preserving deep learning”, *CCS’15*;
- Abadi et al., “Deep learning with differential privacy”, *CCS’16*.

The security domain and the privacy domain typically have been considered separately!

Adversarial Examples (Evasion Attacks)

- **Adversarial goal**: cause model misclassifications at test time by add small perturbations to inputs.



Goodfellow et al., “Explaining and Harnessing Adversarial Examples”, *ICLR’15*

Robustness against Adversarial Examples

- Natural training to minimize prediction loss of model F_θ.

$$\min_\theta \frac{1}{|D_{train}|} \sum_{(x,y)\in D_{train}} \ell(F_\theta(x), y)$$

- Adversarial example to maximize loss under the constraint Δ (e.g., $||\Delta||_\infty \leq \varepsilon$).

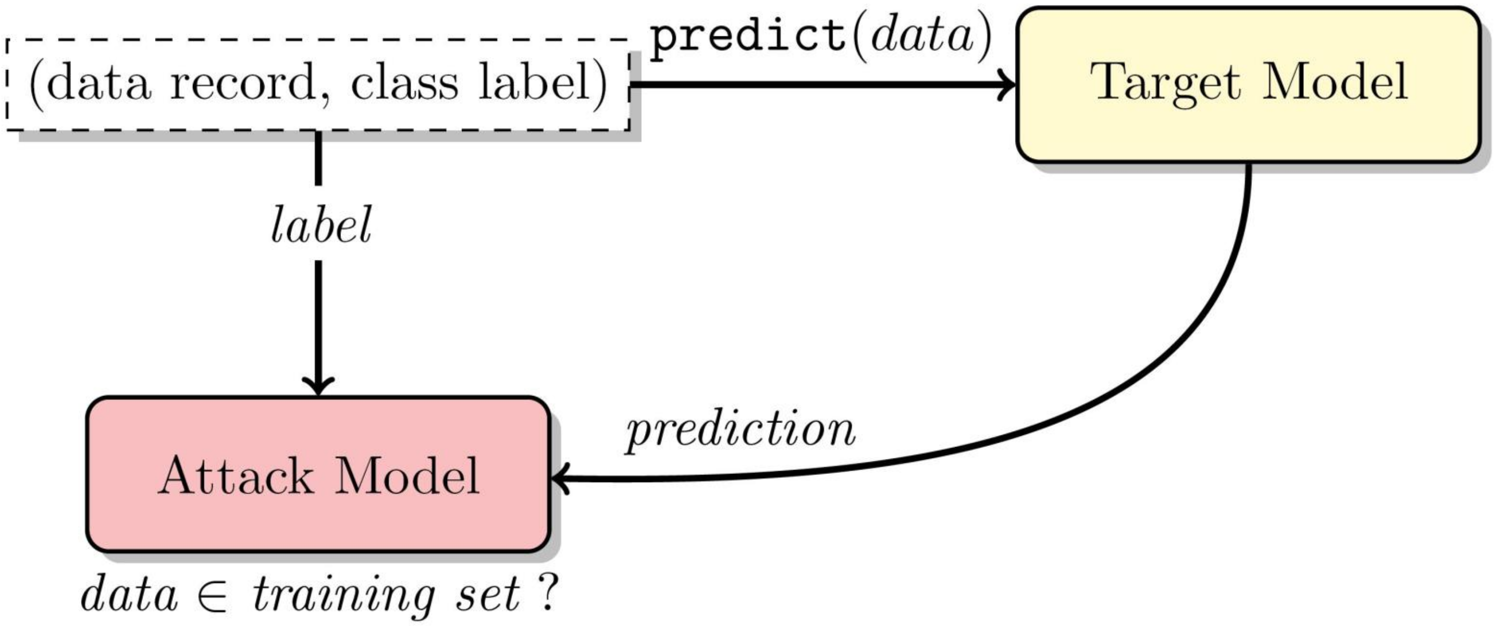
$$\max_{\delta\in\Delta} \ell(F_\theta(x+\delta), y)$$

- Robust training to minimize adversarial loss.

$$\min_\theta \frac{1}{|D_{train}|} \sum_{(x,y)\in D_{train}} \max_{\delta\in\Delta} \ell(F_\theta(x+\delta), y)$$

Membership Inference

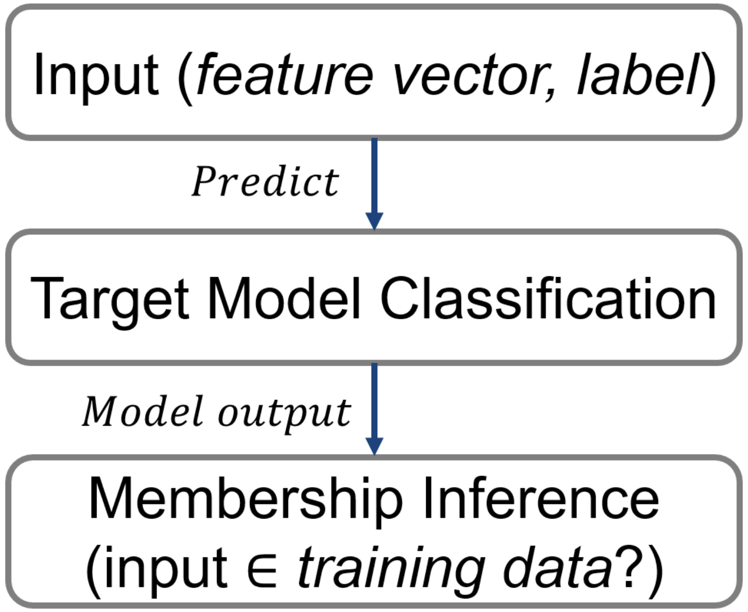
- **Adversarial goal**: guess whether an input example was used to train the target model or not.



Shokri et al., "Membership Inference Attacks Against Machine Learning Models", *S&P' 17*

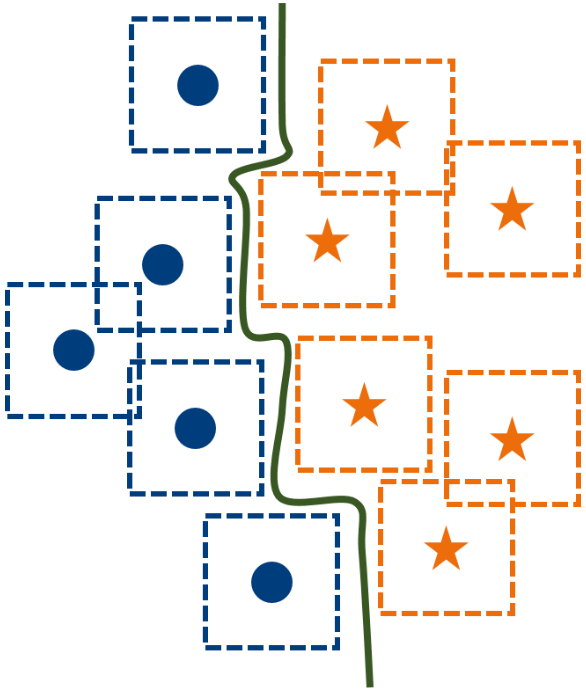
Membership Inference Attacks against Adversarially Robust Models

Membership Inference Attack



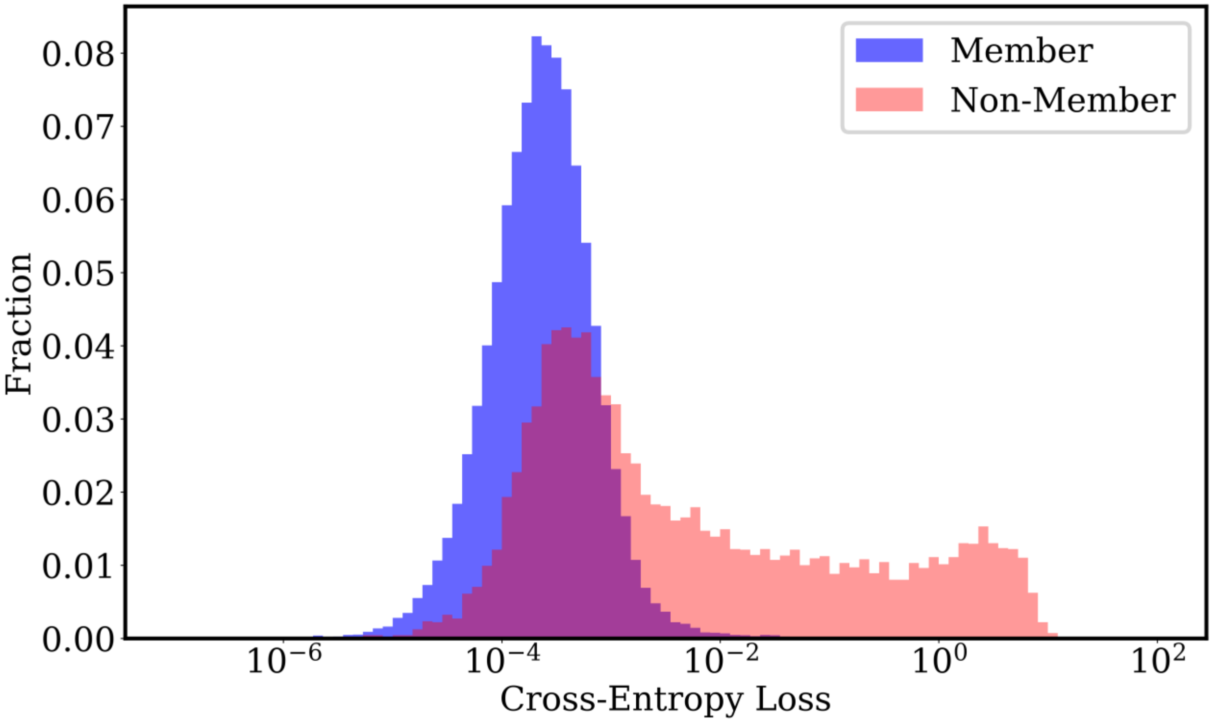
- ❑ Highly related to target model's overfitting.
- ❑ Also measured by model's sensitivity as to training data.

Adversarial Robustness

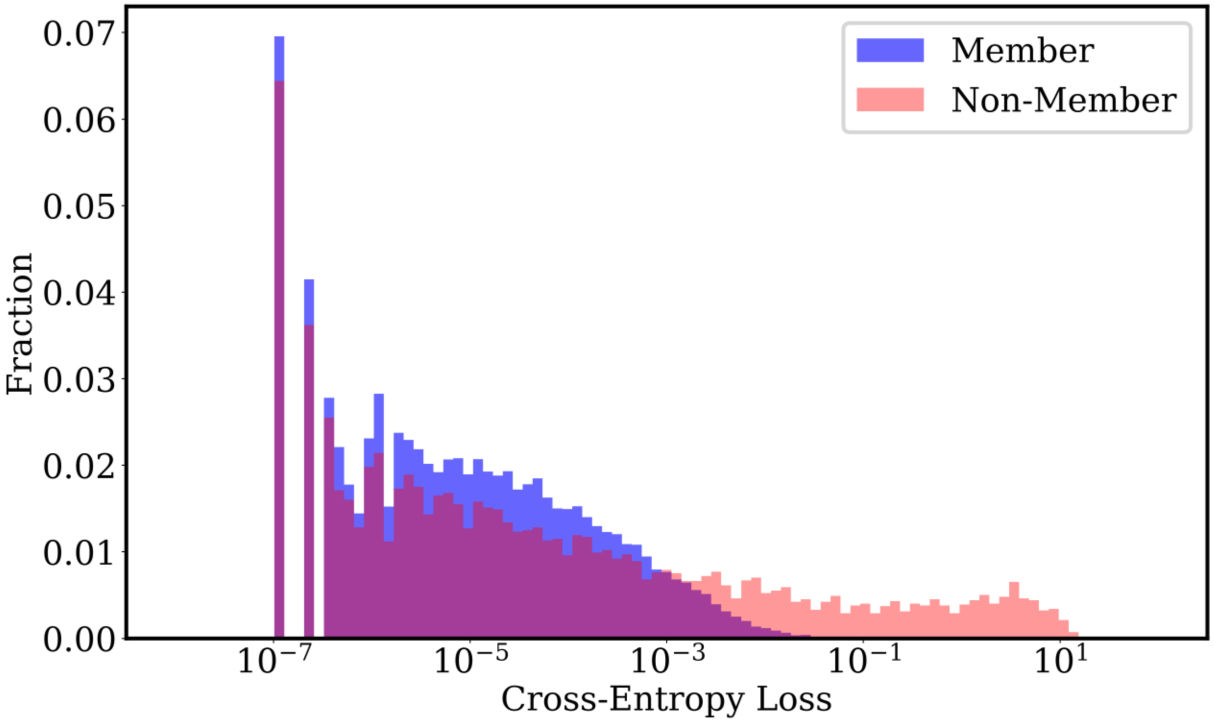


- ❑ May result in more overfitting and larger model sensitivity.
- ❑ Make the model more susceptible to membership inference attacks.

Adversarially robust models may leak more privacy



Robust CIFAR10 classifier (Madry et al., *ICLR'18*)

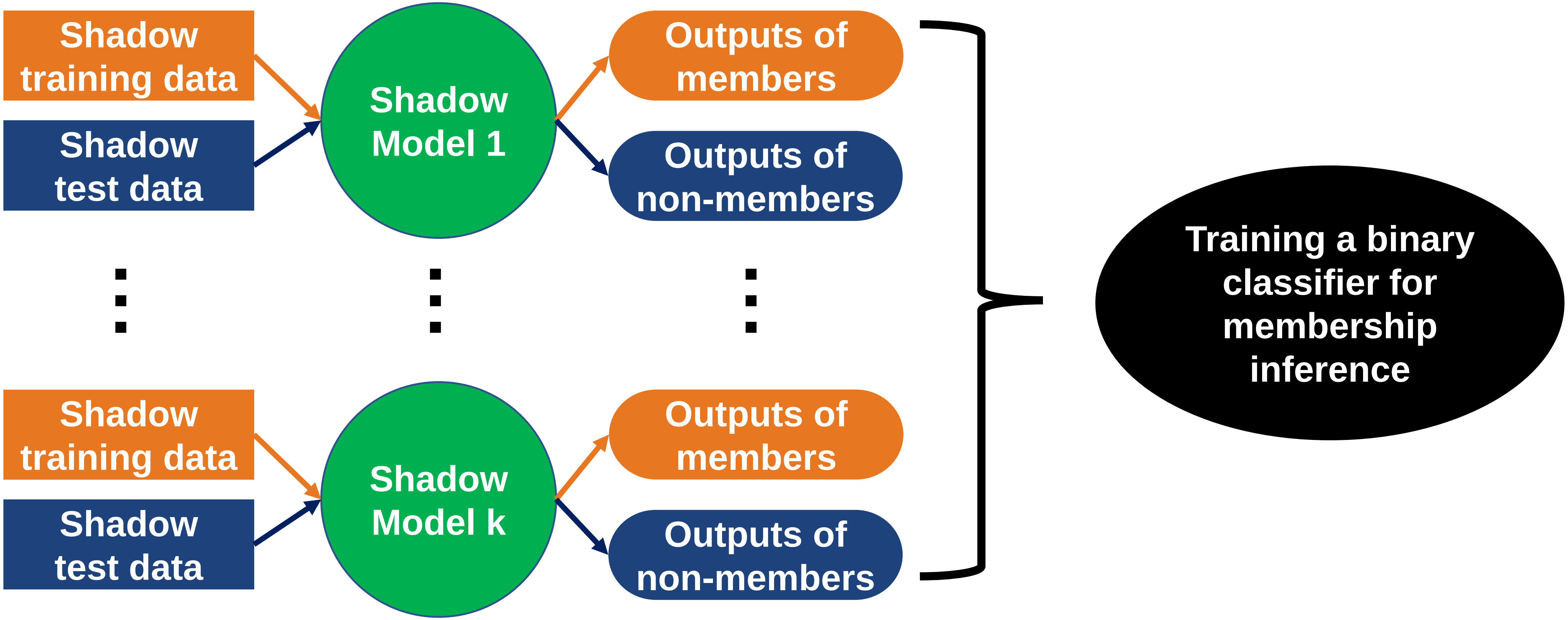


Natural (undefended) CIFAR10 classifier

The robust model has a larger divergence between loss distributions over members (training data) and non-members (test data).

Membership Inference Attacks (black-box setting)

- Inference based on shadow training (Shokri et al., *S&P'17*)



Membership Inference Attacks (Our Choice)

- Inference based on **prediction confidence** (Yeom et al., *CSF'18*)

$$\mathcal{I}(\mathcal{F}, (\boldsymbol{x}, y)) = \begin{cases} \text{member}, & \text{if } \mathcal{F}_y(\boldsymbol{x}) \geq \tau; \\ \text{non–member}, & \text{otherwise} \end{cases}$$

- Evaluate the **worst-case inference risk** by setting the threshold τ to achieve highest inference accuracy, which could be learned using shadow training in practice.

Membership Inference Attacks

- Sample the input $(\boldsymbol{x}, y)$ from either training dataset or test dataset with an equal **50%** probability.

- Evaluation Metrics: **inference accuracy, precision, recall**.

- Random guessing strategy results in 50% inference accuracy and 50% precision.

- Targeted adversarially robust models: **adversarial training** (Madry et al., *ICLR'18*), and **provable defense** (Wong & Kolter, *ICML'18*).

Inference Attacks against Adversarial Training (Madry et al., *ICLR'18*)

- Adversarial training makes models more susceptible to inference attack.
  - CIFAR10 dataset: wide ResNet, robustly trained with the l_∞ constraint $\varepsilon = 8/255$
  - SVHN dataset: wide ResNet, robustly trained with the l_∞ constraint $\varepsilon = 4/255$

Models	Train Acc	Test Acc	Adv-Train Acc	Adv-Test Acc	Infer Acc	Precision	Recall
CIFAR10 (natural)	100%	95.01%	0%	0%	**57.37%**	54.16%	96.00%
CIFAR10 (robust)	99.99%	87.25%	96.07%	46.59%	**74.86%**	69.08%	90.00%
SVHN (natural)	99.99%	95.64%	6.53%	3.86%	**56.79%**	53.72%	98.00%
SVHN (robust)	99.99%	93.91%	99.74%	72.17%	**64.30%**	59.70%	88.00%

Relation with Adversarial Perturbation Budget

Datasets	Perturbation Budget	Infer Acc
CIFAR10	2/255	**64.40%**
CIFAR10	4/255	**69.34%**
CIFAR10	8/255	**74.86%**
SVHN	2/255	**60.69%**
SVHN	4/255	**64.30%**
SVHN	8/255	**68.09%**

The robust model trained with a larger perturbation budget has an increased risk against membership inference attacks.

Inference Attacks against Provable Defense (Wong & Kolter, *ICML'18*)

- Provable defense does not increase membership inference accuracy, with a cost of accuracy degradation.
  - CIFAR10 dataset: ResNet, robustly trained with the l_∞ constraint $\varepsilon = 2/255$
  - SVHN dataset: CNN, robustly trained with the l_∞ constraint $\varepsilon = 0.1$

Models	Train Acc	Test Acc	Adv-Train Acc	Adv-Test Acc	Infer Acc	Precision	Recall
CIFAR10 (natural)	92.80%	85.15%	12.89%	12.63%	**54.37%**	52.67%	86.00%
CIFAR10 (robust)	68.57%	66.33%	61.25%	58.43%	**51.11%**	50.78%	72.00%
SVHN (natural)	98.86%	84.01%	20.38%	16.64%	**57.85%**	54.45%	96.00%
SVHN (robust)	82.06%	79.62%	68.55%	66.15%	**51.00%**	51.27%	40.00%

Summary

- **Combine both security and privacy domains** for machine learning by measuring membership information leakage of adversarially robust deep learning models.
  - Adversarial Training
    - More susceptible to membership inference attacks.
    - Privacy leakage related to model’s robustness performance.
  - Provable Defense
    - No increase of vulnerability to membership inference attacks, with a significant drop in the model’s predictive power.
- **Think about security and privacy together**.