2019 IEEE Security and Privacy Workshops (SPW)

Membership Inference Attacks against Adversarially Robust Deep Learning Models

Liwei Song
liweis@princeton.edu
Princeton University

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

Prateek Mittal
pmittal@princeton.edu
Princeton University

***Abstract*—In recent years, the research community has increasingly focused on understanding the security and privacy challenges posed by deep learning models. However, the security domain and the privacy domain have typically been considered *separately*. It is thus unclear whether the 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 deep learning models when the two domains are combined together. We do this by *measuring the success of membership inference attacks against two state-of-the-art adversarial defense methods that mitigate evasion attacks: adversarial training and provable defense*. On the one hand, membership inference attacks aim to infer an individual's participation in the target model's training dataset and are known to be correlated with target model's overfitting. On the other hand, adversarial defense methods aim to enhance the robustness of target models by ensuring that model predictions are unchanged for a small area around each sample in the training dataset. Intuitively, adversarial defenses may rely more on the training dataset and be more vulnerable to membership inference attacks. By performing empirical membership inference attacks on both adversarially robust models and corresponding undefended models, we find that the adversarial training method is indeed more susceptible to membership inference attacks, and the privacy leakage is directly correlated with model robustness. We also find that the provable defense approach does not lead to enhanced success of membership inference attacks. However, this is achieved by significantly sacrificing the accuracy of the model on benign data points, indicating that privacy, security, and prediction accuracy are not jointly achieved in these two approaches.**

I. Introduction

The security and privacy issues of deep learning models have come to a forefront in recent years, as these models were not originally designed to be robust in adversarial settings [1].

From the security perspective, an adversary's objective is to cause the target machine learning model to misbehave. Existing attack methods can be divided into two categories: poisoning attacks and evasion attacks [2]. Poisoning attacks manipulate part of training data to compromise the trained deep learning model [3]–[5]. Evasion attacks, also called adversarial examples, find vulnerabilities in deep learning models trained on benign data and directly perturb test inputs to induce misclassifications [6]–[10].

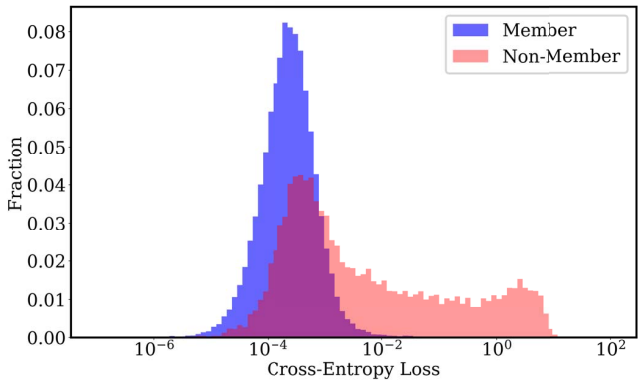
From the privacy perspective, an adversary's objective is to infer private information about the target model itself or its training data. Well-known privacy issues include *membership inference* to infer whether an input is part of the model's training dataset [11]–[13]; property inference to infer global properties of training dataset, such as the fraction of a certain class [14]; model inversion to reconstruct the model's input from model predictions [15]; and training data memorization by adversarially modifying the training algorithm to memorize sensitive training data information [16].

Along with finding novel attacks again deep learning models, the research community has also proposed defense approaches to resolve both security issues [17]–[21] and privacy issues [22]–[25]. 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 some unexpected impact on the other domain.

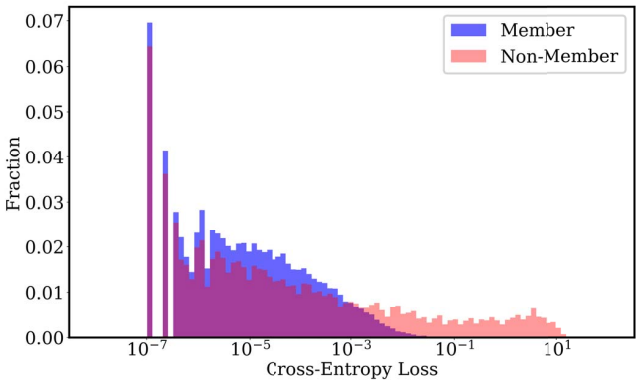
In this paper, we take a first step towards enhancing our understanding of deep learning models when both the security and privacy domains combined. In particular, we seek to understand the impact of robust machine learning algorithms on the privacy of sensitive training data. Specifically, we evaluate *membership inference attacks against adversarially robust deep learning models*, which aim to mitigate the threat of adversarial examples. The membership inference attack aims to infer whether an input to the deep learning model is part of its training dataset or not. The success of membership inference attacks, in the black-box setting, is shown to be highly related to the target model's overfitting [11], [12]. Adversarially robust models [17]–[19] aim to enhance the robustness of target models by ensuring that model predictions are unchanged for a small area (such as l_∞ ball) around each (training) example. Intuitively, adversarially robust models magnify the influence of the training data on the model, resulting in an enhanced risk of membership inference attacks.

We measure the success of membership inference attacks against two state-of-the-art adversarial defense methods, *adversarial training* [17] and *provable defense* [18], [19]. Our experimental results show that compared to undefended naturally trained models, *adversarially trained models are indeed more vulnerable to membership inference attacks.* Moreover, *an increased robustness of the adversarially trained model (model trained with larger adversarial perturbations) is correlated with an increase in the success of the membership inference attack.* An example is shown in Fig. 1, where we plot the distributions of training examples' prediction cross-entropy loss values and test examples' loss values for both the adversarially trained CIFAR10 model and the naturally trained CIFAR10 model. It is clear that members (training examples)

DOI 10.1109/SPW.2019.00021



(a) Adversarially trained model [17], with 99% train accuracy and 87% test accuracy.



(b) Naturally trained model, with 100% train accuracy and 95% test accuracy. Around 23% training and test examples have zero loss.

Fig. 1: Histogram of CIFAR10 models' loss values of training data (members) and test data (non-members). The large divergence between the loss distribution over members and non-members increases the privacy risk of adversarially trained models.

and non-members (test examples) can be distinguished more easily for the adversarially trained model, compared to the naturally trained (undefended) model.

We also find that the provable defense approach [18], [19] does not significantly increase the models' vulnerability to membership inference attacks. However, this is achieved by significantly sacrificing the accuracy of the model on benign data points, indicating that privacy, security, and accuracy are not achieved in these two approaches.

II. Background and Related Work

A. Robustness against Adversarial Examples

For a standard classification task with the training dataset D_{train} over pairs of inputs $\boldsymbol{x}$ and corresponding labels y, the natural training algorithm tries to learn a model that minimizes the prediction loss over all training examples, which can be formulated as a minimization problem:

$$\min_{\theta} \frac{1}{|D_{train}|} \sum_{(\boldsymbol{x},y)\in D_{train}} \mathcal{L}(F_\theta(\boldsymbol{x}), y), \tag{1}$$

where $F_\theta(\cdot)$ is the prediction function of the learning model with parameters θ, and $\mathcal{L}$ is an appropriate loss function, such as the cross-entropy loss for neural networks.

Adversarial examples: Although deep learning models have achieved tremendous success in many classification scenarios, they have been found to be easily fooled by adversary examples [6]–[8], which induce misclassifications by the models via the addition of imperceptible perturbations to input examples. Corresponding to the learning algorithm shown in Equation (1), the generation of adversarial perturbation can be expressed as a maximization problem:

$$\max_{\boldsymbol{\delta}\in\boldsymbol{\Delta}} \mathcal{L}(F_\theta(\boldsymbol{x}+\boldsymbol{\delta}), y), \tag{2}$$

where $\boldsymbol{\Delta}$ represents the constraint of allowed adversarial perturbation, such as the l_∞-ball within a small distance value ($||\boldsymbol{\Delta}||_\infty \leq \epsilon$) in the image classification task [8], [10].

To defend against adversarial examples, a robust training algorithm can be formulated as a min-max optimization problem by taking the adversarial attack into consideration [17]–[19]:

$$\min_{\theta} \frac{1}{|D_{train}|} \sum_{(\boldsymbol{x},y)\in D_{train}} \max_{\boldsymbol{\delta}\in\boldsymbol{\Delta}} \mathcal{L}(F_\theta(\boldsymbol{x}+\boldsymbol{\delta}), y). \tag{3}$$

However, it is usually hard to find the global maximum of the inner maximization problem for deep neural networks due to the highly non-concave function with many local maxima [17]. Adversarial training [17] and provable defense algorithm [18], [19] try to solve Equation (3) in different ways.

Adversarial training: Madry et al. [17] adopt the adversarial training method to train robust models by approximating the inner maximization problem in Equation (3) via the adversarial perturbations generated from a multi-step projected gradient descent (PGD) attack method, i.e.,

$$\boldsymbol{\delta}^{t+1} = \Pi_{\boldsymbol{\Delta}}(\boldsymbol{\delta}^t + \alpha \ \text{sign}(\nabla_{\boldsymbol{x}+\boldsymbol{\delta}^t}\mathcal{L}(F_\theta(\boldsymbol{x}+\boldsymbol{\delta}^t), y))), \tag{4}$$

where α is the value of step size, ∇ denotes the gradient computation, and $\Pi_{\boldsymbol{\Delta}}$ means the projection onto the perturbation constraint. Compared to the other defense algorithms, the obtained models from adversarial training have been shown to be the most robust models against the majority of the adversarial attack methods [17], [26].

Provable defense: Different from the empirical defense strategy of adversarial training [17], Wong et al. [18], [19] propose a provable defense method for robust training by finding an upper bound of the inner maximization problem in Equation (3) via its relaxed dual problem. Due to the space limit, we refer interested readers to Wong et al. [18], [19] for more details. The approach computes an upper bound of the loss value in the adversarial setting, yielding a quantification of the robust error bound for the defended model. So, it bounds the fraction of input examples that can be adversarially perturbed under the predefined perturbation constraint. One shortcoming of this defense method is that the trained model usually

has much reduced accuracy performance on *benign data*: the provably defended CIFAR10 model with the l_∞ perturbation budget of 8/255 has 29% test accuracy [19], compared with 87% test accuracy for the adversarially trained model [17].

B. Privacy against Membership Inference Attacks

For a target deep learning model, the membership inference attacks aim to determine whether a given data point was used to train the model or not [11]. The attack poses a serious privacy risk, as the participation of a sample in the training data can correspond to an individual's sensitive information, such as in the setting of health analytics [12].

In this paper, we focus on membership inference attacks in the black-box setting, where for a certain input, the adversary only has the knowledge of the target model's final output. We do not cover the analysis of white-box attacks, where the adversary has access to the target model's parameters [13].

Shokri et al. [11] design a membership inference attack method based on the *shadow training technique*: (1) an 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 binary labeled (member vs non-member) dataset, and (3) the adversary finally trains a neural network model using the labeled dataset to perform membership inference attack against the target model.

This method can be further simplified. Yeom et al. [12] suggest comparing the classification loss value of a target example with a preset threshold (equivalent to shadow models as a linear classifier of loss values). Small loss indicates membership. The experiment results show that the inference strategy of using a threshold on the prediction confidence is very effective and achieves membership inference accuracy close to that of the shadow training method. In this paper, we follow this simple approach, by using a linear classifier (using a threshold) as the inference attack.

C. Other Related Work

To the best of our knowledge, there is no previous work trying to analyze privacy issues for adversarially robust models. The closest work to our paper is Schmidt et al. [27], where the authors show that although the adversarially trained model generalizes well in the standard classification setting, it overfits in the adversarial setting: for adversarially perturbed input examples, the training accuracy is much larger than the test accuracy. However, as shown in Fig. 1a, we find that even without adversarial perturbations, the adversarially trained model is vulnerable to membership inference attacks based on predictions on benign/clean inputs.

III. Problem Statement

In this section, we provide a detailed description of our membership inference adversary and the metrics adopted to measure the privacy leakage.

A. Threat Model

In this paper, we consider membership inference attacks in the **black-box setting** [11]. Let $\mathcal{F}(\cdot)$ denote the classification function of the target model, where $\mathcal{F}_i(\cdot)$ means the prediction probability of class i with $\sum_i \mathcal{F}_i(\cdot) = 1$. For each labelled input ($\boldsymbol{x}$, y), the adversary only knows the final prediction vector $\mathcal{F}(\boldsymbol{x})$ and tries to guess whether the input is in the model's training dataset (member) or not (non-member).

For the adversary's membership inference strategy, we choose the threshold inference method (linear classifier) based on the classification's confidence value, which can be expressed as following.

$$\mathcal{I}(\mathcal{F}, (\boldsymbol{x}, y), \tau) = \begin{cases} \text{member}, & \text{if } \mathcal{F}_y(\boldsymbol{x}) \geq \tau; \\ \text{non-member}, & \text{if } \mathcal{F}_y(\boldsymbol{x}) < \tau, \end{cases} \tag{5}$$

where $\mathcal{I}(\cdot)$ represents the inference strategy and τ is a certain confidence threshold. The input example ($\boldsymbol{x}$, y) will be inferred as a member of the target model's training dataset if model's prediction confidence $\mathcal{F}_y(\boldsymbol{x})$ is larger than (or equal to) the threshold, and a non-member otherwise. In our experiments, we assume the threshold is an input hyperparameter to the attack, which could have been learned using the shadow training method in practice.

B. Metrics for Evaluating Membership Inference Attacks

For the evaluation, we sample the input example ($\boldsymbol{x}$, y) from either the target model's training dataset or test dataset with an equal 50% probability. We use the following metrics to evaluate our membership inference attacks against the target deep learning models.

Inference accuracy: The inference accuracy corresponds to the fraction of correct membership predictions made by the adversary. The random guessing strategy would result in a baseline accuracy value of 50%.

Precision: Precision is calculated as the fraction of examples inferred as members that are indeed members of the target model's training dataset. The baseline precision value with a random guessing strategy is also 50%.

Recall: Recall is calculated as the fraction of training examples that are inferred as members correctly. Given our threshold inference strategy, it corresponds to the probability that a training example has its prediction confidence value larger than (or equal to) the preset threshold.

Area under the precision-recall curve (AUPRC): We use different confidence threshold values to obtain the precision-recall curve and compute the total area under the curve. A larger AUPRC value corresponds to more leakage. The baseline AUPRC with a random guessing strategy is 0.5.

Kullback–Leibler (KL) divergence: The KL divergence value captures how the distribution of the prediction cross-entropy loss over training examples is different from that of test examples (see Fig. 1 for an illustration of these distributions). We compute the distribution of entropy loss, which is the negative logarithm value of prediction confidence. A larger divergence value means that it is easier to distinguish training data (member) and test data (non-member).

TABLE I: Membership inference attacks against adversarially trained models and corresponding naturally trained models. ϵ is the l_∞ perturbation budget used for robust training. 'adv-train accuracy' and 'adv-test accuracy' are computed with PGD attacks under the same ϵ constraint. 'adv-train' and 'nat-train' represent adversarial training [17] and natural training, respectively.

Target Models				Accuracy Performance				Membership Inference Adversary				
Dataset	**Architecture**	**Train Method**	ϵ	**Train Accuracy**	**Test Accuracy**	**Adv-Train Accuracy**	**Adv-Test Accuracy**	**Inference Accuracy**	**Precision**	**Recall**	**KL Divergence**	**AUPRC**
CIFAR10	WRN-34-10	adv-train [17]	8/255	99.99%	87.25%	96.07%	46.59%	74.86%	69.08%	90.00%	0.72	0.76
CIFAR10	WRN-34-10	nat-train	N.A.	100%	95.01%	0.00%	0.00%	57.37%	54.16%	96.00%	0.14	0.52
SVHN	WRN-34-4	adv-train [17]	4/255	99.99%	93.91%	99.74%	72.17%	64.30%	59.70%	88.00%	0.33	0.67
SVHN	WRN-34-4	nat-train	N.A.	99.99%	95.64%	6.53%	3.86%	56.79%	53.72%	98.00%	0.13	0.53

Note that a well-generalized machine learning model with no membership inference risk will have inference accuracy, precision, and AUPRC values all equal to 0.5, and the KL divergence equal to 0. Also note that the inference accuracy, precision and recall depend on the choice of confidence threshold, while AUPRC and KL divergence do not. *In our experiments, we set the value of confidence threshold to achieve the highest inference accuracy value.*

IV. Experiments Results

In this section, we measure the success of membership inference attacks against adversarially robust models. All experiments are performed on a GPU cluster with 8 NVIDIA P100 GPUs.

A. Target Deep Learning Models

We use the code released by Madry et al. [17] and Wong et al. [18], [19] to train the adversarially trained models[1] and provably defended models[2] on the CIFAR10 dataset and the SVHN dataset.

We train the adversarially robust models using l_∞ perturbation budget as an input parameter. We also train corresponding baseline models with the natural training method (undefended) for comparison of their privacy properties with robust training methods. The details about the model architectures and training parameters are provided below.

Adversarial training: The code released by Madry et al. [17] adopts a wide residual network (WRN) architecture [28] for the adversarially trained CIFAR10 model with a perturbation budget (ϵ) equal to 8/255. The WRN model contains 4 groups of residual layers with filter sizes (16, 160, 320, 640) and 5 residual blocks for each group. The architecture is named as WRN-34-10, following the notation of Zagoruyko et al. [28]. As suggested by Schmidt et al. [27], we use a similar WRN architecture with filter sizes (16, 64, 128, 256) to train the robust SVHN model (WRN-34-4) with the ϵ value of 4/255.

Provable defense: The code released by Wong et al. [18], [19] provides different model architectures for CIFAR10 and SVHN datasets. For the CIFAR10 dataset, the residual network is also adopted but with a narrower architecture due to scalability issues – 4 groups of residual layers with filter sizes (16, 16, 32, 64) and just one residual block for each group (WRN-10-1). Two defended CIFAR10 models with different ϵ values (2/255 and 8/255) are provided; we report on the model with the smaller ϵ, as the other model has the classification error higher than 70%. For the SVHN dataset, a simple 4-layer convolution neural network (CNN) architecture is adopted with the adversarial perturbation value equal to $2.55/255$.

[1]https://github.com/MadryLab/cifar10_challenge
[2]https://github.com/locuslab/convex_adversarial

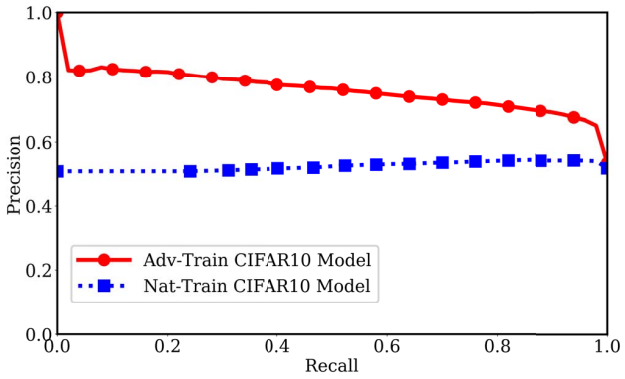


Fig. 2: The precision-recall curve of the membership inference attack against adversarially trained and naturally trained CIFAR10 models. This is obtained by varying the confidence threshold value τ in Equation (5).

B. Membership Inferences against Adversarial Training

According to Table I, the adversary can have a membership inference accuracy of 74.86% on the adversarially trained CIFAR10 model, compared to the inference accuracy of 57.37% on the naturally trained CIFAR10 model. Similarly adversarial training for the SVHN model also causes the membership inference accuracy to increase from 56.79% to 64.30%. Combined with Fig. 1, we can find that for the adversarially trained CIFAR10 model, the loss distribution of training dataset differs greatly from that of the test dataset with the KL divergence value of 0.72. While the two distributions for the naturally trained CIFAR10 model are quite close with the KL divergence value of 0.14. Thus, **adversarial training increases the information leakage about the training data**, as it makes members and non-members more distinguishable. This is true for all different attack threshold values, as presented in the precision-recall curve in Fig. 2.

Adversarially trained models generalize well (and have high test accuracy) in the standard benign setting. However, as

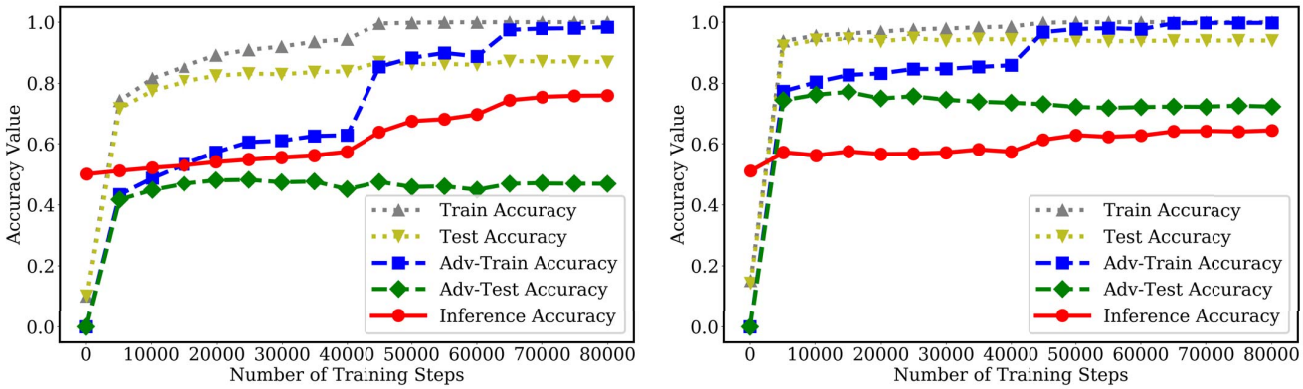


Fig. 3: Accuracy performance and privacy leakage over different training steps for the adversarially trained CIFAR10 (left) and SVHN (right) models. We can see that for adversarially trained models, although the train and test accuracy do not diverge much during the training process, the adv-train accuracy and the adv-test accuracy have a larger gap, which correlates with a higher membership inference accuracy.

shown in Table I, they fail to generalize in the adversarial setting (and have much reduced adv-test accuracy values) where we perturb the input examples using the PGD attack method in Equation (4). For example, by applying PGD attacks on the adversarially trained CIFAR10 model, the adv-train accuracy is 96.07%, however, the adv-test accuracy is reduced to 46.59%. Thus, **the membership information leakage seems to be directly related to the generalization of the robust training algorithm.**

To better understand the relation between privacy leakage and the generalization of robustness in the defended CIFAR10 and SVHN models, we compute their benign accuracy and adversarial accuracy performance along with membership inference accuracy over different training steps, plotted in Fig. 3. We can clearly see that the privacy leakage has a strong correlation with the (lack of) generalization of adversarial robustness : *as the number of training steps increases, (1) the train accuracy and the test accuracy values are close to each other, while (2) the adv-train accuracy and the adv-test accuracy have a larger divergence, and (3) the membership inference accuracy is correspondingly higher.*

Table II illustrates the membership inference attack results for varying perturbation budgets during adversarial training. We obtain three adversarially trained models with ϵ equals to $2/255$, $4/255$, $8/255$ for both CIFAR10 and SVHN datasets.

Recall that a model trained with a larger ϵ value is more robust since it can defend against larger adversarial perturbations. **We find that more robust models leak more information about the training data.** With a larger ϵ value, the adversarially trained model relies on a larger l_∞ ball around each training point and will overfit more, leading to a higher membership inference attack accuracy.

C. Membership Inferences against Provable Defense

The results for provably defended models are provided in Table III. We can see that both provably defended models and naturally trained models leak negligible information about training data membership. In fact, the provably defended models have inference accuracy values closer to 50%, at the cost of much reduced standard training and test accuracy performance. For example, the provably trained CIFAR10 model with $\epsilon = 2/255$ has both train and test accuracy values around 67%, but the naturally trained CIFAR10 model has standard accuracy values higher than 85%. We also find that as opposed to the adversarially trained models, the provably defended models have similar adv-train and adv-test accuracy values under the PGD attacks, which may explain why the provable defense method does not incur more privacy leakage.

TABLE II: Membership inference attacks against adversarially trained models with different robustness budgets.

Dataset	Perturbation Budget	Inference Accuracy	AUPRC
CIFAR10	$2/255$	64.40%	0.62
CIFAR10	$4/255$	69.34%	0.69
CIFAR10	$8/255$	74.86%	0.76
SVHN	$2/255$	60.69%	0.61
SVHN	$4/255$	64.30%	0.67
SVHN	$8/255$	68.09%	0.70

Furthermore, for the provably defended CIFAR10 model, we also measure its standard and adversarial accuracy performance along with membership inference accuracy over different training steps. As shown in Fig. 4, as the number of training steps increases, (1) the training accuracy and the test accuracy curves in both standard and adversarial settings do not diverge much, and (2) the adversary cannot achieve a high inference accuracy. We note that this property comes at the cost of relatively low benign accuracy values.

Thus, **the provable defense method does not increase the vulnerability of robust models to membership inference attacks, in the black-box adversary setting. However, this**

TABLE III: Membership inference attacks against provably defended models and corresponding naturally trained models. ϵ is the l_∞ perturbation budget used for robust training. 'adv-train accuracy' and 'adv-test accuracy' are computed with PGD attacks under the same ϵ constraint. 'pro-train' and 'nat-train' denote provable training [18], [19] and natural training, respectively.

Target Models				Accuracy Performance				Membership Inference Adversary				
Dataset	Architecture	Train Method	ϵ	Train Accuracy	Test Accuracy	Adv-Train Accuracy	Adv-Test Accuracy	Inference Accuracy	Precision	Recall	KL Divergence	AUPRC
CIFAR10	WRN-10-1	pro-train [19]	2/255	68.57%	66.33%	61.25%	58.43%	51.11%	50.78%	72.00%	0.01	0.51
CIFAR10	WRN-10-1	nat-train	N.A.	92.80%	85.15%	12.89%	12.63%	54.37%	52.67%	86.00%	0.04	0.51
SVHN	4-layer CNN	pro-train [18]	2.55/255	82.06%	79.62%	68.55%	66.15%	51.00%	51.27%	40.00%	0.01	0.51
SVHN	4-layer CNN	nat-train	N.A.	98.86%	84.01%	20.38%	16.64%	57.85%	54.45%	96.00%	0.15	0.54

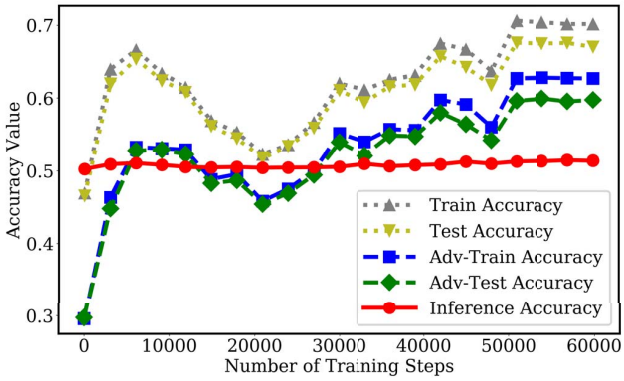


Fig. 4: Accuracy performance and privacy leakage over different training steps for the provably defended CIFAR10 model. We can see that the gap between train accuracy and test accuracy is small in both benign and adversarial settings, leading to low membership inference accuracy (close to random guess accuracy of 50%).

comes at the cost of a significant drop in model's accuracy (for benign data).

V. Discussion

Rethinking generalization from the privacy perspective: For machine learning, generalization means the ability of the learned model to fit on unseen instances. Usually the gap between train and test accuracy is used to show the generalization performance, which is not sufficient from the privacy perspective. As shown in Fig. 1a, even with a high test accuracy value, the target model may still leak its membership information through the prediction loss. To guarantee privacy, the model should avoid any differences between the output performance of training examples and that of test examples. Nasr et al. [24] and Hayes et al. [25] design privacy mechanisms to minimize the difference between the model's prediction distribution over training and test data.

Membership inference attacks in the white-box setting: Recent work has considered membership inference attacks in the white-box setting. Nasr et al. [13] show that simply adding all hidden layers' outputs as additional features does not help to enhance the membership inference accuracy. Instead, they find that the gradients with regard to each layer's parameters can increase the membership inference performance in the white-box setting. Measuring white-box membership inference risks would be an interesting direction for the future work.

VI. Conclusion

Security and privacy are two important domains of computer systems. In this paper, we have connected both domains together for deep learning systems by asking the following problem: *are adversarially robust models more vulnerable to membership inference attacks compared to undefended models?* Using experimental evaluation of black-box membership inference attacks, we find that: (1) the adversarially trained model is more susceptible to membership inference attacks, and the privacy leakage is correlated with the target model's robustness and generalization performance. (2) The provable defense method does not increase the target model's vulnerability to membership inference attacks, yet at the cost of a significant drop in the model's predictive power.

VII. Acknowledgments

This work was supported in part by the National Science Foundation under grants CNS-1553437, CNS-1704105, CIF-1617286 and EARS-1642962, by the Office of Naval Research Young Investigator Award, by the Army Research Office Young Investigator Prize, by Faculty research awards from Intel and IBM, and by the National Research Foundation, Prime Ministers Office, Singapore under its Strategic Capability Research Centres Funding Initiative.

References

[1] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognition*, vol. 84, pp. 317–331, 2018.

[2] L. Huang, A. D. Joseph, B. Nelson, B. I. Rubinstein, and J. Tygar, "Adversarial machine learning," in *ACM Workshop on Artificial Intelligence and Security (AISec)*, 2011, pp. 43–58.

[3] B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," in *International Conference on Machine Learning (ICML)*, 2012, pp. 1467–1474.

[4] Y. Liu, S. Ma, Y. Aafer, W.-C. Lee, J. Zhai, W. Wang, and X. Zhang, "Trojaning attack on neural networks," in *Network and Distributed Systems Security (NDSS) Symposium*, 2018.

[5] A. Shafahi, W. R. Huang, M. Najibi, O. Suciu, C. Studer, T. Dumitras, and T. Goldstein, "Poison frogs! targeted clean-label poisoning attacks on neural networks," in *Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

[6] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. 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)*, 2013, pp. 387–402.

[7] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *International Conference on Learning Representations (ICLR)*, 2014.

[8] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations (ICLR)*, 2015.

[9] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *ACM Asia Conference on Computer and Communications Security (AsiaCCS)*, 2017, pp. 506–519.

[10] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *IEEE Symposium on Security and Privacy (S&P)*, 2017, pp. 39–57.

[11] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *IEEE Symposium on Security and Privacy (S&P)*, 2017, pp. 3–18.

[12] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," in *IEEE Computer Security Foundations Symposium (CSF)*, 2018, pp. 268–282.

[13] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in *IEEE Symposium on Security and Privacy (S&P)*, 2019.

[14] K. Ganju, Q. Wang, W. Yang, C. A. Gunter, and N. Borisov, "Property inference attacks on fully connected neural networks using permutation invariant representations," in *ACM Conference on Computer and Communications Security (CCS)*, 2018, pp. 619–633.

[15] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *ACM Conference on Computer and Communications Security (CCS)*, 2015, pp. 1322–1333.

[16] C. Song, T. Ristenpart, and V. Shmatikov, "Machine learning models that remember too much," in *ACM Conference on Computer and Communications Security (CCS)*, 2017, pp. 587–601.

[17] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations (ICLR)*, 2018.

[18] E. Wong and Z. Kolter, "Provable defenses against adversarial examples via the convex outer adversarial polytope," in *International Conference on Machine Learning (ICML)*, 2018, pp. 5283–5292.

[19] E. Wong, F. Schmidt, J. H. Metzen, and J. Z. Kolter, "Scaling provable adversarial defenses," in *Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

[20] J. Steinhardt, P. W. W. Koh, and P. S. Liang, "Certified defenses for data poisoning attacks," in *Conference on Neural Information Processing Systems (NeurIPS)*, 2017, pp. 3517–3529.

[21] M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru, and B. Li, "Manipulating machine learning: Poisoning attacks and countermeasures for regression learning," in *IEEE Symposium on Security and Privacy (S&P)*, 2018.

[22] R. Shokri and V. Shmatikov, "Privacy-preserving deep learning," in *ACM Conference on Computer and Communications Security (CCS)*, 2015, pp. 1310–1321.

[23] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *ACM Conference on Computer and Communications Security (CCS)*, 2016, pp. 308–318.

[24] M. Nasr, R. Shokri, and A. Houmansadr, "Machine learning with membership privacy using adversarial regularization," in *ACM Conference on Computer and Communications Security (CCS)*, 2018.

[25] J. Hayes and O. Ohrimenko, "Contamination attacks and mitigation in multi-party machine learning," in *Conference on Neural Information Processing Systems (NeurIPS)*, 2018, pp. 6602–6614.

[26] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *International Conference on Machine Learning (ICML)*, 2018.

[27] L. Schmidt, S. Santurkar, D. Tsipras, K. Talwar, and A. Madry, "Adversarially robust generalization requires more data," in *Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

[28] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2016.