Leveraging Adversarial Examples to Quantify Membership Information Leakage

Ganesh Del Grosso [1,*] Hamid Jalalzai [1,*] Georg Pichler [2,*] Catuscia Palamidessi [1] Pablo Piantanida [3]
[1] Inria - Laboratoire d'Informatique de l'École polytechnique, France
[2] TU Wien, Austria
[3] International Laboratory on Learning Systems (ILLS)
McGill - ETS - MILA - CNRS - Université Paris-Saclay - CentraleSupélec, Canada
[1]{ganesh.del-grosso-guzman, hamid.jalalzai, catuscia.palamidessi}@inria.fr
[2]georg.pichler@ieee.org [3]piantani@mila.quebec

Abstract

The use of personal data for training machine learning systems comes with a privacy threat and measuring the level of privacy of a model is one of the major challenges in machine learning today. Identifying training data based on a trained model is a standard way of measuring the privacy risks induced by the model. We develop a novel approach to address the problem of membership inference in pattern recognition models, relying on information provided by adversarial examples. The strategy we propose consists of measuring the magnitude of a perturbation necessary to build an adversarial example. Indeed, we argue that this quantity reflects the likelihood of belonging to the training data. Extensive numerical experiments on multivariate data and an array of state-of-the-art target models show that our method performs comparable or even outperforms state-of-the-art strategies, but without requiring any additional training samples.

1. Introduction

With the deluge of data and increase of computational power within the last decades, performance of modern machine learning shows dramatic improvement in a wide range of applications such as computer vision [15, 23, 42] and natural language processing [6, 13, 62]. Along with this improvements new methods emerge, leading to remarkable change in societal applications ranging from industry [7, 25] to modern medicine [24, 35, 59] to art [16, 22], all of which may be considered *sensitive* domains, given the nature of the data.

While the benefits of machine learning are set upfront, other societal aspects such as fairness [38, 57] or safety [5] should not be trampled on [4]. It is common consensus that models require vast amounts of training data [2, 12] to reach state-of-the-art performance; meanwhile they do not necessarily guarantee the anonymity of the data provider [20]. This represents a serious privacy issue and controlling such leakage of information with modern regulation presents a new challenge [18, 37]. With recent data protection regulations [1, 43] personal data are required to be protected while being used by machine learning models [54]. To improve safety in machine learning, the study of attack strategies that exploit models in order to infer training data, or even corrupt them, has become an active area of research.

In this paper we investigate membership inference attacks (MIAs) [32–34, 46, 48–51, 55, 63], in which an attacker tries to determine whether or not a sample was part of the training set of a target model [49]. By leveraging adversarial attacks, we propose an MIA strategy that achieves similar performance to the state-of-the-art, but without using training samples to construct the attack. We only require accessing the target model and the testing sample.

Adversarial attacks maximize the loss function of a model with respect to an input sample, in order to find a perturbation that changes the class predicted by the model. Interestingly, we empirically observed that changing the predicted class requires a larger perturbation for samples that are part of the training set since the model was tuned to minimize the empirical loss function computed using these samples. Hence the idea is to measure this perturbation, i.e., the distance between an adversarial example and its original counterpart, and test whether it is lager than a certain threshold. We call this measure *Adversarial Distance*. Figure 1 provides an overview of our strategy[1].

*Equal contribution.

[1]The illustrations for the pipeline's input and adversarial noise are provided by [17]. The noise illustrated in Figure 1 is obtained with a fast

In contrast to other recent works [40, 44, 49], which provide the attacker with a subset of the training set of the target model, our approach does not require any training data. Intuitively, if a model is susceptible to MIAs without resorting to training resources, we would expect it to be even more vulnerable in presence of additional data. As a matter of fact, we will show that in many cases additional samples are not necessary in order to accurately determine the membership of target samples.

1.1. Contributions

Below we list the contributions of this work:

- We propose a novel MIA (Sec. 3) that performs consistently well regardless of the architecture of the target model, and does not require training samples. This strategy exploits the distance between adversarial examples and the corresponding raw inputs.
- We perform a thorough revision of MIA strategies previously proposed in the literature and evaluate their performance (Sec. 5). Through this evaluation we show that several well-known machine learning models are vulnerable to MIAs.
- Empirically, we show that in most of the investigated scenarios the proposed MIA outperforms, or it is at least competing with, state-of-the-art methods that rely on a large amount of data samples to perform the attack. On the other hand, for large models (*e.g.*, DenseNet), we observe that training samples can grant a significant advantage to the attacker.

1.2. Related Work

Along the review cycle, the reviewers brought to our attention the work from the authors of [31] where they propose a similar strategy for MIAs to the one depicted in this article. We thank the reviewers for pointing out such concurrent work. However, their work differs from ours in two crucial ways: First, our attack exploits white-box access to the target model, which allows for different, more powerful adversarial strategies. Namely, they use HopSkipJump [10] and QEBA [30], while we use Auto-PGD [11]. Second, we evaluate the performance of our strategy on real-world models, testing our attack strategies and those of our competitors on pre-trained state-of-the-art ML models for image classification (AlexNet, DenseNet, ResNet, ResNext), while their work uses its own target model, which makes it difficult to compare the results directly.

Membership Inference Attacks. The work [49] introduced MIAs against machine learning models. The attacks proposed in [49] consist of training an attack model that observes the input-output relation of the target model. In this black-box scenario, the attacker can access the target model only by querying it. In order to be trained, the attacker requires a part of the training set of the target model. When this is not available, the attacker resorts to training its own *shadow models*, which share the same architecture as the target model, but provide the attacker with full knowledge of their training set. This seminal work was further extended by [40], which considers an attacker with white-box access to the target model, using intermediate outputs of the target model and gradients of the loss function as input to the attacker.

A more recent contribution [50] proposes the use of entropy and "Modified Entropy" in MIAs, while [44] studies the use of gradients with respect to the input samples, gradients with respect to model parameters, intermediate outputs and "distance to the decision boundary" in MIAs. Note that our Adversarial Distance strategy does not attempt to estimate the distance to the decision boundary, but to estimate the magnitude of the perturbation required to produce an adversarial example, *i.e.* a sample that is classified incorrectly with very high confidence. Remarkably, while [44] suggests that MIAs are ineffective against machine learning models, our revision of state-of-the-art MIA strategies provides evidence to the contrary. It is important to note that we repeated some of the same experiments of [44], obtaining different results.

The aforementioned works are taken as baselines for assessing the performance of our method. We reproduce their results and, like those works, we use the output of the target model, loss value, norm of the gradient and modified entropy. However, we directly implement MIAs as binary decision tests, without training attack models.

The authors of [52] propose to exploit the model's predictions on adversarial examples to perform MIAs. Their method consists of using the predictions on adversarial examples to distinguish members from non-members of the training set. This strategy is found to be significantly more effective against models that are robust against adversarial attacks. In our work we take a different approach, by measuring the size of the perturbation necessary to produce an adversarial example, rather than looking at the model's prediction on an adversarial example. In contrast to [52], our aim is to develop effective MIAs, and not to compare the vulnerability of different models.

In [47] the reader can find an extensive discussion on the success and complexity of MIAs based on the difficulty of the underlying machine learning task. In [55] a strategy similar to [49] is used, but the emphasis is on investigating how the vulnerability of the model is influenced by the model choice and dataset, rather than providing novel attack strategies. The work [64] studies the connection between

adversarial example against GoogLeNet's classification algorithm. The added noise changes the classifier's output from class "*panda*" to class "*gibbon*".

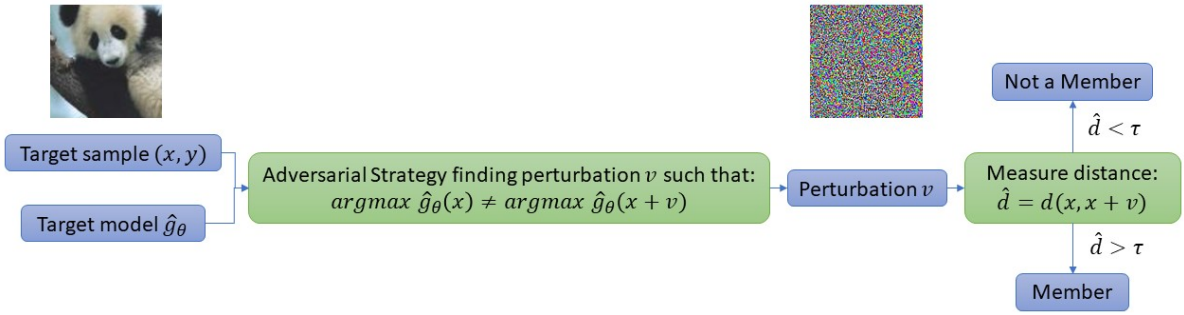


Figure 1. Illustrative diagram of our method to perform membership inference attack based on adversarial perturbations.

MIAs, attribute inference, differential privacy and overfitting, and proposes an MIA method that uses the value of the loss function to distinguish members from non-members of the training set. We use a similar strategy, but without requiring the average value of the loss of the target model over its training set. Finally, a comprehensive study of MIAs against GANs and other generative models is provided by [9]. Although similar attack strategies can be used in this context, generative models are beyond the scope of the present work.

Adversarial Attacks. Adversarial examples were first introduced in [53], showing that most ML models, and more specifically neural networks, are vulnerable to minor changes to their inputs [17]. Since then, a myriad of works has emerged in this field [8, 39, 45, 60] and opened the path for a better understanding of broader issues in generative models and mechanisms to fool algorithms, including deepfake technologies [26, 58]. We will use the adversarial attack proposed by [11], as it is highly adaptable and does not require fine-tuning of additional parameters.

2. Definitions and Preliminaries

In this section, we introduce the framework and notation used throughout the rest of the paper and the formal definitions for membership inference and adversarial attacks.

2.1. Pattern Recognition

Among the classical frameworks of machine learning, pattern recognition [14] consists of predicting a discrete random variable $Y \in \mathcal{Y}$ based on observing some multivariate random variable $X \in \mathcal{X}$, using a classifier g_θ from a class of classifiers $\mathcal{G}$, parameterized by θ. The classifier outputs a probability distribution on the label set $\mathcal{Y}$ (*i.e.* for any $g_\theta \in \mathcal{G}$, $g_\theta : \mathcal{X} \mapsto \mathcal{P}(\mathcal{Y})$) to predict Y with low classification risk $R(g_\theta) = \mathbb{E}\{\ell(g_\theta(X), Y)\}$, where loss function ℓ measures the error occurring between the true label Y and the one provided by the classification rule g_θ. As the joint distribution P_{XY} is unknown in practice, in order to select a possible classifier, one relies on $\mathcal{D}_n = \{(x_i, y_i)\}_{i=1}^n$ composed of $n \geq 1$ realizations of (X, Y). The empirical risk minimization paradigm [56] suggests to select the classifier minimizing the empirical risk $\widehat{R}(g_\theta) = \frac{1}{n}\sum_{i=1}^n \ell(g_\theta(x_i), y_i)$, among all the possible classifiers in $\mathcal{G}$. Let $\widehat{g}_\theta$ denote the classifier minimizing $\widehat{R}$ for a training set $\mathcal{D}_n$.

2.2. Membership Inference Attacks

Membership inference attacks (MIAs) can be used to measure the privacy leakage of ML models [49]. The goal of a MIA is to determine whether or not a sample (or group of samples) belongs to the training set of target model. Formally, MIAs can be stated as a binary decision test. Given a test sample $(x_{\text{test}}, y_{\text{test}})$, and the target model $\widehat{g}_\theta$ defined above, the goal of the attacker is to determine if $(x_{\text{test}}, y_{\text{test}}) \in \mathcal{D}_n$. Let φ be a scoring criteria, which takes as input the target model, the test sample and outputs a prediction score. This prediction score can be compared to a threshold $\tau \in \mathbb{R}_+$ to predict if the test sample belongs to the training set of the target model. Formally,

$$\begin{aligned} &\text{if} \quad \varphi((x_{\text{test}}, y_{\text{test}}), \widehat{g}_\theta) \geq \tau \quad \text{then} \quad (x_{\text{test}}, y_{\text{test}}) \in \mathcal{D}_n \\ &\text{otherwise} \quad (x_{\text{test}}, y_{\text{test}}) \notin \mathcal{D}_n \ . \end{aligned} \tag{1}$$

The hyper parameter τ of our approach selects the operating point in ROC curve. In practice, we make our analysis independent of τ by comparing performance for the whole range of possible τ values.

Hereafter, we consider different scoring criteria.

Softmax Response. Our main claim is that models tend to give more confident predictions over samples that belongs to their training set. This strategy aims to exploit the confidence of the predictions to identify members of the training set of the target model:

$$\varphi((x_{\text{test}}, y_{\text{test}}), \widehat{g}_\theta) = \max_{i \in \mathcal{Y}} \widehat{g}_\theta^i(x_{\text{test}}) \ , \tag{2}$$

where $\widehat{g}_\theta^i$ is the i-th component of the output of the model parametrized by θ. This observation has previously been used to build MIAs in [40, 49], training an attack model, while [52] directly compares the score to a threshold.

Modified Entropy. An alternative idea is to look at the uncertainty of the model. Intuitively, this should be lower for samples that were present in the training set. [50] proposes a metric called modified entropy, which decreases with the prediction probability of the correct class and increases with the prediction probability of any other class:

$$\varphi((x_{\text{test}}, y_{\text{test}}), \widehat{g}_\theta) = -\left(1 - \widehat{g}_\theta^{y_{\text{test}}}(x_{\text{test}})\right) \log\left(\widehat{g}_\theta^{y_{\text{test}}}(x_{\text{test}})\right) - \sum_{i \neq y_{\text{test}}} \widehat{g}_\theta^{i}(x_{\text{test}}) \log\left(1 - \widehat{g}_\theta^{i}(x_{\text{test}})\right) . \tag{3}$$

Unlike other metrics, modified entropy (3) takes into account whether the target model is predicting the correct class.

Loss. The learning objective of ML models is to minimize a loss function,

$$\varphi((x_{\text{test}}, y_{\text{test}}), \widehat{g}_\theta) = -\ell\left(y_{\text{test}}, \widehat{g}_\theta(x_{\text{test}})\right) , \tag{4}$$

over samples from the training set. Hence, we expect the value of the loss to be lower for samples in the training set. The minus sign in front of the loss is added to make this definition consistent with Eq. (1). An attack proposed in [64] compares the loss on the test sample to the average loss on the training set. The idea was also exploited in [49].

Gradient Norm. The loss function is minimized via Stochastic Gradient Descent, or similar iterative optimization algorithms. Around the optimal points, the gradient of the loss function with respect to its model parameters should approach 0. This attack strategy measures the ℓ_2 norm of the gradient of the loss function w.r.t. to the model parameters over different samples and expects this norm to be smaller for members of the training set,

$$\varphi((x_{\text{test}}, y_{\text{test}}), \widehat{g}_\theta) = -\|\nabla_\theta \ell\left(y_{\text{test}}, \widehat{g}_\theta(x_{\text{test}})\right)\|_2^2 . \tag{5}$$

Since we expect the norm of the gradient to be smaller for members of the training set, the minus sign is added to make this definition consistent with Eq. (1). This observation was first used in [49] as part of their MIA.

Although these ideas are not novel, most of them have not been used to make a binary decision test. Our aim is to assess and compare in a systematic way the power of these observations and whether or not it is possible to perform MIAs with them without requiring a training set for the attacker.

2.3. Adversarial Examples

The framework of untargeted adversarial examples can be set as follows: Given an input $x \in \mathcal{X}$ and target model $\widehat{g}_\theta : \mathcal{X} \mapsto \mathcal{P}(\mathcal{Y})$, the goal of adversarial strategy $\psi_{p,\epsilon}$ is to produce some perturbation $v \in \mathcal{X}$ such that the prediction provided by $\widehat{g}_\theta(x+v)$ changes from that provided by $\widehat{g}_\theta(x)$. Additionally, we require that the target model is confident on its prediction of the adversarial example. Formally, we define an untargeted adversarial strategy for a classifier $g \in \mathcal{G}$ as a function $\psi_{p,\epsilon} \colon \mathcal{X} \to \mathcal{X}$ on the input space $\mathcal{X}$, such that for any $x \in \mathcal{X}$ it obtains $v \stackrel{\text{def}}{=} \psi_{p,\epsilon}(x) \in \mathcal{X}$ with

$$\arg\max_{i \in \mathcal{Y}} g^i(x+v) \neq \arg\max_{i \in \mathcal{Y}} g^i(x) \text{ and} \tag{6}$$

$$\|v\|_p \leq \epsilon , \tag{7}$$

i.e., the constrained perturbation v changes the prediction of the target model to a wrong class.

Adversarial examples are computed constraining $\|v\|_p < \epsilon$, with $\epsilon \in \mathbb{R}_+$ and the ℓ_p norm $\|\cdot\|_p$ (see [3] for an extensive review on adversarial strategies). The purpose of this constraint is twofold: to perturb the original image in a way that is imperceptible for the human eye and to control the power of the attacker. In our case, the goal is not to produce subtle perturbations, the adversarial examples may be significantly different from their original counterparts. Indeed, our goal is to observe the size of the perturbation necessary to force the target model to drastically change its prediction, and use it as a criteria to distinguish members from non-members of the training set. Since $\psi_{p,\epsilon}$ will tend to compute the smallest perturbation possible such that $\widehat{g}_\theta$ changes its prediction, arbitrarily high ϵ can be allowed while still observing a significant difference between the size of the perturbation of samples in and outside the training set.

For our experiments we use *Auto-Attack* to build adversarial examples[2] [11]. The Auto-Attack library offers an ensemble of different strategies to compute adversarial examples. Particularly, we use auto Projected Gradient Descent (Auto-PGD). Given an objective function for the adversary $f_a : \mathcal{X} \mapsto \mathbb{R}$ and a constraint in the form $\mathcal{S} \subset \mathcal{X}$, Auto-PGD iteratively solves $\max_{x \in \mathcal{S}} f_a(x)$ by applying $x^{(k+1)} = P_{\mathcal{S}}\left(x^{(k)} + \eta^{(k)} \nabla_{x^{(k)}} f_a(x^{(k)})\right)$, for $k = [1, \ldots, N_{\text{iter}}]$, where $P_{\mathcal{S}}$ is the projection onto the surface of $\mathcal{S}$, and typically $f_a(x) = \ell(y, \widehat{g}_\theta(x))$. In the original algorithm introduced by [29, 36], the step size $\eta^{(k)}$ is fixed, while Auto-PGD uses an adaptive step size which improves the performance and makes the algorithm model-agnostic.

3. Membership Inference Attacks from Adversarial Examples

In this section, the core elements of this paper are discussed as we show how the adversarial distance bridges the gap between MIA and adversarial examples. We introduce our attack and describe the resulting algorithm in detail.

During training, the target model minimizes the loss over samples from the training set. The objective of Projected Gradient Descent [29,36] and other algorithms derived from

[2]Code available at https://github.com/fra31/auto-attack.

it (e.g., Auto-PGD [11]) is to maximize the very same loss. Hence, we expect this process to require larger perturbations for members of the training set, compared to samples that were not observed during training. We exploit this feature to perform MIAs against machine learning models.

Our membership inference strategy measures the distance between an adversarial example and its original counterpart, i.e., the size of the perturbation, and uses this as a criteria to distinguish members of the training set,

$$\varphi((x_{\text{test}}, y_{\text{test}}), \widehat{g}_\theta) = \|\psi_{p,\epsilon}((x_{\text{test}}, y_{\text{test}}), \widehat{g}_\theta)\|_p \ ,$$

where $\|\cdot\|_p$ measures the length of the perturbation. In our experiments, we use either l_1, l_2 or l_∞ norm (i.e., $p \in \{1, 2, \infty\}$) and the same norm is used to constraint the size of the perturbation produced by the adversarial strategy, guaranteeing that $\varphi((x_{\text{test}}, y_{\text{test}}), \widehat{g}_\theta) \leq \epsilon$ (see Algorithm 1).

Algorithm 1

Require: Target sample $(x_{\text{test}}, y_{\text{test}})$, target model $\widehat{g}_\theta$, adversarial strategy $\psi_{p,\epsilon}$, $p \in \{1, 2, \infty\}$, $\epsilon > 0$ and, $\tau \in \mathbb{R}_+$.

1. $v \leftarrow \psi_{p,\epsilon}((x_{\text{test}}, y_{\text{test}}), \widehat{g}_\theta)$
   \\Adversarial perturbation v.
2. **return** $\mathbb{1}\{\|v\|_p \geq \tau\}$
   \\Is the distance between the adv. ex. and the original input x_{test} greater than τ?

Since we are not interested in producing subtle perturbations that preserve the perspective of a human, we let the adversarial attacker generate arbitrarily large perturbations (constrained only by the dynamic range of the image). However, as shown in Fig. 2 and as demonstrated in the experimental section, there is a significant shift in the distribution of the size of perturbations, depending on whether (or not) the samples are part of the training set.

4. Review of MIAs Relying on Training Data

Most of the MIA strategies proposed in the literature combine sets of features of the target sample and target model. Combining these features into a single score that can be used for a binary decision test is a challenging task. A common strategy is to train a machine learning model that learns to combine these features and predict whether the target sample belongs to the training set or not. Naturally, the attack model requires a set of samples that are labeled as either part of the training set of the target model or outside of the training set of the target model. In this section we present a short review of attack models previously proposed in the literature:

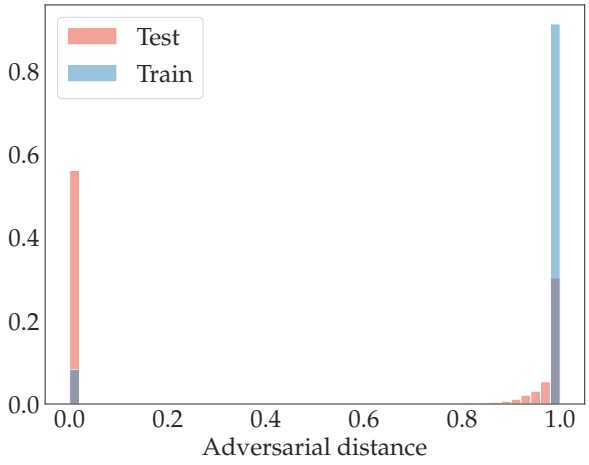


Figure 2. Histogram of adversarial distances over $50k$ samples from the training set (blue) superposed to the same histogram over $10k$ samples from the test set. The adversarial examples are computed for AlexNet, trained on CIFAR-100, based on the $\|\cdot\|_\infty$ norm and $\epsilon = 1$.

Grad x and Grad w Attack Models: In [44], they propose an attack model that uses an array of statistics from the gradient of the loss of target model. The statistics considered are the ℓ_1 norm, ℓ_2 norm, maximum value, mean, skewness, kurtosis and absolute minimum of the gradient. These statistics are combined by a logistic regression model trained on labeled data. We implement this attack model and reproduce the results in our setting. When the gradient is taken with respect to the model parameters, we refer to the attack model as 'Grad w'. On the other hand, when the gradient is taken with respect to the input sample, we refer to the attack model as 'Grad x'.

Intermediate Outputs: The authors of [44] also consider an attack model that uses the intermediate outputs of the target model. For the models considered in this work, the attacker uses the outputs of the last two layers of the target model. This attack model is also implemented as described in the original paper and evaluated in our setting. This model is later abbreviated as Int. Outs.

White-Box [40]: The attack model proposed by [40] utilizes the gradients of the loss function with respect to model parameters at the target sample, the value of the loss at the target sample, intermediate outputs of the target model and the one hot encoded labels of the target sample. To our knowledge this was the first work to propose using the gradient of the loss w.r.t. model parameters as a criteria to infer membership. We implement this attack strategy and reproduce the results presented in their paper. This attack strategy is referred to as WB [40].

Ensemble Attacker We propose an ensemble attacker that takes as input the Softmax response of the target model, the

value of its loss function, the norm of the gradient of the loss with respect to the model parameters, the norm of the gradient of the loss with respect to the input sample, the modified entropy, and the adversarial distance. This model outperforms the state-of-the-art against AlexNet, and achieves similar performance against ResNext. The details of the model are presented in the supplementary material [3] (Appendix A). The results for this model are also detailed in the supplementary material (Appendix C).

5. Numerical Experiments

Hereafter, we first present the experimental setting for MIAs and then provide numerical results on real world data. The code necessary to reproduce these experiments is available in our repository[4]. Further intuition and results are presented in the supplementary material (Appendix B).

5.1. CIFAR10 and CIFAR100 Datasets

These datasets are a standard benchmark for image recognition tasks [27]. They contain 60000 32×32 pixels, color (RGB) images split amongst 10, 100 distinct classes, respectively. In standard libraries, like PyTorch [41], these datasets are usually divided into a training set containing $50k$ images and a test set containing the remaining $10k$ images. The standard training set provided by PyTorch is used to train the target models we consider, the rest is used as outside-the-training-set data.

5.2. Target Models

State-of-the-art models for image recognition. We consider popular models for image recognition, pre-trained and publicly available[5]. Namely, the models considered are AlexNet [28], ResNet [19], ResNext [61] and DenseNet [21], trained for image classification on CIFAR-100. These are the same pretrained models considered in [40, 44].

5.3. Comparison of MIA strategies

To evaluate a membership inference strategy, two groups of samples are needed: samples from the training set and samples outside the training set of the target model. The pre-trained models considered in this work are trained on $50k$ samples from the CIFAR-100 dataset. The remaining $10k$ samples constitute the test set.

We perform membership inference attacks using nine different strategies. The Softmax Response, Modified Entropy and Loss strategies are black-box strategies, since the attacker only requires access to the target sample, its label and the output of the model (either the logits or Softmax response of the model). On the other hand, the Gradient Norm and Adversarial Distance strategies are white-box, as the attacker requires access to the model parameters in order to compute gradients of the loss function. In addition to white-box access to the target model, for some of the strategies we consider, the attacker requires its own training set of samples, including a subset of the training set used by the target model. This is the case for the Grad w, Grad x, Intermediate Outputs (Int. Outs) and the White-Box (WB) attacker.

In our analysis we consider a balanced evaluation set and report the AUROC score and the maximum accuracy achieved by each strategy. In this setting, a subset of $10k$ samples from the training set is selected uniformly as *in-training* data and the entire test set ($10k$ samples) is selected as *out-of-training* data. Since the choice of the subset of the training set influences our results, the experiments are repeated 20 times, choosing a different subset each time. All the quantities reported are averaged over these 20 runs of the experiment and the error reported is the empirical standard deviation. The results of this analysis are reported in Tab. 1. For each target model, the best performing attack strategies are highlighted in boldface. Note that the upper part of the table corresponds to strategies that do not require to train an attack model, nor require any additional samples, while the bottom part corresponds to strategies that require training an attack model. The best performing strategy when no additional samples are available is the adversarial distance strategy. Note that this strategy performs consistently across all target models and even surpasses the more resource hungry strategies for the case of AlexNet and ResNext. When additional samples are available, the Intermediate Outputs strategy is the most effective against the ResNet and DenseNet models. It is worth to mention that it might be infeasible for the attacker to obtain enough samples from the training set of the target model to launch this attack, in which case Adversarial Distance might be a better alternative. Figure 3 shows the ROC curves for the different strategies mentioned in the upper portion of Tab. 1. Note the jump on the ROC curve of the adversarial distance strategy. Due to a large group of samples having a score of 0, increasing the threshold slightly above results in a sudden increase in the true positive rate (TPR). Indeed, this strategy takes advantage of the fact that most misclassified samples lie outside of the training set.

Additionally, to demonstrate that our results are not biased by our analysis we perform the same analysis as in [44]. In this setting the whole training set ($50k$ samples) and the whole test set ($10k$ samples) are used, resulting in an unbalanced evaluation set. To alleviate this issue, the metrics considered are robust to the training:test sample ratio.

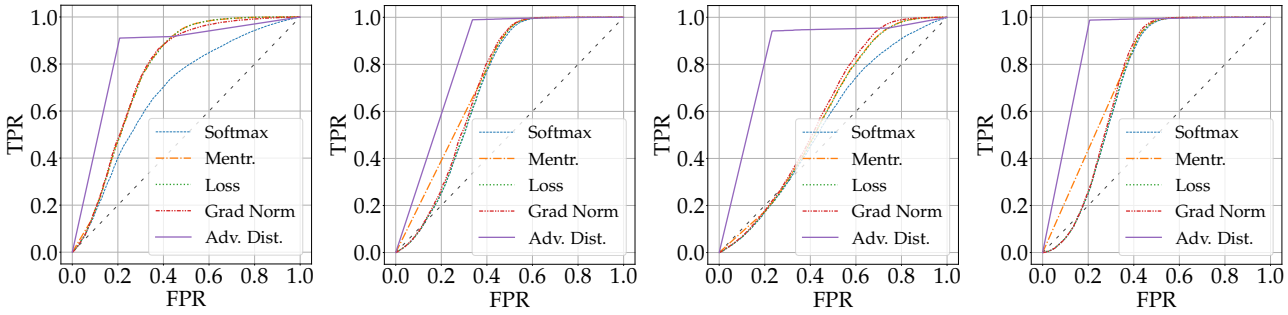
[3] Available at https://arxiv.org/abs/2203.09566

[4] https://github.com/ganeshdg95/Leveraging-Adversarial-Examples-to-Quantify-Membership-Information-Leakage

[5] Model implementations, pre-trained weights and code to train the models available at https://github.com/bearpaw/pytorch-classification

MIA Strategy	AlexNet		ResNet		ResNext		DenseNet	
	AUROC	Accuracy	AUROC	Accuracy	AUROC	Accuracy	AUROC	Accuracy
Softmax	68.00 ± 0.16	65.34 ± 0.14	55.45 ± 0.15	57.40 ± 0.13	72.37 ± 0.07	74.84 ± 0.11	70.52 ± 0.09	72.11 ± 0.10
Mentr. [50]	77.11 ± 0.10	74.16 ± 0.11	59.10 ± 0.13	61.39 ± 0.11	76.87 ± 0.08	75.28 ± 0.11	74.21 ± 0.10	72.69 ± 0.10
Loss	76.69 ± 0.10	74.14 ± 0.13	58.66 ± 0.13	61.29 ± 0.11	72.57 ± 0.07	75.17 ± 0.11	70.85 ± 0.09	72.61 ± 0.10
Grad Norm	76.58 ± 0.10	74.19 ± 0.12	59.93 ± 0.13	62.56 ± 0.09	73.06 ± 0.07	75.74 ± 0.11	71.30 ± 0.09	73.81 ± 0.09
Adv. Dist. (ours)	$\mathbf{84.35 \pm 0.13}$	$\mathbf{85.12 \pm 0.18}$	$\mathbf{84.53 \pm 0.16}$	$\mathbf{85.45 \pm 0.11}$	$\mathbf{89.24 \pm 0.03}$	$\mathbf{89.10 \pm 0.05}$	$\mathbf{82.76 \pm 0.03}$	$\mathbf{82.63 \pm 0.05}$
Grad w* [44]	78.76 ± 0.30	$\mathbf{74.32 \pm 0.28}$	61.98 ± 0.38	62.72 ± 0.27	77.80 ± 0.30	73.47 ± 0.57	73.12 ± 1.42	72.59 ± 0.55
Grad x* [44]	77.20 ± 0.26	73.43 ± 0.26	68.48 ± 0.27	63.58 ± 0.22	77.54 ± 0.61	73.47 ± 0.57	75.81 ± 0.43	71.81 ± 0.40
Int. Outs* [44]	57.92 ± 0.50	56.36 ± 0.41	$\mathbf{96.59 \pm 0.29}$	$\mathbf{91.57 \pm 0.43}$	$\mathbf{93.62 \pm 0.39}$	$\mathbf{86.38 \pm 0.37}$	$\mathbf{99.17 \pm 0.10}$	$\mathbf{97.68 \pm 0.14}$
WB* [40]	$\mathbf{80.33 \pm 1.21}$	74.03 ± 0.71	87.51 ± 0.41	79.73 ± 0.30	84.52 ± 1.95	76.46 ± 1.82	79.38 ± 1.16	71.92 ± 0.97

Table 1. Comparison of different MIA Techniques. The Accuracy(%) and AUROC score (%) on a balanced evaluation set are reported. $10k$ are uniformly selected from the training set (members) and the whole $10k$ samples from the testing set are selected (non-members). All the data selected is used for evaluation. Techniques with a (*) require training. In this case, only 60% of the data is used for evaluation and rest is used for training.



(a) AlexNet (b) DenseNet (c) ResNet (d) ResNext

Figure 3. ROC curves of different MIA strategies against AlexNet (3a), DenseNet (3b), ResNet (3c), ResNext (3d). These are computed on a balanced evaluation set and averaged over 20 iterations, as described for Tab. 1.

Namely, the False Positive Ratio (FPR) and balanced accuracy are considered. Table 2 presents the results of this analysis. Note that these results are consistent with Tab. 1; in particular, Adversarial Distance remains the best performing strategy across all models when no additional samples are available, also outperforming the state-of-the-art against AlexNet and ResNext. On the other hand, Intermediate Outputs is the best option against ResNet and DenseNet out of the models that require additional information and training.

To demonstrate that the effectiveness of our approach is not limited to one dataset, we repeat our experiments on CIFAR10. The target model considered is AlexNet and both the analyses from Tab. 1 and from Tab. 2 are considered. The results, shown in Tab. 3, are consistent with the results for CIFAR100, in the sense that our strategy remains the most effective. However, the performance of MIAs in general is worsen in comparison to CIFAR100. This is tied to the fact that the classification problem at hand has less classes, a trend that was previously observed in the literature [55]. Finally, we test how robust the AUROC score is to a change in the proportion of training/testing samples. For this purpose we perform membership inference attacks using the Adversarial Distance strategy, and measure the AUROC score in three different settings, where we change the training:testing ratio. The results presented in Tab. 4 show that the AUROC score measures the effectiveness of the attack consistently, regardless of the ratio of members to non-members present in the evaluation set.

6. Summary and Concluding Remarks

We have proposed a novel strategy for membership inference that pushes forward the state of the art. We have compared our strategy to the best performing membership inference attacks in the literature using the same setting (standalone scenario with popular image classification models on CIFAR-100) and found that our strategy is the most consistent across all target models. Despite not using additional resources (no additional samples from the training set of the target model and no computational resources to train an attacker), our model outperforms the more resource demanding state-of-the-art methods against AlexNet and ResNext, and remains relevant in other cases.

We found that MIAs are highly effective against machine learning models and even without additional knowledge of samples from the training set of the target model, MIAs can reliably distinguish training samples from non-training ones

MIA Strategy	AlexNet		ResNet		ResNext		DenseNet	
	Accuracy	FPR	Accuracy	FPR	Accuracy	FPR	Accuracy	FPR
Softmax	65.17 ± 0.48	43.01 ± 1.00	57.32 ± 0.43	63.64 ± 0.81	74.97 ± 0.52	45.82 ± 0.99	71.98 ± 0.51	51.25 ± 1.00
Mentr. [50]	74.04 ± 0.46	40.96 ± 1.23	61.38 ± 0.30	67.43 ± 0.84	75.44 ± 0.56	44.59 ± 1.12	72.59 ± 0.53	50.47 ± 1.03
Loss	74.00 ± 0.47	40.88 ± 1.20	61.28 ± 0.34	67.95 ± 0.63	75.31 ± 0.54	45.15 ± 1.04	72.49 ± 0.51	50.67 ± 0.84
Grad Norm	74.10 ± 0.52	39.19 ± 1.14	62.55 ± 0.31	67.81 ± 0.83	75.81 ± 0.58	43.78 ± 0.97	73.06 ± 0.50	49.56 ± 0.95
Adv. Dist. (ours)	$\mathbf{85.21 \pm 0.45}$	$\mathbf{20.60 \pm 0.87}$	$\mathbf{85.56 \pm 0.36}$	$\mathbf{22.98 \pm 0.77}$	$\mathbf{89.19 \pm 0.35}$	$\mathbf{20.39 \pm 0.70}$	$\mathbf{82.78 \pm 0.49}$	$\mathbf{33.37 \pm 0.97}$
Grad w* [44]	68.35 ± 0.50	60.74 ± 1.04	59.98 ± 0.34	78.86 ± 0.67	66.59 ± 0.42	66.74 ± 0.86	64.54 ± 0.37	70.87 ± 0.72
Grad x* [44]	62.84 ± 0.45	71.18 ± 1.00	59.76 ± 0.43	79.51 ± 0.85	65.37 ± 0.48	69.22 ± 0.96	63.48 ± 0.34	73.01 ± 0.69
Int. Outs* [44]	53.90 ± 0.45	77.05 ± 2.43	$\mathbf{87.97 \pm 1.34}$	$\mathbf{19.92 \pm 3.03}$	$\mathbf{85.67 \pm 0.91}$	$\mathbf{24.56 \pm 2.32}$	$\mathbf{96.45 \pm 0.88}$	$\mathbf{5.63 \pm 1.85}$
WB* [40]	$\mathbf{70.76 \pm 0.80}$	$\mathbf{56.54 \pm 1.91}$	65.75 ± 3.48	65.90 ± 7.70	69.33 ± 2.12	59.68 ± 4.81	65.35 ± 2.07	68.25 ± 4.64

Table 2. Comparison of different MIA Techniques. The balanced accuracy (%) and FPR (%) on an imbalanced evaluation set are reported. The whole training set ($50k$ samples) and testing set ($10k$ samples) are used; thus ratio between members and non-members is 5:1. 80% of the total data is used to find the threshold that maximizes accuracy and the other 20% is used to evaluate this threshold. Techniques with a (*) require training. In this case, instead of using 80% of the total data to determine the threshold, this 80% is used for training and the threshold is set at 0.5.

MIA Strategy	Analysis 1		Analysis 2	
	AUROC	Accuracy	FPR	Accuracy
Softmax	54.24 ± 0.21	54.17 ± 0.14	60.94 ± 1.01	54.13 ± 0.50
Mentr. [50]	57.05 ± 0.22	56.96 ± 0.17	69.58 ± 0.96	56.85 ± 0.45
Loss	56.99 ± 0.27	56.90 ± 0.17	70.25 ± 1.59	56.77 ± 0.45
Grad Norm	56.91 ± 0.20	56.95 ± 0.17	68.30 ± 0.86	56.91 ± 0.46
Adv. Dist. (ours)	$\mathbf{77.78 \pm 0.24}$	$\mathbf{80.03 \pm 0.17}$	$\mathbf{29.36 \pm 0.69}$	$\mathbf{80.03 \pm 0.36}$
Grad w* [44]	$\mathbf{60.36 \pm 0.34}$	$\mathbf{57.62 \pm 0.26}$	97.61 ± 0.29	51.03 ± 0.14
Grad x* [44]	60.08 ± 0.33	57.53 ± 0.34	99.51 ± 0.12	50.18 ± 0.06
Int. Outs* [44]	58.65 ± 0.58	56.65 ± 0.42	$\mathbf{80.38 \pm 2.93}$	$\mathbf{54.34 \pm 0.72}$
WB* [40]	55.57 ± 4.77	54.29 ± 3.48	92.17 ± 1.43	53.47 ± 0.60

Table 3. Comparison of different MIA techniques against the AlexNet model trained on CIFAR10. The best accuracy (%) and AUROC score (%) on a balanced evaluation set are reported (marked as "Analysis 1"). The best balanced accuracy (%) and FPR (%) at the same threshold on an unbalanced evaluation set are also computed (marked as "Analysis 2"). Techniques with a (*) require training.

Target Model	Training:Test ratio		
	5:1	1:1	1:5
AlexNet	84.36	84.35 ± 0.13	84.39 ± 0.41
ResNet	84.55	84.53 ± 0.16	84.36 ± 0.34
ResNext	89.24	89.24 ± 0.03	89.24 ± 0.08
DenseNet	82.76	82.76 ± 0.03	82.75 ± 0.08

Table 4. Influence of evaluation set on performance for the Adversarial Distance strategy. The AUROC (%) for different evaluation sets is reported. When the ratio is 5:1, the whole training is selected. When the ratio is 1:1, $10k$ samples from the training set are uniformly selected. When the ratio is 1:5, $2k$ samples from the training set are uniformly selected. In all cases the whole test set is also selected for evaluation.

(above 82% accuracy regardless of the target model).

6.1. Limitations

In this work we only consider target models trained on CIFAR10 and CIFAR100 as there is a lack of standard pre-trained models available for other data sets. Thence, comparing MIA strategies on such datasets is more difficult. In the future, we will extend our work to target models trained on other datasets.

On the other hand, always using the same pre-trained model for a particular architecture might result in a biased assessment of the privacy risks of that given architecture. Thus, it is equally important to evaluate the potential privacy risks on different shots of training, over the same dataset with the same architecture.

In this paper, we focused on the standalone scenario, *i.e.* the target model is trained by a single entity which has access to the whole training set and the attacker does not observe the model during training. This setup was chosen since it is the simplest setting. The ideas in this work can be extended to other scenarios (*e.g.*, federated learning).

6.2. Discussion of (Potential) Negative Impact

This article describes a novel attack strategy to retrieve information from the training data from any classification model. A direct potential negative impact of the work would be the improvement of attacks against machine learning models in production. Nevertheless, reliable and effective MIA strategies are needed to assess a model's privacy risks.

7. Acknowledgement

This research was supported by DATAIA "Programme d'Investissement d'Avenir" (ANR-17-CONV-0003) and by the ERC project Hypatia under the European Unions Horizon 2020 research and innovation program. Grant agreement No. 835294.

References

[1] The california consumer privacy act. 2018. 1

[2] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016. 1

[3] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018. 4

[4] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016. 1

[5] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery. 1

[6] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *The journal of machine learning research*, 3:1137–1155, 2003. 1

[7] Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(8):1207–1223, 2006. 1

[8] Y. Carmon, Aditi Raghunathan, Ludwig Schmidt, Percy Liang, and John C. Duchi. Unlabeled data improves adversarial robustness. In *NeurIPS*, 2019. 3

[9] Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. Gan-leaks: A taxonomy of membership inference attacks against generative models. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, CCS '20, page 343–362, New York, NY, USA, 2020. Association for Computing Machinery. 3

[10] Jianbo Chen, Michael I. Jordan, and Martin J. Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. In *2020 IEEE Symposium on Security and Privacy, SP 2020, San Francisco, CA, USA, May 18-21, 2020*, pages 1277–1294. IEEE, 2020. 2

[11] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2206–2216. PMLR, 13–18 Jul 2020. 2, 3, 4, 5

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[14] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013. 3

[15] David Forsyth and Jean Ponce. *Computer vision: A modern approach.* Prentice hall, 2011. 1

[16] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 1

[17] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1, 3

[18] Elizabeth Liz Harding, Jarno J Vanto, Reece Clark, L Hannah Ji, and Sara C Ainsworth. Understanding the scope and impact of the california consumer privacy act of 2018. *Journal of Data Protection & Privacy*, 2(3):234–253, 2019. 1

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[20] A Hern. Royal free breached uk data law in 1.6 m patient deal with google's deepmind', guardian, 3 july 2017, 2017. 1

[21] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 6

[22] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017. 1

[23] Simon Jegou, Michal Drozdzal, David Vazquez, Adriana Romero, and Yoshua Bengio. The one hundred

layers tiramisu: Fully convolutional densenets for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017. 1

[24] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021. 1

[25] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. 1

[26] David Khachaturov, Ilia Shumailov, Yiren Zhao, Nicolas Papernot, and Ross Anderson. Markpainting: Adversarial machine learning meets inpainting. *arXiv preprint arXiv:2106.00660*, 2021. 3

[27] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009. 6

[28] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 6

[29] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *CoRR*, abs/1607.02533, 2016. 4

[30] Huichen Li, Xiaojun Xu, Xiaolu Zhang, Shuang Yang, and Bo Li. Qeba: Query-efficient boundary-based blackbox attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[31] Zheng Li and Yang Zhang. Membership leakage in label-only exposures. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, CCS '21, page 880–895, New York, NY, USA, 2021. Association for Computing Machinery. 2

[32] Gaoyang Liu, Chen Wang, Kai Peng, Haojun Huang, Yutong Li, and Wenqing Cheng. Socinf: Membership inference attacks on social media health data with machine learning. *IEEE Transactions on Computational Social Systems*, 6(5):907–921, 2019. 1

[33] Yunhui Long, Vincent Bindschaedler, and Carl A Gunter. Towards measuring membership privacy. *arXiv preprint arXiv:1712.09136*, 2017. 1

[34] Yunhui Long, Vincent Bindschaedler, Lei Wang, Diyue Bu, Xiaofeng Wang, Haixu Tang, Carl A Gunter, and Kai Chen. Understanding membership inferences on well-generalized learning models. *arXiv preprint arXiv:1802.04889*, 2018. 1

[35] Michael Lustig, David L Donoho, Juan M Santos, and John M Pauly. Compressed sensing mri. *IEEE signal processing magazine*, 25(2):72–82, 2008. 1

[36] 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*, 2018. 4

[37] Deven McGraw and Kenneth D Mandl. Privacy protections to encourage use of health-relevant digital data in a learning health system. *npj Digital Medicine*, 4(1):1–11, 2021. 1

[38] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021. 1

[39] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 3

[40] Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE Symposium on Security and Privacy, SP 2019, San Francisco, CA, USA, May 19-23, 2019*, pages 739–753. IEEE, 2019. 2, 3, 5, 6, 7, 8

[41] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 6

[42] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 1

[43] General Data Protection Regulation. Regulation eu 2016/679 of the european parliament and of the council of 27 april 2016. *Official Journal of the European Union*, 2016. 1

[44] S. Rezaei and X. Liu. On the difficulty of membership inference attacks. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7888–7896, Los Alamitos, CA, USA, jun 2021. IEEE Computer Society. 2, 5, 6, 7, 8

[45] Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8093–8104. PMLR, 13–18 Jul 2020. 3

[46] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. *arXiv preprint arXiv:1806.01246*, 2018. 1

[47] Avital Shafran, Shmuel Peleg, and Yedid Hoshen. Membership inference attacks are easier on difficult problems. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14820–14829, 2021. 2

[48] Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1310–1321, 2015. 1

[49] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2017. 1, 2, 3, 4

[50] Liwei Song and Prateek Mittal. Systematic evaluation of privacy risks of machine learning models. In *USENIX Security Symposium*, 2021. 1, 2, 4, 7, 8

[51] Liwei Song, Reza Shokri, and Prateek Mittal. Privacy risks of securing machine learning models against adversarial examples. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 241–257, 2019. 1

[52] Liwei Song, Reza Shokri, and Prateek Mittal. Privacy risks of securing machine learning models against adversarial examples. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, CCS '19, page 241–257, New York, NY, USA, 2019. Association for Computing Machinery. 2, 3

[53] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 3

[54] Colin Tankard. What the gdpr means for businesses. *Network Security*, 2016(6):5–8, 2016. 1

[55] Stacey Truex, Ling Liu, Mehmet Emre Gursoy, Lei Yu, and Wenqi Wei. Demystifying membership inference attacks in machine learning as a service. *IEEE Transactions on Services Computing*, 2019. 1, 2, 7

[56] Vladimir Vapnik. Principles of risk minimization for learning theory. In *Advances in neural information processing systems*, pages 831–838, 1992. 3

[57] Robin Vogel, Aurélien Bellet, Magnet Team, Stephan Clémençon, and Télécom Paris LTCI. Learning fair scoring functions: Fairness definitions, algorithms and generalization bounds for bipartite ranking. *stat*, 1050:9, 2020. 1

[58] Mika Westerlund. The emergence of deepfake technology: A review. *Technology Innovation Management Review*, 9(11), 2019. 3

[59] Derek Wong and Stephen Yip. Machine learning classifies cancer, 2018. 1

[60] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 2958–2969. Curran Associates, Inc., 2020. 3

[61] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 6

[62] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019. 1

[63] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pages 268–282. IEEE, 2018. 1

[64] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pages 268–282, 2018. 2, 4