On the Difficulty of Membership Inference Attacks

Shahbaz Rezaei Xin Liu
University of California
Davis, CA, USA
{srezaei,xinliu}@ucdavis.edu

Abstract

Recent studies propose membership inference (MI) attacks on deep models, where the goal is to infer if a sample has been used in the training process. Despite their apparent success, these studies only report accuracy, precision, and recall of the positive class (member class). Hence, the performance of these attacks have not been clearly reported on negative class (non-member class). In this paper, we show that the way the MI attack performance has been reported is often misleading because they suffer from high false positive rate or false alarm rate (FAR) that has not been reported. FAR shows how often the attack model mislabel non-training samples (non-member) as training (member) ones. The high FAR makes MI attacks fundamentally impractical, which is particularly more significant for tasks such as membership inference where the majority of samples in reality belong to the negative (non-training) class. Moreover, we show that the current MI attack models can only identify the membership of misclassified samples with mediocre accuracy at best, which only constitute a very small portion of training samples.

We analyze several new features that have not been comprehensively explored for membership inference before, including distance to the decision boundary and gradient norms, and conclude that deep models' responses are mostly similar among train and non-train samples. We conduct several experiments on image classification tasks, including MNIST, CIFAR-10, CIFAR-100, and ImageNet, using various model architecture, including LeNet, AlexNet, ResNet, etc. We show that the current state-of-the-art MI attacks cannot achieve high accuracy and low FAR at the same time, even when the attacker is given several advantages. The source code is available at https://github.com/shrezaei/MI-Attack.

1. Introduction

There is an extensive recent literature on membership inference (MI) attacks on deep learning models that achieve high MI attack accuracy [20, 19, 12, 22, 13, 24, 14, 26].

Table 1: Complete evaluation of CIFAR-100 with three different target models. Almost all papers report the third section that includes accuracy, precision, recall, and F1 score. The second section, including train and test accuracy of the target victim model, is missing in many papers in literature despite its usefulness in evaluating the generalization gap (and degree of overfitting). The last section that includes balanced accuracy and FAR has never been reported, but it is of paramount importance for understanding the performance of attack models in practice.

Dataset Model	Cifar-100 AlexNet	Cifar-100 ResNet	Cifar-100 DenseNet
Target Model Train Acc.	92.48%	95.80%	99.98%
Target Model Test Acc.	43.87%	74.14%	82.83%
Attack Acc.	82.62%	79.13%	87.74%
Attack Precision	91.90%	87.3%	86.97%
Attack Recall	86.92%	87.85%	98.29%
Attack F1	89.23%	87.45%	92.26%
Attack Bal. Acc.	74.02%	61.70%	66.65%
Attack FAR	38.89%	64.45%	65.00%

These MI attack models often use confidence values of the target model to infer the membership of an input sample. High MI attack accuracy is often justified by claiming that deep learning models are more confident towards the training (member) samples than the samples they have not seen during training[1]. Consequently, MI attack accuracy is reported to be highly correlated to model's overfitting or generalization gap [19, 20, 22] because an overfitted model should perhaps behave even more confident towards training samples.

In this paper, we show that the way the previous papers report the attack performance do not reveal how exactly these attacks perform in practice and can be misleading. First, many papers do not provide the train and test accuracy of the target victim models. Hence, it is not clear whether the target

[1]In this paper, we use training samples, member samples, and positive samples interchangeably. Likewise, we also use non-training, test, non-member, and negative samples interchangeably.

model is well-trained. We only find a handful of papers that report such metrics [20, 13, 17, 8, 9]. Even in these cases, one can clearly spot impractical target models where generalization gap is sometimes larger than 35% [20, 13], 50% [17], or 80% [8]. Clearly, such extremely overfitted models have no practical use and the results on such models should not be generalized to well-trained models. Second, all papers we have examined limit their reporting to accuracy, precision, and recall. Such a reporting does not reveal the performance of attack models on negative samples (non-member), especially how many negative samples are misclassified as positive (false positive). For binary classification tasks, this is of crucial importance, especially when the negative class can significantly outnumber the positive class (e.g., all possible images vs. the limited number of images used to train an image classification model). A good practice, which is also common in other fields such as intrusion detection systems [25], is to report false positive rate (FPR) or false alarm rate (FAR) alongside the other metrics. A good attack model should have a low FAR.

To evaluate the feasibility of MI attack, we tried to reproduce the results in [17] for CIFAR-100 dataset, presented in Table 1[2]. Although the generalization gap of AlexNet is high ($\sim 48\%$), we keep the results for the sake of comparison. As it is shown, commonly used metrics, including accuracy, precision, and recall, do not reveal how attack models really perform on non-member samples. The high FAR of these attacks make them unreliable. Interestingly, even the attack on an extremely overfitted model such as AlexNet still suffers from high FAR.

In this paper, we first elaborate on why previous reporting practices are misleading in membership inference research. Second, we provide a comprehensive evaluation of membership inference attacks on deep learning models. We give as much advantage as possible to an attacker and we show that a reliable MI attack with high accuracy and low FAR is hard to achieve. We show that the reason MI attacks often fail is not because attack models are trained poorly. The reason is that the statistical properties of the features used in MI attacks are not clearly different and distinguishable for training and non-training sample.

To provide an insight on why membership inference of some samples are possible, we separate datasets into two parts: correctly classified samples (by the target victim model) and misclassified samples. In general, we find that membership inference of correctly classified samples, independent of what dataset or model is used, is a more difficult task than the membership inference of misclassified samples. This is because deep learning models often behave similarly on train and non-train samples when they are correctly classified. This observation sheds light on the difficulty of membership inference on deep models.

Our contributions are summarized as follows:

- We show that attack accuracy, precision, and recall are not enough to show the performance of MI attacks, particularly on negative (non-member) samples. Instead, we should also report FAR (or other substitutes explained in Sec.3) and train/test accuracy of target models to better demonstrate the performance of MI attacks. Moreover, we study the performance of correctly classified samples and misclassified samples separately. We show that membership inference of correctly classified samples, to which the majority of training samples belong, is a very difficult task.

- We perform MI attack on various image datasets (including MNIST, CIFAR-10, CIFAR-100, and ImageNet), and models (LeNet, AlexNet, ResNet, DenseNet, InceptionV3, Xception, etc), some of which are studied for the first time in the MI context. We conduct experiments such that they give a lot more advantages to the attacker than in any previous work. Even in this case, we show that a meaningful membership inference attack with high accuracy and low FAR is often not achievable.

- In addition to confidence values of the target (victim) model, we extensively analyze and use other information available from the target model, including values from intermediate layers, the gradient w.r.t input, gradient w.r.t to model weights, and distance to the decision boundary. In some cases, these types of information slightly leak more membership status than confidence values, but they are still not sufficient for a reliable MI attack in practice. Surprisingly, all evidence suggests that deep models often behave similarly on train and non-train samples across all these metrics. The only considerable difference appears between correctly classified samples and misclassified samples, not between the train and non-train samples.

In summary, our experiments, including the reproduction of results in the literature, suggest membership inference of correctly classified samples, to which the majority of training samples belong, is a difficult task. Clearly more research is needed and we are hesitant to generalize our results to all scenarios, some of which are as follows:

- We mainly focus on vision tasks with high dimensional input. Membership inference for other tasks with low dimensional input may culminate in a different result.

- We do not extend our conclusion to generative models. High capacity generative models can often memorize

[2] In literature, we have only found two papers with public source codes [22, 19]. We run their implementation on CIFAR-10 and observed the same problem. They both suffer from high FAR, which has not been reported before.

Table 2: Performance evaluation of an MI attack model when balancedness of the evaluation set is changed.

Balancedness	Attack Model	Positive (member) Class		Negative (non-member) Class		Balanced Accuracy	FAR
-	-	Precision	Recall	Precision	Recall	-	-
5:1	MI Attack	87.30%	87.85%	38.18%	35.55%	61.70%	64.45%
5:1	ZeroR	83.33%	100.0%	0.0%	0.0%	50.0%	100.0%
1:1	MI Attack	57.68%	87.42%	74.49%	35.42%	61.22%	64.82%
1:1	ZeroR	50.0%	100.0%	0.0%	0.0%	50.0%	100.0%
1:5	MI Attack	21.41%	87.82%	93.57%	35.73%	61.28%	64.42%
1:5	ZeroR	16.66%	100.0%	0.0%	0.0%	50.0%	100.0%

training samples, which can be retrieved at inference time, as shown in [2]. However, there is no trivial method to retrieve memorized samples from a discriminative model even if it memorizes training samples.

- We do not extend the conclusion to any attack that can be launched during the training phase, such as data poisoning, model/training manipulation [21], etc. We only study membership inference on naturally trained models and natural datasets. Deep models may behave very differently if any unnatural manipulation appears during the training phase.

- In each dataset, there often exists outliers. The MI attack maybe more successful on these samples [14], whether they are classified correctly or not.

2. Related Work

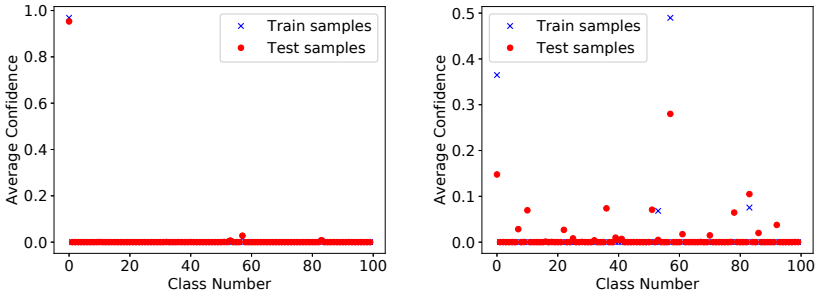
The first membership inference attack on deep models is proposed by Shokri et al. in [20]. The key idea is to build a machine learning attack model that takes the target model's output (confidence values) to infer the membership of the target model's input. To train the attack models, membership dataset containing (x_{conf}, y_{mem}) pairs is needed where x_{conf} represents the confidence values obtained by the target model for each sample and y_{mem} is a binary variable indicating whether the sample is used in target model's training or not. To build the membership dataset, a set of shadow models are trained for which the training and non-training samples are known. The attack is possible under two assumptions [19]: (a) the shadow models share the same structure as the target model, and (b) the training dataset used to train the shadow models share the same distribution as the one used to train the target model. To mitigate these limitations, Salem et al. [19] relax the second assumption by showing the attack is possible using different datasets and the first assumption by proposing a threshold-based attack that does not require a training procedure. To further relax these assumptions, several studies [12, 22, 13, 24, 14, 26] introduce better dataset generating procedures for shadow models, and extend the experiments to various scenarios and datasets. However, all studies share the same idea of using target model's output for membership inference. In this paper, we show that such attacks that rely on confidence values either suffer from high FAR or low accuracy. Moreover, we show that the output of deep learning models are often similar for correctly labeled samples, whether they were used in training or not.

3. Better MI Attack Reporting

As discussed in Section I, the common approach of reporting accuracy, precision, and recall of MI attacks does not truly reveal the performance of them on negative (non-member) samples. By providing false alarm rate (FAR) in Table 1, we show that these attacks suffer from high false positive ratio. The reason why high false positive ratio does not significantly affect the reported precision ($\frac{TP}{TP+FP}$) is due to the MI dataset imbalancedness. In a typical machine learning training, majority of samples are used for training and, consequently, the holdout (test) set is considerably smaller. For instance, the training:test (or member:non-member) ratio of CIFAR dataset is 5:1 and, consequently, the MI dataset to train/evaluate the MI attack model has the same imbalancedness ratio. As a result, the total number of FPs is small despite the high false positive ratio. This is clearly illustrated in Table 2 where the member:non-member ratio varies from 5:1, 1:1, to 1:5 and the precision on the positive class dropped from 87% to 21% for the same MI attack model. Note the MI attack model is the same for all experiments which was trained on a balanced dataset. We only change the balancedness of the evaluation set in Table 2.

We emphasize that reporting performance only on positive samples can be misleading. In Table 2, we report the precision and recall for both positive and negative classes. When the 5:1 ratio of balancedness is kept, the MI attack shows high precision and recall on positive samples, but low precision and recall on negative samples (which in general has not been reported). To stress this message, we also report ZeroR as a baseline. ZeroR is a classifier that always predicts the positive class (member). As shown in Table 2, ZeroR performance is high and close to the MI attack when **only** the precision and recall of positive samples are considered. This clearly shows that one should also report performance on negative samples as well.

The training:test ratio can significantly change the performance metrics of a same model, as shown in Table 2.



(a) Correctly labeled (b) Misclassified

Figure 1: Distribution of average confidence values for CIFAR 100 dataset (ResNet) (class 0)

So what is the right way to treat this ratio? Previous MI papers often keep the radio of 5:1, or in some cases 1:1 [17]. However, in practice, the ratio can reverse drastically: for example, a random sample, chosen from the distribution of all natural images, has a significantly higher probability of being a non-member sample than a member sample. However, there is no practical way to estimate the true value of this ratio. Furthermore, even if the ratio is known, due to the limited number of samples available for evaluation, it might not be practical to obtain more non-member samples to achieve the true ratio. As a result, in this paper, we keep the ratio as 5:1, as in most existing papers in the evaluation set. Instead, we report performance metrics that are less sensitive to the balancedness ratio, i.e, balanced accuracy and FAR[3].

There is another way to show the ineffectiveness of current MI attacks using a rather simple baseline. In [11], such a baseline is introduced, called naive attack. The idea is to predict a sample as member if it is correctly labeled by the target model, and to predict it as non-member if misclassified. Clearly, the FAR of naive attack is high because it classifies all correctly classified samples as members. However, this impractical attack can also achieve high accuracy on positive samples. Since this naive attack is obviously ineffective in practice, one can compare the accuracy gap between the naive attack and a MI attack to conclude the effectiveness of an MI attack. This approach has been used in [11]. We report the accuracy of this naive attack for completeness in Table 3, but we rely on accuracy/FAR pair to evaluate an attack's success.

Despite their low performance, MI attacks still outperform the random guess. To shed light on why MI attacks are effective on some samples, we report the behavior of target models on correctly classified and misclassified samples, separately. We show that deep learning models often behave similarly when they correctly label a sample, whether it is a training (member) or test (non-member) sample. In comparison, deep learning models demonstrate slightly different behavior on misclassified samples, which can be exploited by MI attack models. Hence, we believe that separating correctly classified and misclassified samples, and reporting the MI attack on them separately provides a better insight on how MI attacks work.

In summary, the way MI attack has been reported in literature does not provide a complete picture of their performance. Relying only on precision and recall of the positive class can present a delusion of successful attack, particularly when the imbalancedness issue is ignored. Instead, we should report performance on both positive and negative samples, i.e., adding FAR, or precision and recall on negative samples. Furthermore, simple baselines, such as ZeroR and naive attack, can be used for comparison. Last, separating correctly labeled and incorrectly labeled samples provide a better insight on how these MI attacks work. Therefore, in this paper, we report balanced accuracy and FAR, and we also report the performance on correctly classified and misclassified samples, separately, when possible.

4. Methodology

Threat model and assumptions In this paper, we give an attacker as much advantage as possible to show that even in such cases membership inference cannot significantly outperform the baseline. We assume a white-box access to the model and unlimited number of queries. Moreover, we give the membership status of up to 80% of training samples and test samples to the attackers and we only ask the attack model to predict the membership inference of the remaining samples. Hence, the attack performance we report in this paper is as good as or better than any proposed attack based on shadow models [20, 13] or transferred or synthesized data [12, 19] [4]. In addition to confidence values of the target

[3] A more elaborated argument for why FAR is important in tasks where one class hugely outnumber the other class is the base-rate fallcay [1].

[4] Note that the use of these methods are beneficial when the dataset is small for training a MI attack because one can potentially multiply the MI training dataset by obtaining multiple shadow models. However, in our

model, which have been used extensively for membership inference attack in the past, we also study the output of intermediate layers, distance to the decision boundary and a set of gradient norms to better understand if deep models behave differently on training and test samples.

Confidence values Confidence values, or the output of Softmax layer, have been widely used for membership inference [20, 19, 12, 22, 13, 24, 14, 26]. Figure 1 shows the distribution of average confidence for correctly classified samples and misclassified samples of a ResNet model trained on CIFAR-100. As it is shown, misclassified samples often show different distribution for training samples and non-training samples. However, correctly classified samples often saturate the true class confidence value and zero out other confidence values. We show in Section 5 that membership inference attack models are often fail to considerably outperform a coin toss for correctly labeled samples.

Output of intermediate layers In deep models, first layers often extract general and simple features that are not specific to training samples. As suggested in [17], the last layer and layers close to the last one contain more sample-specific information. Hence, we also examine the output of the fully connected layers before the Softmax for membership inference attacks.

Algorithm 1 FGM-based algorithm to find distance to the decision boundary

Input: S (maximum number of steps), $\mathbf{x}$ (input sample) and y (the sample ground-truth), f_{cls} (an interface to the target model which returns predicted class), f_{conf} (an interface to the target model which returns the confidence value of the predicted class), L (target model loss function), θ (confidence threshold indicating when the algorithm stops optimizing):

procedure DISTANCE_TO_BOUNDARY($\mathbf{x}$, f)
  $\mathbf{x}_0 = \mathbf{x}$
  for t from 0 to S **do**
    $\mathbf{x}_{t+1} = \mathbf{x}_t + \varepsilon \frac{\nabla_x L(x_t, y)}{||\nabla_x L(x_t, y)||_2}$
    if $f_{cls}(\mathbf{x}_{t+1}) \neq f_{cls}(\mathbf{x}_t)$ **then**
      while $|f_{conf}(\mathbf{x}_{t+1}) - f_{conf}(\mathbf{x}_t)| > \theta$ **do**
        $\mathbf{x}_m = \frac{\mathbf{x}_{t+1} + \mathbf{x}_t}{2}$
        if $f_{cls}(\mathbf{x}_m) = f_{cls}(\mathbf{x}_t)$ **then**
          $\mathbf{x}_t = \mathbf{x}_m$
        else if $f_{cls}(\mathbf{x}_m) = f_{cls}(\mathbf{x}_{t+1})$ **then**
          $\mathbf{x}_{t+1} = \mathbf{x}_m$
        else
          return *Error!*
      return $||\mathbf{x}_0 - \mathbf{x}_t||_2$
  return *Optimization failed!*

Distance to the decision boundary Some research focuses on understanding decision boundary of deep models [15, 10] or geometry and space of deep models [5, 16] to often understand the nature of adversarial examples or to improve robustness. In this paper, we investigate whether the distance to boundary is a distinguishable feature for membership inference. To find the distance to the decision boundary, we use FGM [4] optimization procedure to craft an image on the other side of the decision boundary (Algorithm 1). Then, we perform a binary search to find an instance for which the model's confidence for two classes are almost equal, that is, the difference between two confidences is smaller than a small threshold, similar to [10]. Finally, we obtain the L_2 distance between the original sample and the crafted samples as a measure of distance to the boundary.

study, such methods are not necessary since MNIST, CIFAR, and ImageNet datasets have abundant samples.

Table 3: Accuracy of various datasets, target models, and MI attack models

Dataset(Model)	Train	Test	Attack Acc	Naive Attack
MNIST (LeNet)	99.74%	99.05%	50.04%	50.07%
C-10 (AlexNet)	91.80%	77.22%	57.43%	57.87%
C-10 (ResNet)	99.43%	93.89%	54.11%	52.56%
C-10 (DenseNet)	100.00%	95.46%	56.05%	52.45%
C-100 (AlexNet)	92.48%	43.87%	74.02%	70.23%
C-100 (ResNet)	95.80%	74.14%	61.70%	65.68%
C-100 (DenseNet)	99.98%	82.83%	66.65%	71.97%
I (InceptionV3)	87.91%	79.98%	50.03%	54.12%
I (Xception)	87.77%	80.70%	51.18%	53.37%

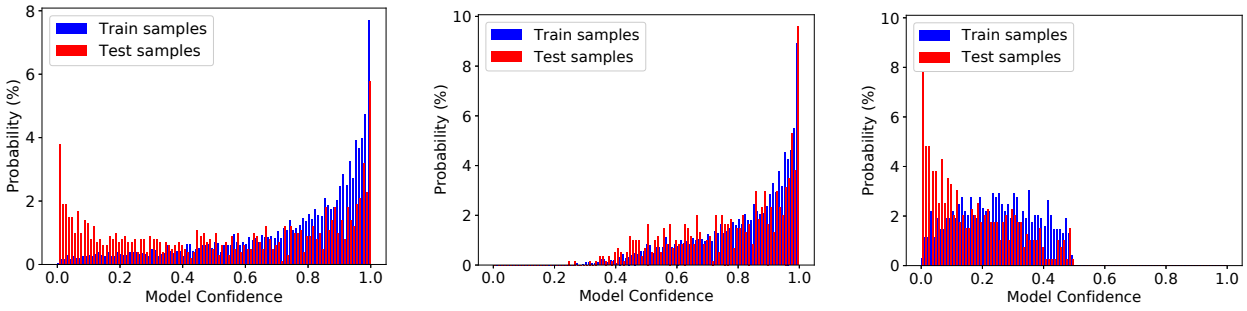
Gradient norm It has been shown that the gradient of loss with respect to model parameters, $\frac{\partial L}{\partial w}$, is often smaller for training samples than non-training samples [17] and it can be used for membership inference attack in federated learning scenario. In this paper, we study the gradient of loss with respect to model parameters, $\frac{\partial L}{\partial w}$, and also the gradient of loss with respect to model input, $\frac{\partial L}{\partial x}$, in a non-federated learning setting. The large value for the former indicates that major re-tuning of model parameters is needed for that sample, and hence, it can be an indication of a non-member sample. The large value of the latter indicates that there are input samples with more confident output in the vicinity of that sample, and hence, it can be an indication of a non-training sample. Both $\frac{\partial L}{\partial w}$ and $\frac{\partial L}{\partial x}$ are extremely high dimensional. Thus, we adopt the seven norms used in [18], originally used for analysis of deep model's uncertainty, namely L_1, L_2, absolute minimum, L_∞, mean, Skewness, and Kurtosis.

5. Experimental Evidence

Target models and datasets We launch MI attack on various CNN-based models on four image classification tasks: MNIST, CIFAR-10 (C-10), CIFAR-100 (C-100), ImageNet (I). For MNIST, we train a LeNet model. For CIFAR-10 and CIFAR-100, we use a set of trained models used in

Table 4: Membership attack results based on confidence values

Dataset(Model)	-	Attack Accuracy	Attack FAR	Train Confidence	Test Confidence
MNIST	All data	50.04% ± 0.11	50.01% ± 47.81	99.61 ± 4.57	98.90 ± 9.14
	Correctly classified	49.98% ± 0.01	50.21% ± 48.38	99.81 ± 2.04	99.75 ± 2.51
	Misclassified	62.30% ± 17.93	40.83% ± 35.44	77.61 ± 16.05	87.09 ± 15.93
CIFAR-10 (AlexNet)	All data	57.43% ± 2.59	71.4% ± 9.29	85.26 ± 23.08	72.56 ± 34.75
	Correctly classified	50.54% ± 0.82	91.89% ± 5.43	90.77 ± 13.96	89.57 ± 15.52
	Misclassified	52.16% ± 2.57	5.45% ± 5.62	60.17 ± 16.68	66.89 ± 19.68
CIFAR-10 (ResNet)	All data	54.11% ± 1.92	86.60% ± 6.49	98.66 ± 7.03	92.74 ± 22.02
	Correctly classified	51.81% ± 0.84	91.63% ± 4.05	99.08 ± 4.33	97.98 ± 7.33
	Misclassified	60.83% ± 19.74	0.0% ± 0.0	66.23 ± 15.66	79.71 ± 18.61
CIFAR-10 (DenseNet)	All data	56.05% ± 3.88	77.05% ± 26.5	99.97 ± 0.49	94.78 ± 19.33
	Correctly classified	54.0% ± 2.64	81.15% ± 27.45	99.97 ± 0.29	98.77 ± 5.64
	Misclassified	100.0% ± 0.0	0.0% ± 0.0	-	82.83 ± 17.63
CIFAR-100 (AlexNet)	All data	74.02% ± 8.27	38.89% ± 16.69	84.59 ± 24.31	40.58 ± 42.09
	Correctly classified	55.13% ± 7.39	83.12% ± 14.94	89.9 ± 15.82	85.05 ± 19.95
	Misclassified	55.11% ± 11.28	2.01% ± 4.33	50.29 ± 19.45	60.96 ± 23.81
CIFAR-100 (ResNet)	All data	61.7% ± 6.55	64.45% ± 14.81	91.14 ± 18.44	70.19 ± 38.10
	Correctly classified	53.96% ± 5.25	83.7% ± 11.01	94.15 ± 11.43	90.88 ± 15.74
	Misclassified	54.37% ± 16.0	2.66% ± 8.08	57.82 ± 17.69	64.3 ± 21.71
CIFAR-100 (DenseNet)	All data	66.65% ± 8.36	65.0% ± 18.16	99.95 ± 1.07	79.99 ± 34.70
	Correctly classified	60.58% ± 5.98	77.17% ± 15.28	99.96 ± 0.5	94.67 ± 13.06
	Misclassified	98.25% ± 9.2	0.0% ± 0.0	65.01% ± 11.52	67.25% ± 24.61
ImageNet (InceptionV3)	All data	50.03% ± 0.28	45.96% ± 44.27	76.03 ± 25.52	68.62 ± 29.43
	Correctly classified	50.03% ± 0.31	45.46% ± 47.86	83.99 ± 15.55	81.85 ± 16.3
	Misclassified	51.5% ± 8.57	51.69% ± 44.69	13.57 ± 11.44	10.8 ± 9.38
ImageNet (Xception)	All data	51.18% ± 3.56	50.81% ± 44.33	73.56 ± 25.56	66.92 ± 28.58
	Correctly classified	50.72% ± 3.57	50.42% ± 48.80	81.24 ± 16.81	79.08 ± 17.18
	Misclassified	51.95% ± 19.29	52.15% ± 44.08	13.48 ± 10.94	11.27 ± 8.94



(a) All samples (b) Correctly classified samples (c) Misclassified samples

Figure 2: Distribution of the confidence of the true class for CIFAR-10 (AlexNet) class #3. Although the distribution of all samples seems to be distinguishable, it is only the manifestation of accuracy gap between train and test which is exploited by naive and other attack. When correctly classified and misclassified samples are depicted separately, sample difference of train and test sets is vividly less distinguishable.

[17][5], including AlexNet, ResNet [6], and DenseNet [7]. For ImageNet, we use pre-trained InceptionV3 [23] and Xception [3] models without any re-training from Keras package[6]. We deliberately choose to launch MI attacks on models trained by others for two reasons: 1) to make it easier for readers to compare the attack with papers that use the same trained models, such as [17], and 2) to avoid any intentional/unintentional model training practices that bias our results. To the best of our knowledge, the models trained on CIFAR-10 and CIFAR-100, used in [17], has not adopted the early stopping technique during the training phase. In supplementary material, we train all models on CIFAR-10 and CIFAR-100 from scratch and we show that using early stopping techniques can make MI attacks even harder.

[5]https://github.com/bearpaw/pytorch-classification

[6]https://keras.io/api/applications/

Table 5: MI attack performance based on the output of intermediate layers. C and I represents CIFAR and ImageNet, respectively

Dataset (Model)	Layer	All Data		Correctly Classified		Misclassified	
-	-	Accuracy	FAR	Accuracy	FAR		
MNIST	-1	47.65% ± 2.6	59.55% ± 17.95	47.58% ± 2.6	59.69% ± 17.98	55.45% ± 20.49	43.33% ± 40.1
MNIST	-2	47.96% ± 3.01	60.77% ± 11.33	47.99% ± 2.93	60.65% ± 11.28	47.42% ± 25.3	69.17% ± 41.51
C-10 (AlexNet)	-1	55.38% ± 2.18	47.45% ± 9.55	51.34% ± 2.46	58.47% ± 12.79	55.47% ± 3.69	14.49% ± 9.95
C-10 (ResNet)	-1	53.89% ± 2.62	45.1% ± 16.25	52.74% ± 2.2	47.63% ± 17.43	60.75% ± 20.8	5.51% ± 7.21
C-10 (DenseNet)	-1	54.4% ± 3.63	48.7% ± 6.99	53.12% ± 3.03	51.28% ± 7.54	100.0% ± 0.0	0.0% ± 0.0
C-100 (AlexNet)	-1	61.34% ± 7.91	38.95% ± 12.48	55.39% ± 10.45	52.46% ± 19.99	57.65% ± 12.98	29.32% ± 15.65
C-100 (ResNet)	-1	53.81% ± 7.24	47.2% ± 16.65	51.46% ± 7.72	52.85% ± 18.99	52.39% ± 23.72	31.2% ± 27.8
C-100 (DenseNet)	-1	64.76% ± 9.99	37.35% ± 14.67	61.7% ± 9.4	43.48% ± 14.99	92.02% ± 20.78	38.73% ± 31.23
I (InceptionV3)	-1	58.37% ± 7.8	40.2% ± 15.68	57.83% ± 9.4	42.58% ± 19.0	55.86% ± 17.26	36.25% ± 35.4
I (Xception)	-1	57.44% ± 8.59	42.3% ± 17.08	57.35% ± 9.36	43.92% ± 18.61	55.52% ± 18.65	40.38% ± 39.49

MI attack models In most cases, we fit three types of attack models: FC neural network (NN), random forest (RF), and XGBoost. For the NN model, we train a model with 2 hidden layers of size 128 and 64. For RF and XGBoost, we perform a random search over a large set of hyper-parameters. For ImageNet, we conduct experiments with only 100 classes out of 1000 classes due to the limited computational and time budget. Moreover, we perform random under-sampling of member class and oversampling of non-member class to balance the training dataset on separate experiments. In this paper, we only report the seemingly best MI attack accuracy we achieve over all attack models and hyper-parameters. It is worth mentioning that by under-sampling or over-sampling of training data, or by chainging the decision threshold of the MI attack, we could decrease FAR at the cost of accuracy. In any case, we could not find a good MI attack model with relatively high accuracy and low FAR. The input of attack models varies which is described in each following subsection. The accuracy of target and MI attack models are shown in Table 3. Note that even the best MI attack models can barely outperform the naive attack. In the following sections, we show that separating correctly classified and misclassified samples, and reporting accuracy and FAR gives more insight on the performance of MI attacks.

5.1. Confidence Values

Confidence values have been extensively used for MI attacks. As shown in Table 4, the MI attacks are more successful on inferring membership of misclassified samples, which often consist a small portion of training samples. Interestingly, the state-of-the-art target models on ImageNet does not even leak membership status of misclassified samples. The best attack performance on correctly classified samples is observed on DenseNet model trained on CIFAR-100, which is 60.58% that still suffers from very high FAR (77.17%). Note that the MI attacks on misclassified samples of DenseNet model may not be meaningful because there are no misclassified samples in the training set of CIFAR-10 and there are only 10 misclassified samples in the training set of CIFAR-100.

To better understand why MI attacks fail, it is better to investigate the average confidence value of target models, shown in the fifth and sixth columns of Table 4. As shown, the average confidence values of train samples (members) are often close to the test samples (non-members). MI attacks are only partially successful when average confidence values between train and test samples are far apart and the standard deviation is low. As shown in Figure 2, by separating correctly classified samples and misclassified sample, we can observe that sample distribution is very close, particularly for correctly classifies samples.

5.2. Output of Intermediate Layers

The attack accuracy based on the output of intermediate layers are shown in Table 5. We only launch an attack on the output of FC or flattened layers. The layer column shows the number of layers we go back from the Softmax layer. Only the MI attack on ImageNet is more successful in terms of both accuracy and FAR. Nevertheless, the FAR is still high and accuracy is not considerably better than a random guess.

5.3. Distance to the Boundary

Since the distance to the decision boundary is one-dimensional, we only fit a logistic regression to the samples. As shown in Table 6, all MI attacks fail. By looking at the average distance and their standard deviation, it is evident that the distance to boundary is not a distinguishable feature for membership inference. Note that finding the exact distance to the boundary is a computationally heavy task for high dimensional data. What is clear in Table 6 is that the

Table 6: Performance of attack models based on the distance to the decision boundary.

Dataset (Model)	Correctly Classified				Misclassified			
-	MI Attack		Average Distance		MI Attack		Average Distance	
-	Accuracy	FAR	Train	Test	Accuracy	FAR	Train	Test
MNIST	49.86% ± 6.0	52.97% ± 8.2	1.372 ± 0.40	1.371 ± 0.45	49.81% ± 9.0	48.3% ± 24.0	.0103 ± .0016	.0108 ± .0024
C-10 (AlexNet)	52.36% ± 4.1	53.19% ± 17	51.59 ± 15.2	50.31 ± 14.7	52.42% ± 4.9	49.53% ± 7.1	48.84 ± 15.8	50.5 ± 16.0
C-10 (ResNet)	51.05% ± 5.5	49.57% ± 8.2	51.3 ± 15.2	50.3 ± 14.7	46.05% ± 4.8	60.49% ± 8.6	51.2 ± 15.2	50.5 ± 16.0
C-10 (DenseNet)	50.14% ± 5.9	50.17% ± 7.8	51.4 ± 15.3	50.2 ± 14.6	100.0% ± 0.0	0.0% ± 0.0	-	51.2 ± 16.4
C-100 (AlexNet)	50.45% ± 9.1	47.77% ± 6.3	53.6 ± 16.3	54.3 ± 15.1	53.78% ± 11.6	45.94% ± 17.7	48.2 ± 16.5	52.6 ± 16.6
C-100 (ResNet)	49.08% ± 6.4	49.28% ± 6.8	52.6 ± 16.4	53.3 ± 15.9	47.24% ± 12.4	52.05% ± 21.4	51.9 ± 16.6	51.6 ± 16.8
C-100 (DenseNet)	49.74% ± 6.41	48.57% ± 7.04	52.8 ± 16.3	53.3 ± 16.0	98.25% ± 9.2	0.0% ± 0.0	51.5 ± 10.4	56.8 ± 16.5
I (InceptionV3)	51.75% ± 8.2	50.43% ± 26.3	212.4 ± 64.8	218.6 ± 63.0	44.95% ± 11.9	46.31% ± 18.8	215.8 ± 67.0	215.1 ± 61.5
I (Xception)	53.01% ± 9.5	46.1% ± 23.7	214.5 ± 64.6	216.2 ± 61.9	49.29% ± 13.5	44.88% ± 17.1	212.4 ± 67.0	214.7 ± 64.0

Table 7: Performance of attack models based on gradient norms with respect to input (x) and weights (w)

Dataset (Model)	Correctly Classified				Misclassified			
-	Grad w.r.t w		Grad w.r.t x		Grad w.r.t w		Grad w.r.t x	
-	Accuracy	FAR	Accuracy	FAR	Accuracy	FAR	Accuracy	FAR
MNIST	52.06% ± 3.7	42.75% ± 28.5	53.19% ± 3.5	34.5% ± 23.7	57.84% ± 26.8	38.48% ± 20.0	52.02% ± 22.2	41.79% ± 16.8
C-10 (AlexNet)	51.94% ± 4.1	41.38% ± 7.9	51.81% ± 4.4	39.38% ± 10.3	61.36% ± 6.6	39.27% ± 8.2	58.69% ± 5.4	35.92% ± 7.2
C-10 (ResNet)	52.75% ± 3.0	21.75% ± 12.2	49.88% ± 3.8	36.25% ± 13.9	50.85% ± 10.6	66.49% ± 26.8	50.26% ± 16.1	50.36% ± 31.3
C-10 (DenseNet)	54.88% ± 2.6	16.38% ± 5.4	54.37% ± 3.1	13.25% ± 8.3	100.0% ± 0.0	0.0% ± 0.0	100.0% ± 0.0	0.0% ± 0.0
C-100 (AlexNet)	57.71% ± 9.6	31.45% ± 14.6	56.65% ± 11.5	31.46% ± 9.9	67.32% ± 11.3	34.51% ± 19.9	59.12% ± 13.8	38.52% ± 23.5
C-100 (ResNet)	52.98% ± 6.2	37.72% ± 21.1	54.31% ± 6.9	29.71% ± 11.3	55.4% ± 17.9	54.87% ± 34.6	61.38% ± 17.3	36.58% ± 27.1
C-100 (DenseNet)	69.7% ± 7.8	8.03% ± 6.2	70.22% ± 7.4	6.12% ± 3.6	98.25% ± 9.2	0.0% ± 0.0	98.25% ± 9.2	0.0 ± 0.0
I (InceptionV3)	49.25% ± 11.7	42.5% ± 19.0	52.1% ± 9.6	48.91% ± 14.7	58.48% ± 15.3	30.8% ± 20.3	50.55% ± 16.5	42.75% ± 17.3
I (Xception)	53.48% ± 9.0	36.53% ± 19.3	53.26% ± 11.4	48.05% ± 14.2	47.89% ± 18.2	43.69% ± 17.7	49.65% ± 15.6	41.25% ± 17.4

FGM-based approximation of distance to boundary does not provide a distinguishable feature. A more accurate approach may reveal more membership information.

5.4. Gradient Norm

Table 7 shows the performance of MI attack models based on gradient norms. For each case, we fit a logistic regression to the 7 norms introduced in Section 4. Except for ImageNet, all MI attacks on correctly classified samples achieve almost the same accuracy while having a lower FAR, in comparison with MI attacks based on confidence values (Table 4). Gradient information of misclassified samples leak less membership information than confidence values.

6. Conclusion

In this paper, we show that commonly-used MI attacks based on confidence values of deep models are not as reliable as it has been reported before. By reporting accuracy and FAR together, we show that MI attacks that achieve higher accuracy suffers from higher FAR. Previous MI attacks extensively rely on confidence values of the target model for membership inference. We show that such a membership inference in general is a difficult task because the distribution of confidence values are similar for member and non-member samples. We report the attack accuracy on correctly classified samples and misclassified samples separately to show that misclassified samples slightly leak more information for membership inference. Additionally, we analyze several other features of input samples, including the distance to the decision boundary and gradient norms, to further illustrate the difficult nature of reliable membership inference attack on deep models. In summary, we find that naturally trained deep models often behave similarly across training and test samples and, hence, an accurate membership inference attack on all training samples in practice is a difficult and inaccurate task under current attack models unless a new revolutionary approach is introduced.

Acknowledgement. This work was supported by the National Science Foundation (NSF) under Grant CNS-1547461, Grant CNS-1718901, and Grant IIS-1838207

References

[1] Stefan Axelsson. The base-rate fallacy and the difficulty of intrusion detection. *ACM Transactions on Information and System Security (TISSEC)*, 3(3):186–205, 2000.

[2] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pages 267–284, 2019.

[3] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[4] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018.

[5] Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard, and Stefano Soatto. Empirical study of the topology and geometry of deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3762–3770, 2018.

[6] 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.

[7] 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.

[8] Bargav Jayaraman and David Evans. Evaluating differentially private machine learning in practice. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pages 1895–1912, 2019.

[9] Jinyuan Jia, Ahmed Salem, Michael Backes, Yang Zhang, and Neil Zhenqiang Gong. Memguard: Defending against black-box membership inference attacks via adversarial examples. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 259–274, 2019.

[10] Hamid Karimi, Tyler Derr, and Jiliang Tang. Characterizing the decision boundary of deep neural networks. *arXiv preprint arXiv:1912.11460*, 2019.

[11] Klas Leino and Matt Fredrikson. Stolen memories: Leveraging model memorization for calibrated white-box membership inference. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*, pages 1605–1622, 2020.

[12] 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.

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

[14] 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.

[15] David Mickisch, Felix Assion, Florens Greßner, Wiebke Günther, and Mariele Motta. Understanding the decision boundary of deep neural networks: An empirical study. *arXiv preprint arXiv:2002.01810*, 2020.

[16] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Jonathan Uesato, and Pascal Frossard. Robustness via curvature regularization, and vice versa. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9078–9086, 2019.

[17] Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning. In *2019 ieee symposium on security and privacy*, 2019.

[18] Philipp Oberdiek, Matthias Rottmann, and Hanno Gottschalk. Classification uncertainty of deep neural networks based on gradient information. In *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*, pages 113–125. Springer, 2018.

[19] 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.

[20] 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.

[21] Congzheng Song, Thomas Ristenpart, and Vitaly Shmatikov. Machine learning models that remember too much. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 587–601, 2017.

[22] 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.

[23] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[24] 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.

[25] Kehe Wu, Zuge Chen, and Wei Li. A novel intrusion detection model for a massive network using convolutional neural networks. *IEEE Access*, 6:50850–50859, 2018.

[26] 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.