Privacy vs Robustness (against Adversarial Examples) in Machine Learning *

Liwei Song
Princeton University
liweis@princeton.edu

Reza Shokri
National University of Singapore
reza@comp.nus.edu.sg

Prateek Mittal
Princeton University
pmittal@princeton.edu

Abstract

Research into challenges of machine learning models typically considers the security domain and the privacy domain separately. It is thus unclear whether the defenses in one domain will have any unexpected impact on the other domain. In this paper, we combine two domains together by investigating the interplay between membership inference and robustness against adversarial examples in machine learning. By performing membership inference attacks against both robust models and natural (undefended) models, we find that the adversarial defense methods, although increase the model robustness against adversarial examples, also make the model more vulnerable to membership inference attacks, indicating a potential conflict between privacy and robustness in machine learning.

1 Introduction

The security and privacy vulnerabilities of machine learning models have come to a forefront in recent years, together with the arms race between attacks and defenses [3, 7, 13]. From the security perspective, the adversary aims to induce misclassifications to the target model with either test-time evasion attacks (also known as adversarial examples) [1, 17, 6] or training-time poisoning attacks [2, 9]. From the privacy perspective, the adversary aims to infer private information about target model's training data [14, 4] or the target model itself [18, 19]. The research community has proposed defenses to resolve both security issues [11, 16] and privacy issues [12, 8]. However, these defense approaches typically focus solely on either the security domain or the privacy domain, and it is unclear whether defense methods in one domain will have any unexpected impact on the other domain.

In this paper, we take a step towards enhancing our understanding of machine learning models when both the security and privacy domains are combined. In particular, we investigate the interplay between privacy and adversarial robustness in machine learning by measuring the success of membership inference attacks on defense methods that mitigate the threat of adversarial examples.

Membership inference attacks determine whether a data point is from the target model's training set or not [14, 22]. Adversarial defense methods enhance model robustness against adversarial examples by ensuring that model predictions remain unchanged for a small area around each input [11, 20, 24]. However, this objective is optimized on training set, increasing each training point's influence on the model. This makes the robust model more susceptible to membership inference attacks.

We perform membership inference attacks against robust models trained with one of state-of-the-art adversarial defenses: adversarial training proposed by Madry et al. [11]. Our experiment results show that the robust models indeed leak more membership information, compared to natural models. We can further enhance membership inference attacks by exploiting the structural properties of robust models on adversarially perturbed data. We refer interested readers to the full version of this paper [15] for membership inference results with other adversarial defense methods.

*The full version of this paper [15] is accepted by ACM CCS 2019.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

2 Background

2.1 Adversarial examples and defenses

For a classification task with the training set D_{train} over pairs of inputs $\mathbf{x}$ and labels y, the natural training algorithm learns a model F_θ by minimizing the prediction loss ℓ over all training examples

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

where $|\cdot|$ measures the size of a dataset.

However, machine learning models can be easily fooled by adversary examples [1, 17, 6], which induce model misclassifications via the addition of imperceptible perturbations to benign inputs

$$\operatorname*{argmax}_i F_\theta(\tilde{\mathbf{x}})_i \neq y, \quad \text{such that } \tilde{\mathbf{x}} \in \mathcal{B}_\epsilon(\mathbf{x}), \tag{2}$$

where $\mathcal{B}_\epsilon(\mathbf{x})$ denotes the set of points around $\mathbf{x}$ within the perturbation budget of ϵ. The solution to Equation (2) is called an "untargeted adversarial example" as the adversarial goal is to achieve any misclassification. In comparison, a "targeted adversarial example" ensures that the model prediction is a specified incorrect label y', which is not equal to y.

$$\operatorname*{argmax}_i F_\theta(\tilde{\mathbf{x}})_i = y', \quad \text{such that } \tilde{\mathbf{x}} \in \mathcal{B}_\epsilon(\mathbf{x}). \tag{3}$$

To provide robustness against adversarial examples, a robust training algorithm is adopted to train the model by taking the adversarial attack into consideration [11, 20, 24]

$$\min_{\theta} \frac{1}{|D_{\text{train}}|} \sum_{(\mathbf{x},y)\in D_{\text{train}}} \alpha \cdot \ell(F_\theta(\mathbf{x}), y) + (1-\alpha) \cdot \max_{\tilde{\mathbf{x}}\in\mathcal{B}_\epsilon(x)} \ell(F_\theta(\tilde{\mathbf{x}}), y). \tag{4}$$

However, it is usually hard to find the exact solution to the inner maximization problem. Therefore, the adversarial defense methods propose different ways to approximate the robust loss. In particular, Madry et al. [11] propose one of the most effective defense methods by training purely on adversarial examples ($\alpha = 0$) generated from a multi-step projected gradient descent (PGD) method

$$\tilde{\mathbf{x}}^{t+1} = \Pi_{\mathcal{B}_\epsilon(\mathbf{x})}[\tilde{\mathbf{x}}^t + \eta \cdot \text{sign}(\nabla_{\tilde{\mathbf{x}}^t} \ell(F_\theta(\tilde{\mathbf{x}}^t), y))], \tag{5}$$

where η is the step size, ∇ denotes the gradient computation, and $\Pi_{\mathcal{B}_\epsilon(\mathbf{x})}$ means the projection onto the perturbation constraint.

2.2 Membership inference

Shokri et al. [14] design a membership inference attack method based on training an inference model to distinguish between predictions on training set members versus non-members. To train the inference model, they introduce the shadow training technique: (1) the adversary first trains multiple "shadow models" which simulate the behavior of the target model, (2) based on the shadow models' outputs on their own training and test examples, the adversary obtains a labeled (member vs non-member) dataset, and (3) finally trains the inference model as a neural network to perform membership inference attack against the target model.

A simpler inference model, such as a linear classifier, can also distinguish significantly vulnerable members from non-members. Yeom et al. [22] suggest comparing the prediction confidence value of a target example with a threshold. Large confidence indicates membership. Their results show that such a simple confidence-thresholding method is reasonably effective and achieves membership inference accuracy close to that of a complex neural network classifier learned from shadow training.

3 Membership Inference Attacks against Robust Models

3.1 Membership inference performance

For a machine learning model F (we skip its parameters θ for simplicity) robustly trained with the perturbation constraint $\mathcal{B}_\epsilon$, the membership inference attacks aim to determine whether a given input

$(\mathbf{x}, y)$ is in its training set D_{train} or not. We use the inference accuracy to evaluate the success of membership inference attacks, and sample an input $(\mathbf{x}, y)$ from either training set D_{train} or test set D_{test} with an equal 50% probability. Thus a random guessing strategy will lead to a 50% inference accuracy. We denote the inference strategy as $\mathcal{I}(F, \mathcal{B}_\epsilon, (\mathbf{x}, y))$, which codes members as 1, and non-members as 0. The membership inference accuracy can be expressed as

$$A_{inf}(F, \mathcal{B}_\epsilon, \mathcal{I}) = \frac{\sum_{(\mathbf{x},y)\in D_{\text{train}}} \mathcal{I}(F, \mathcal{B}_\epsilon, (\mathbf{x}, y))}{2 \cdot |D_{\text{train}}|} + \frac{\sum_{(\mathbf{x},y)\in D_{\text{test}}} 1 - \mathcal{I}(F, \mathcal{B}_\epsilon, (\mathbf{x}, y))}{2 \cdot |D_{\text{test}}|}. \quad (6)$$

3.2 Exploiting the model's predictions on benign examples

We adopt the confidence-thresholding method [22] due to its simplicity and effectiveness:

$$\mathcal{I}_{\text{B}}(F, \mathcal{B}_\epsilon, (\mathbf{x}, y)) = \mathbb{1}\{F(\mathbf{x})_y \geq \tau_{\text{B}}\}, \quad (7)$$

where $\mathbb{1}\{\cdot\}$ is the indicator function, an input $(\mathbf{x}, y)$ is inferred as member if its prediction confidence $F(\mathbf{x})_y$ is larger than (or equal to) a preset threshold τ_{B}. In this paper, we evaluate the worst case inference risks by choosing τ_{B} to achieve the highest inference accuracy (Equation (6)). In practice, an adversary can learn the threshold via the shadow training technique [14].

3.3 Exploiting the model's predictions on adversarial examples

We further leverage the structural properties of robust models to enhance membership inference attacks. Specifically, we use the PGD attack method (Equation (5)) to generate an untargeted adversarial example $\mathbf{x}_{adv}$ under $\mathcal{B}_\epsilon$, and use a threshold on the model's prediction confidence on $\mathbf{x}_{adv}$

$$\mathcal{I}_{\text{A}}(F, \mathcal{B}_\epsilon, (\mathbf{x}, y)) = \mathbb{1}\{F(\mathbf{x}_{adv})_y \geq \tau_{\text{A}}\}. \quad (8)$$

Similarly, we choose the preset threshold τ_{A} to achieve the highest inference accuracy.

3.3.1 Targeted adversarial examples

We extend the attack to exploiting targeted adversarial examples. Targeted adversarial examples contain information about distance of the benign input to each label's decision boundary, and are expected to leak more membership information. We adapt the PGD attack to generate targeted adversarial examples by iteratively minizing the targeted prediction loss.

$$\tilde{\mathbf{x}}^{t+1} = \Pi_{\mathcal{B}_\epsilon(\mathbf{x})}[\tilde{\mathbf{x}}^t - \eta \cdot \text{sign}(\nabla_{\tilde{\mathbf{x}}^t} \ell(F_\theta(\tilde{\mathbf{x}}^t), y'))]. \quad (9)$$

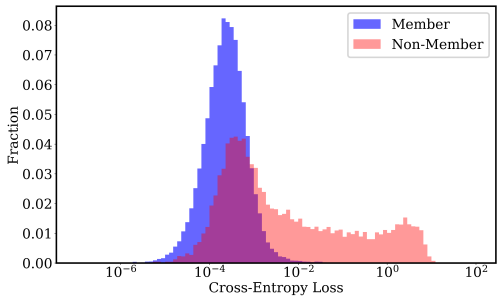
The confidence-thresholding inference strategy does not apply here because there exist $k-1$ targeted adversarial examples for each input. Instead, following Shokri et al. [14], we train neural network models for membership inferences. For each class label, we first choose a fraction of training and test points and generate corresponding targeted adversarial examples. Next, we compute model predictions on the targeted adversarial examples, and use them to train the membership inference classifier. Finally, we perform inference attacks using the remaining training and test points.

4 Experiment Results

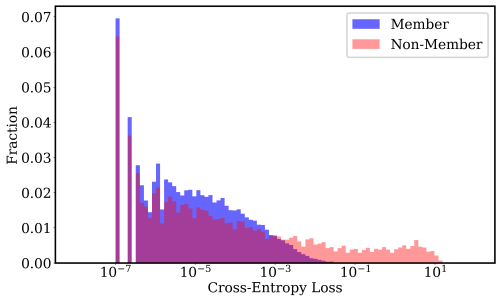
We follow the method of Madry et al. [11] to train robust classifiers with l_∞ perturbation constraints ($\mathcal{B}_\epsilon(\mathbf{x}) = \{\mathbf{x}' \mid \|\mathbf{x}' - \mathbf{x}\|_\infty \leq \epsilon\}$) on Yale Face dataset [5, 10], Fashion-MNIST dataset [21], and CIFAR10 dataset. Details about experiment setup can be found in the full version of this paper [15].

We present the membership inference results in Table 1 and visualize the divergence between the loss distribution over members and non-members on robust and natural CIFAR10 classifiers in Figure 1. We can see that compared to natural models, **robust models are more vulnerable to membership inference attacks with much higher inference accuracy**. We also notice that compared with benign inputs, **leveraging (untargeted) adversarial examples increases the inference accuracy**.

We use the robust CIFAR10 classifier [11] as an example to show the gain of using targeted adversarial examples for membership inferences. For each class label, we learn a dedicated inference model (a 3-layer MLP) with predictions of targeted adversarial examples generated from 500 training points



(a) Robust CIFAR10 classifier from Madry et al. [11], with 99% train accuracy and 87% test accuracy.



(b) Natural CIFAR10 classifier, with 100% train accuracy and 95% test accuracy.

Figure 1: Histogram of models' loss values of training data (members) and test data (non-members).

Table 1: Membership inference attacks against natural and robust models [11].

dataset	Target Models: model architecture	training method	ϵ	Accuracy Performance: train accuracy	test accuracy	adv-train accuracy	adv-test accuracy	Membership Inference: inference accuracy ($\mathcal{I}_{\mathrm{B}}$)	inference accuracy ($\mathcal{I}_{\mathrm{A}}$)
Yale Face	10-layer CNN	natural	N.A.	100%	98.25%	4.53%	2.92%	**55.85%**	54.27%
Yale Face	10-layer CNN	robust [11]	8/255	99.89%	96.69%	99.00%	77.63%	61.69%	**68.83%**
Fashion MNIST	8-layer CNN	natural	N.A.	100%	92.18%	4.35%	4.14%	**57.12%**	50.95%
Fashion MNIST	8-layer CNN	robust [11]	0.1	99.93%	90.88%	96.91%	68.06%	58.32%	**64.49%**
CIFAR10	Wide ResNet [23]	natural	N.A.	100%	95.01%	0%	0%	**57.43%**	50.86%
CIFAR10	wide ResNet [23]	robust [11]	8/255	99.99%	87.25%	96.08%	46.61%	74.89%	**75.67%**

and 500 test points, then test the inference model on remaining training and test points. We call this method "model (targeted)". Similarly, we obtain inference models with the same architecture by using either untargeted adversarial examples' predictions or benign examples' predictions.We call these methods "model (untargeted)" and "model (benign)". Finally, we adapt $\mathcal{I}_{\mathrm{B}}$ and $\mathcal{I}_{\mathrm{A}}$ to be class dependent by choosing the threshold according to confidence values from 500 training points and 500 test points. we call the adapted methods "confidence (benign)" and "confidence (untargeted)".

We present the membership inference results with different approaches in Table 2. We can see that **targeted adversarial example based inference strategy "model (targeted)" always has the highest inference accuracy.**

Table 2: Membership inference attacks against robust CIFAR10 classifier [11].

inference method	**class 0**	**class 1**	**class 2**	**class 3**	**class 4**	**class 5**	**class 6**	**class 7**	**class 8**	**class 9**
confidence (benign)	70.88%	63.57%	80.16%	90.43%	82.30%	81.34%	75.34%	69.54%	69.16%	68.13%
model (benign)	71.49%	64.42%	76.74%	90.49%	82.17%	79.84%	70.92%	67.61%	69.57%	66.34%
confidence (untargeted)	72.21%	67.52%	79.71%	87.64%	81.83%	81.57%	77.66%	72.92%	74.36%	71.86%
model (untargeted)	72.70%	67.69%	80.16%	87.83%	81.57%	81.34%	76.97%	72.82%	74.40%	72.06%
model (targeted)	**74.42%**	**68.88%**	**83.58%**	**90.57%**	**84.47%**	**83.02%**	**79.94%**	**72.98%**	**75.33%**	**73.32%**

5 Conclusion

In this paper, we investigate the membership inference privacy risk of defense approaches that mitigate the threat of adversarial examples. Our results indicate a potential conflict between privacy and adversarial robustness, and highlight the importance of thinking about security and privacy together.

References

[1] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, pages 387–402, 2013.

[2] Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning attacks against support vector machines. In *International Conference on Machine Learning (ICML)*, pages 1467–1474, 2012.

[3] Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.

[4] Karan Ganju, Qi Wang, Wei Yang, Carl A Gunter, and Nikita Borisov. Property inference attacks on fully connected neural networks using permutation invariant representations. In *ACM Conference on Computer and Communications Security (CCS)*, pages 619–633, 2018.

[5] Athinodoros S Georghiades, Peter N Belhumeur, and David J Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):643–660, 2001.

[6] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

[7] Ling Huang, Anthony D Joseph, Blaine Nelson, Benjamin IP Rubinstein, and JD Tygar. Adversarial machine learning. In *ACM Workshop on Artificial Intelligence and Security (AISec)*, pages 43–58, 2011.

[8] Manish Kesarwani, Bhaskar Mukhoty, Vijay Arya, and Sameep Mehta. Model extraction warning in mlaas paradigm. In *Proceedings of the 34th Annual Computer Security Applications Conference (ACSAC)*, pages 371–380. ACM, 2018.

[9] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning (ICML)*, pages 1885–1894, 2017.

[10] Kuang-Chih Lee, Jeffrey Ho, and David J Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (5):684–698, 2005.

[11] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018.

[12] Milad Nasr, Reza Shokri, and Amir Houmansadr. Machine learning with membership privacy using adversarial regularization. In *ACM Conference on Computer and Communications Security (CCS)*, 2018.

[13] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael Wellman. Sok: Security and privacy in machine learning. In *IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 399–414, 2018.

[14] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy (S&P)*, pages 3–18, 2017.

[15] Liwei Song, Reza Shokri, and Prateek Mittal. Privacy risks of securing machine learning models against adversarial examples. In *ACM Conference on Computer and Communications Security (CCS)*, 2019.

[16] Jacob Steinhardt, Pang Wei W Koh, and Percy S Liang. Certified defenses for data poisoning attacks. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 3517–3529, 2017.

[17] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

[18] Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction apis. In *USENIX Security Symposium*, pages 601–618, 2016.

[19] Binghui Wang and Neil Zhenqiang Gong. Stealing hyperparameters in machine learning. In *IEEE Symposium on Security and Privacy (S&P)*, 2018.

[20] Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning (ICML)*, pages 5283–5292, 2018.

[21] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[22] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *IEEE Computer Security Foundations Symposium (CSF)*, pages 268–282, 2018.

[23] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2016.

[24] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning (ICML)*, 2019.