ON THE IMPORTANCE OF DIFFICULTY CALIBRATION IN MEMBERSHIP INFERENCE ATTACKS

Anonymous authors
Paper under double-blind review

ABSTRACT

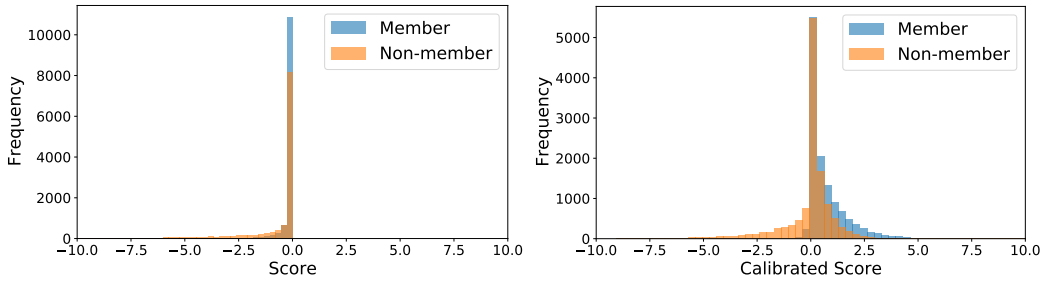
The vulnerability of machine learning models to membership inference attacks has received much attention in recent years. However, existing attacks mostly remain impractical due to having high false positive rates, where non-member samples are often erroneously predicted as members. This type of error makes the predicted membership signal unreliable, especially since most samples are non-members in real world applications. In this work, we argue that membership inference attacks can benefit drastically from *difficulty calibration*, where an attack's predicted membership score is adjusted to the difficulty of correctly classifying the target sample. We show that difficulty calibration can significantly reduce the false positive rate of a variety of existing attacks without a loss in accuracy.

1 INTRODUCTION

Modern applications of machine learning often involve training models on sensitive data such as health records and personal information. Unfortunately, recent studies have found that these models can memorize their training data to a large extent, compromising the privacy of participants in the training dataset (Fredrikson et al., 2014; 2015; Shokri et al., 2017; Carlini et al., 2019). One prominent category of privacy attacks against machine learning is the so-called *membership inference attack* (Shokri et al., 2017; Yeom et al., 2018), where the adversary aims to infer the participation of an individual in the target model's training set. Such an attack is undoubtedly damaging when the status of participation itself is considered sensitive, *e.g.*, if the training dataset consists of health records of cancer patients. Moreover, the ability to infer membership can be viewed as a lower bound for the model's degree of memorization (Yeom et al., 2018), which is useful in itself as an empirical quantifier of privacy loss (Jagielski et al., 2020; Nasr et al., 2021).

The efficacy of membership inference attacks has been improved significantly since the first attempts (Salem et al., 2018; Sablayrolles et al., 2019; Leino & Fredrikson, 2020). However, the most common evaluation metric, attack accuracy, overlooks the crucial factor of the *false positive rate* (FPR) of non-members (Rezaei & Liu, 2021). Indeed, most attacks operate by first defining a membership score $s(h, \boldsymbol{z})$ for the model h and a target input-label pair $\boldsymbol{z} = (\boldsymbol{x}, y)$ that measures how much h memorized the sample $\boldsymbol{z}$. The attack subsequently selects a threshold τ and predicts that $\boldsymbol{z}$ is a member if and only if $s(h, \boldsymbol{z}) > \tau$. For typical choices of the membership scoring function, there is usually a large overlap in the distribution of $s(h, \boldsymbol{z})$ between members and non-members (see Figure 1a). As a result, an attack that determines the membership of $\boldsymbol{z}$ by thresholding on $s(h, \boldsymbol{z})$ will have a high FPR. This drawback renders most existing attacks unreliable since the vast majority of samples likely belong to the non-member class.

In this study, we identify the lack of *difficulty calibration* as a core contributor to the high FPR of existing attacks. Specifically, a non-member sample may have a high membership score simply because it is over-represented in the data distribution. Consequently, an attack that determines a sample is likely to be a member due to having a high score will inevitably fail on these over-represented samples. To remedy this problem, we make the acute observation that if the membership score is measured *in comparison* to a typical model trained on data drawn from the same data distribution, this difference in behavior can serve as a much more reliable membership signal. Indeed, Figure 1b shows the histogram of scores $s(h, \boldsymbol{z})$ after difficulty calibration, where the member and non-member scores have significantly better separation and low FPR is now attainable.



(a) Before calibration

(b) After calibration

Figure 1: Histogram of the negative loss score (*cf.* Equation 1) before and after difficulty calibration. Without calibration, the member and non-member scores overlap significantly, and it is impossible to determine a threshold that results in low FPR. After calibration, the highest scored samples mostly belong to the member class, enabling high precision and low FPR attacks.

We propose difficulty calibration as a general technique for improving score-based membership inference attacks, and modify several membership scoring functions such as confidence (Salem et al., 2018), loss (Yeom et al., 2018), and gradient norm (Nasr et al., 2019) to construct their calibrated variants. Evaluated on a comprehensive suite of benchmark datasets, we show that calibrated attacks achieve a significantly lower FPR compared to prior work. In particular, we measure the trade-off between true positives and false positives using the *area under ROC curve* (AUC) metric, and show that difficulty calibration drastically improves this trade-off compared to uncalibrated attacks, by up to 0.10 AUC on common ML benchmarks. In addition, calibrated attacks also drastically improve the precision-recall trade-off, while remaining on-par with or better than uncalibrated attacks in terms attack accuracy. Our results suggest that it may be important for future work to apply difficulty calibration to design more reliable and practical membership inference attacks.

2 Background

Membership inference attacks are concerned with determining whether a given sample was part of a target model's training set. Homer et al. (2008) showed in their pioneering study that it is possible to infer an individual's presence in a complex genomic DNA mixture, which led to increased caution around releases of DNA data (Zerhouni & Nabel, 2008). Recent interest in member inference attacks was sparked by the work of Shokri et al. (2017), who introduced the *shadow models* method: An adversary trains substitute models (called *shadow models*) to mimic the behavior of the model under attack (called the *target model*). The adversary then observes the behavior of the shadow models when exposed to member and non-member samples, and uses this observation to train an attack meta-model for predicting membership on any given sample. Shokri et al. (2017) evaluated this attack on ML models trained on cloud APIs, and showed that the shadow models approach attains high levels of accuracy.

Score-based attacks. Yeom et al. (2018) discovered a connection between membership inference attacks and overfitting, arguing that in principle, attack accuracy can be determined by how much the target model memorizes (or overfits to) the given sample $(\boldsymbol{x}, y)$. This discovery led to a series of work on quantifying the degree of memorization via *membership scores*, which can be used for predicting that the given sample is a member when the score is high. The membership score can be computed using the loss (Yeom et al., 2018), the gradient norm (GN) (Nasr et al., 2019) or the confidence of the model's prediction (Salem et al., 2018), often yielding state-of-the-art results (Salem et al., 2018; Choquette-Choo et al., 2021). We define these scores for the cross-entropy loss ℓ:

$$s_{\text{loss}}(h, (\boldsymbol{x}, y)) = -\ell(h(\boldsymbol{x}), y) := \log(h(\boldsymbol{x})_y), \tag{1}$$

$$s_{\text{GN}}(h, (\boldsymbol{x}, y)) = -\|\nabla \ell(h(\boldsymbol{x}), y)\|_2, \tag{2}$$

$$s_{\text{confidence}}(h, (\boldsymbol{x}, y)) = -\max_{y'} \ \ell(h(\boldsymbol{x}), y') = \max_{i} \ \log(h(\boldsymbol{x})_i). \tag{3}$$

Label-only attacks. The above score-based attacks rely on continuous-valued predictions from the model in order to define the membership score. To counter these attacks, prior work considered obfuscating the model's output by returning only the top label or modifying the predicted values (Shokri et al., 2017; Jia et al., 2019). However, subsequent studies showed that even with only hard-label output, it is still possible to define membership scores that attain close to state-of-the-art accuracy (Li & Zhang, 2020; Choquette-Choo et al., 2021).

High-precision attacks. Various forms of difficulty calibration have been considered in the context of high-precision attacks. Long et al. (2018) selected samples that differ the most in loss between the target and a set of reference models, and showed that the resulting attack has high precision even for well-generalized target models. Carlini et al. (2020) showed that such privacy attacks are also possible on large-scale language models such as GPT-2 (Radford et al., 2019). Their attack operates by first generating a large number of sentences from the language model and then ranking these sentences by the (log) perplexity, with lower perplexity indicating a more plausible memorized training sample. These perplexity values are then divided by either the z-lib entropy or the perplexity given by a smaller language model to account for the sentence's rarity. In effect, only rare sentences with low perplexity can minimize the resulting score. Both attacks leverage a form of difficulty calibration by comparing the target model's loss with that of reference models, predicting membership only when the difference (or ratio) is large.

Differential privacy as a mitigation. Differential privacy (Dwork et al., 2006) (DP) is a powerful mathematical framework for privacy-preserving data analysis. A randomized algorithm $\mathcal{M}$ satisfies (ϵ, δ)-differential privacy if, given any two datasets $\mathcal{D}$ and $\mathcal{D}'$ that differ in at most one sample, and for any subset R of the output space, we have:

$$\mathbb{P}(\mathcal{M}(\mathcal{D}) \in R) \leq \exp(\epsilon)\mathbb{P}(\mathcal{M}(\mathcal{D}') \in R) + \delta. \tag{4}$$

Under mild assumptions, DP provably protects against a variety of privacy attacks, in particular membership inference (Yeom et al., 2018; Sablayrolles et al., 2019). Abadi et al. (2016) proposed a differentially private version of stochastic gradient descent (SGD), called DP-SGD, to enable differentially private training of generic ML models. Their analysis has further been refined (Mironov, 2017; Mironov et al., 2019) and has been shown experimentally to be tight (Nasr et al., 2021).

3 Difficulty Calibration

As depicted in Figure 1, prior works on score-based membership inference attack are very unreliable for separating *easy-to-predict non-members* from *hard-to-predict members* since both can attain a high membership score. We argue that a simple modification to the score, which we call *difficulty calibration*, can drastically improve the attack's reliability. This approach has been applied to the loss score for high-precision attack against well-generalized models (Long et al., 2018).

Let $s(h, (\boldsymbol{x}, y))$ be the membership score, where higher score indicates a stronger signal that the sample $(\boldsymbol{x}, y)$ is a member. Instead of computing the membership score only on the target model h, we sample multiple "typical" models trained on the same data distribution as h and evaluate the membership score on these models. Doing so calibrates $s(h, (\boldsymbol{x}, y))$ to the difficulty of the sample. For instance, if $(\boldsymbol{x}, y)$ is an easy-to-predict non-member, then $s(h, (\boldsymbol{x}, y))$ is high but typical models g would also perform well on $(\boldsymbol{x}, y)$, hence $s(g, (\boldsymbol{x}, y))$ would also be high. The small gap between $s(h, (\boldsymbol{x}, y))$ and $s(g, (\boldsymbol{x}, y))$ suggests that $(\boldsymbol{x}, y)$ is likely a non-member.

Formally, let $\mathcal{D}_{\text{shadow}}$ be a shadow dataset drawn from the same data distribution as the training set of h, and let $\mathcal{A}$ be a randomized training algorithm that samples from a distribution over models trained on $\mathcal{D}_{\text{shadow}}$. We define the calibrated score as:

$$s^{\text{cal}}(h, (\boldsymbol{x}, y)) = s(h, (\boldsymbol{x}, y)) - \mathbb{E}_{g \leftarrow \mathcal{A}(\mathcal{D}_{\text{shadow}})}[s(g, (\boldsymbol{x}, y))], \tag{5}$$

where the expectation is approximated by sampling one or more models from $\mathcal{A}(\mathcal{D}_{\text{shadow}})$. The membership inference attack proceeds by thresholding on the score $s^{\text{cal}}(h, (\boldsymbol{x}, y))$.

Efficient difficulty calibration via forgetting. Faithfully executing the difficulty calibration algorithm in Equation 5 requires training multiple models g using the randomized training algorithm

$\mathcal{A}$. This can be prohibitively expensive if the model is large, which is the case with modern neural networks such as large-scale transformers (Brown et al., 2020).

An alternative, more efficient approach if given white-box access to the target model h is to warm-start training on h. In fact, this approach establishes an explicit connection between membership inference attacks and *catastrophic forgetting* in neural networks (Goodfellow et al., 2013; Kirkpatrick et al., 2017). Toneva et al. (2018) observed that when a trained model resumes training on a separate dataset drawn from the same data distribution, it is very likely to "forget" about its original training samples. The training examples that are most likely forgotten are ones with abnormal or distinctive features, and this set of examples is stable across different training runs. We leverage this phenomenon to define a more efficient variant of difficulty calibration.

Connection to posterior inference. Under a simplifying assumption of the distribution $\mathcal{A}(\mathcal{D})$ induced by the training algorithm, it is possible to derive difficulty calibration as an approximation to an optimal white-box attack. Sablayrolles et al. (2019) proposed a formal analysis of membership inference, where they assumed that the density function of $\mathcal{A}(\mathcal{D})$ has the form:

$$p(h; \mathcal{D}) \propto e^{-\sum_{(\boldsymbol{x},y)\in\mathcal{D}} \ell(h(\boldsymbol{x}),y)}, \tag{6}$$

which can be formally derived for optimization using SGD under certain conditions (Sato & Nakagawa, 2014). The implied threat model in differential privacy assumes that the adversary has knowledge that the training dataset is either $\mathcal{D}$ or $\mathcal{D}' = \mathcal{D} \cup \{(\boldsymbol{x}, y)\}$ for some inference target $(\boldsymbol{x}, y)$, and $\mathcal{D}$ is known to the adversary. It can be shown that the Bayes-optimal attack strategy under this setting is to threshold on the calibrated score in Equation 5, where $s = s_{\text{loss}}$ and $\mathcal{D} = \mathcal{D}_{\text{shadow}}$ (Sablayrolles et al., 2019). This analysis can be generalized to any scoring function $s(h, (\boldsymbol{x}, y))$, such as gradient norm (Equation 2) and confidence (Equation 3).

4 Experiments

To demonstrate the effect of difficulty calibration, we perform a comprehensive evaluation of several score-based attacks on standard benchmark datasets. We evaluate both variants of difficulty calibration: (i) calibration by training reference models from scratch, and (ii) calibration via forgetting.

4.1 Experimental setup

Datasets. We perform experiments on several benchmark classification datasets: *German Credit*, *Hepatitis* and *Adult* datasets from the UCI Machine Learning Repository (Dua & Graff, 2017), *MNIST* (LeCun et al., 1998), *CIFAR10/100* (Krizhevsky et al., 2009), and *ImageNet* (Deng et al., 2009). The datasets vary in size from 155 to 1,281,167 points and represent both image and non-image classification tasks.

Attack setup. We split the data into two sets: a *private* set, known only to the trainer, and a *public* set, which is used for training reference models and selecting the decision threshold τ. The trainer trains their model h on half of the private set, keeping the other half as non-members. At evaluation time, a sample $(\boldsymbol{x}, y)$ is selected from the private set. The attacker computes a membership score $s(h, (\boldsymbol{x}, y))$ and predicts $(\boldsymbol{x}, y)$ as a member if $s(h, (\boldsymbol{x}, y)) > \tau$. For ImageNet, we sample 10,000 member and non-member samples from the member and non-member sets, while for all other datasets we use the full private set for evaluation. With such a setup, the accuracy of a random adversary is 50%, as is the precision at all levels of recall.

We repeat this attack under the same setup T times, each time performing a different random split into the private/public and member/non-member sets. We set $T = 5$ for all datasets except for ImageNet, where $T = 3$. Thus, in all our results, the reported averages and standard deviations reflect randomness in both model optimization and in the data splits.

Model Architectures. The target model for the German Credit, Adult and Hepatitis datasets is a multi-layer perceptron (MLP) with one hidden layer and ReLU activation followed by a softmax layer. For a dataset with m features, the hidden layer has $2m$ hidden units. For MNIST, we followed the architecture used in Leino & Fredrikson (2020), which is a small convolutional network with two

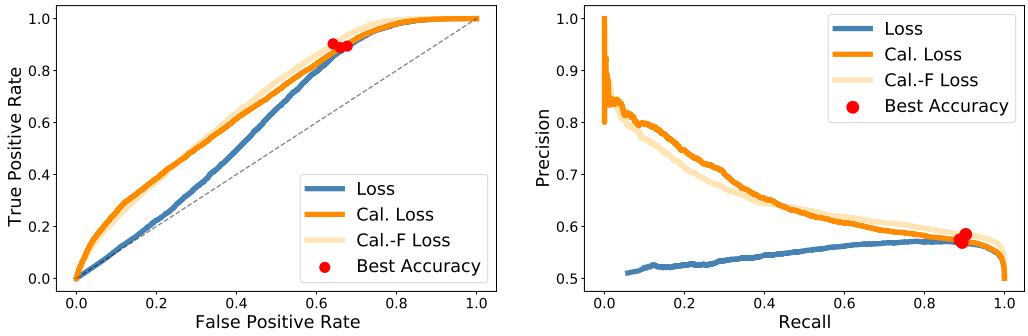


Figure 2: Left: ROC (left) and precision-recall (right) curves of calibrated/uncalibrated loss score attacks on CIFAR10. The threshold τ that optimizes accuracy is shown as a red dot. Calibration yields a higher TPR for the same value of FPR, or equivalently a higher precision at low levels of recall. The precision-recall trade-off surfaces very different behaviors for the two methods that have otherwise very similar accuracy numbers.

convolutional layers with 20 and 50 output channels, respectively, and a kernel size of 5 followed by max pooling layers. The classification head consists of one fully connected layer with 500 hidden units and ReLU activation. Dropout at rate 0.25 is applied after the the max pooling layers and at rate 0.5 after the fully connected layer. For CIFAR10 and CIFAR100, we used a CNN architecture with 4 convolutional and average pooling layers and ReLU activations using 32, 64, 64 and 128 output channels respectively with kernel size 3. For ImageNet, we used the ResNet18 (He et al., 2016) model architecture.

Model training. The target models were trained for between 50 and 200 epochs, with batch sizes varying from 4 (for very small datasets) to 1024. For optimization, we used SGD with a learning rate of 0.1, Nesterov momentum of 0.9 and a cosine learning rate schedule for the CIFAR10/100 and ImageNet datasets. Smaller datasets such as the German Credit dataset also used weight decay of 1×10^{-4}. Differentially private models were trained using the Opacus[1] library. Table 1 presents the target model accuracy statistics.

Reference models for difficulty calibration are trained using the same hyperparameters as the target model. For most experiments, we train a single reference model for calibration, and show the effect of the number of reference models in subsection 4.3. For calibration via forgetting, reference models were tuned over a set of learning rates and batch sizes to ensure that their training and test accuracies were similar to the target model for their largest number of training epochs. For both calibration methods, the reference models were trained until their validation accuracy reached its maximum value, and stopped before their training accuracy converged to its maximum. Due to its warm start, calibration-via-forgetting generally used fewer training epochs for reference models.

Task	Train Acc.	Test Acc.
Credit	0.921	0.751
Hep.	0.998	0.871
Adult	0.904	0.842
MNIST	0.990	0.989
CIFAR10	0.946	0.740
CIFAR100	0.938	0.373
ImageNet	0.773	0.637

Table 1: Average train and test accuracy statistics for the target models.

Threshold selection. When computing AUC and precision/recall metrics, we sweep over a range of values for the threshold τ and measure the resulting attack's FPR/TPR and precision/recall trade-offs. To find a threshold for optimal accuracy, we first split the public set of examples in half again, and treat one half as members, with the rest as non-members. We can then choose the best threshold τ based on membership inference accuracy for this simulated setup. Frequently, the optimal threshold for calibrated attacks was a value only slightly greater than 0 (e.g., 0.0001), which can be used as a default threshold if threshold selection is not performed.

[1] https://github.com/pytorch/opacus

Dataset		Loss			GN			Confidence		
	Gap Atk.	Orig	Cal.	Cal.-F	Orig	Cal.	Cal.-F	Orig	Cal.	Cal.-F
Credit	0.549	0.542	**0.648**	0.594	0.513	0.579	0.559	0.501	0.553	0.543
Hep.	0.535	0.514	0.558	**0.559**	0.515	0.553	0.548	0.513	0.525	0.533
Adult	0.514	0.516	**0.554**	0.529	0.510	0.523	0.515	0.507	0.510	0.512
MNIST	0.506	0.505	**0.519**	0.509	0.504	0.514	0.509	0.503	0.510	0.505
CIFAR10	0.663	0.676	0.731	0.708	0.678	**0.787**	0.762	0.629	0.635	0.631
CIFAR100	0.854	0.911	0.903	0.886	0.912	**0.924**	0.912	0.852	0.707	0.760
ImageNet	0.536	0.547	**0.557**	0.551	0.540	0.511	0.520	0.544	0.515	0.512
CIFAR10 (aug.)	0.603	0.603	0.690	0.674	0.603	0.666	**0.707**	0.560	0.568	0.575
CIFAR100 (aug.)	0.784	0.812	0.844	0.831	0.811	**0.856**	0.843	0.679	0.672	0.647

Table 2: AUC metric for score-based membership attacks before and after difficulty calibration. The standard deviation is up to 0.024 for Credit, up to 0.034 for Hep., up to 0.003 for MNIST and up to 0.015 for all other datasets. Calibration (Cal.) consistently improves the AUC of attacks by a significant margin, while calibration-via-forgetting (Cal.-F) sacrifices a modest amount of improvement for better efficiency.

Dataset		Loss			GN			Confidence		
	Gap Atk.	Orig	Cal.	Cal.-F	Orig	Cal.	Cal.-F	Orig	Cal.	Cal.-F
Credit	0.589	0.617	**0.618**	0.577	0.569	0.573	0.567	0.557	0.554	0.555
Hep.	0.561	0.574	0.575	0.591	0.574	**0.593**	0.585	0.574	0.577	0.576
Adult	0.534	**0.536**	0.534	0.522	0.518	0.518	0.515	0.512	0.511	0.511
MNIST	0.506	0.508	**0.513**	0.509	0.508	0.511	0.508	0.507	0.508	0.506
CIFAR10	0.664	0.712	0.657	0.662	0.719	**0.720**	0.702	0.642	0.623	0.622
CIFAR100	0.854	0.911	0.829	0.862	**0.915**	0.876	0.898	0.820	0.731	0.711
ImageNet	0.536	0.542	0.545	0.542	0.538	0.518	0.521	0.540	0.541	**0.553**
CIFAR10 (aug.)	0.602	0.609	0.626	0.625	0.610	0.610	**0.651**	0.562	0.542	0.558
CIFAR100 (aug.)	0.784	0.785	0.765	0.775	0.788	**0.793**	0.789	0.647	0.641	0.621

Table 3: Attack accuracy for score-based membership attacks before and after difficulty calibration. The standard deviation is up to 0.018 for Credit, up to 0.026 for Hep., up to 0.002 for MNIST and up to 0.005 for all other datasets. In almost all settings, calibrated attacks are on-par with or better than their uncalibrated version in terms of accuracy.

4.2 Main results

We first evaluate attacks that rely on continuous-valued output from the target model: loss score (Equation 1), gradient norm (GN) score (Equation 2), and confidence score (Equation 3). We also consider the gap attack (Yeom et al., 2018) as a baseline to measure progress against.

AUC and accuracy. We compute two primary metrics: area under ROC curve (AUC) for evaluating the FPR/TPR trade-off, and attack accuracy. Table 2 shows the computed AUC for various attacks before and after difficulty calibration. Overall, calibrated attacks often perform drastically better than their uncalibrated counterparts, and in some instances the performance increase can exceed 0.10, such as on the German Credit and CIFAR10 datasets. Table 3 shows the corresponding attack accuracy numbers. Evidently, calibration does not reduce accuracy in most cases, and sometimes brings a modest improvement.

We observe that the confidence score attacks under-perform other methods due to the model predicting non-member points incorrectly with high confidence. For this reason, calibrated confidence did not improve the attack performance on many datasets, as training points that were correctly classified on the target model but were incorrectly classified on the reference model became indistinguishable.

Effect of data augmentation. The last two rows in Tables 2 and 3 show various attacks' performance on the CIFAR10/100 target model trained *with* data augmentation. As expected, attacks are less effective on models with data augmentation due to reduced overfitting. Nevertheless, calibration still improves both the AUC and accuracy of loss and gradient norm based attacks most of the time.

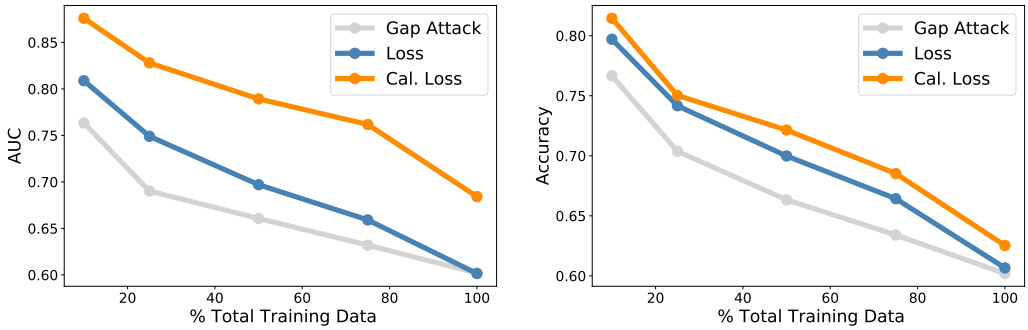


Figure 3: AUC (left) and accuracy (right) of the gap attack and calibrated/uncalibrated loss score attacks against target model trained on varying training set sizes on CIFAR10. Both AUC and accuracy increase as the training dataset size decreases due to more severe overfitting of the target model. Difficulty calibration can effectively leverage this phenomenon to improve the attack's AUC and accuracy across the entire range of training set sizes.

Privacy	AUC			Accuracy		
	Orig.	Cal.	Cal.-F	Orig.	Cal.	Cal.-F
Credit	0.552	**0.640**	0.599	0.616	0.620	**0.638**
Hep.	0.547	0.585	**0.594**	**0.596**	0.588	0.591
Adult	0.512	**0.541**	0.537	0.533	**0.549**	0.539
CIFAR10	0.660	**0.686**	0.620	0.665	**0.666**	0.623
CIFAR100	0.849	0.841	**0.856**	**0.847**	0.841	0.844

Table 4: Average accuracy and AUC of the HopSkipJump label-only attack using $\max(n, 1000)$ datapoints for each of the training and heldout sets, where n is their original datasize. CIFAR models were trained without data augmentation.

ROC and precision-recall curves. Figure 2 shows the ROC and precision-recall curves for the loss score attack on CIFAR10, which demonstrates the effect of difficulty calibration in closer detail. In the left plot, at low FPR values, the uncalibrated attack has close to zero advantage in TPR, behaving essentially the same as random guessing. In stark contrast, calibrated attacks exhibit a large difference in the FPR and TPR values. Furthermore, as shown in the right plot, calibrated attacks can detect a group of member points with very high precision at low recall values, which is not possible for the uncalibrated counterpart. This pattern can also be clearly seen in the membership score distributions of calibrated and uncalibrated attacks shown in Figure 1.

We highlight another interesting phenomenon by computing the threshold τ that achieves optimal accuracy for each attack, shown by the red dot in Figure 2. For all attacks, this threshold lies towards the tail of the ROC and precision-recall curves, where the attack achieves high FPR and high recall but low precision. This observation concurs with prior work, which shows that existing attacks do not predict membership reliably when optimizing for accuracy (Rezaei & Liu, 2021).

Label-only attacks. When the target model returns only a discrete label rather than a continuous-valued output, membership inference can still be performed using label-only attacks. We show that difficulty calibration also improves such attacks. Table 4 reports the performance of calibrated and uncalibrated label-only attacks using the HopSkipJump boundary distance method (Choquette-Choo et al., 2021). We excluded the MNIST and ImageNet datasets as the HopSkipJump attack required too many queries to successfully change the label. Similar to the results in Tables 2 and 3, label-only attacks also benefit from difficulty calibration with increased AUC and improved or similar accuracy. However, the performance increase is noticeably smaller for CIFAR10/100 compared to other score-based attacks.

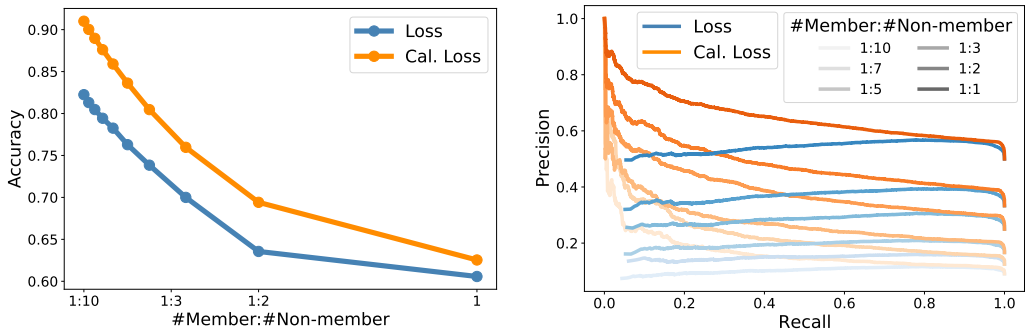


Figure 4: Accuracy (left) and precision-recall curve (right) of the loss score attack when changing the member to non-member ratio. At low ratios, the calibrated attack can still identify members with high precision, whereas the uncalibrated attack is unable to do so despite high accuracy.

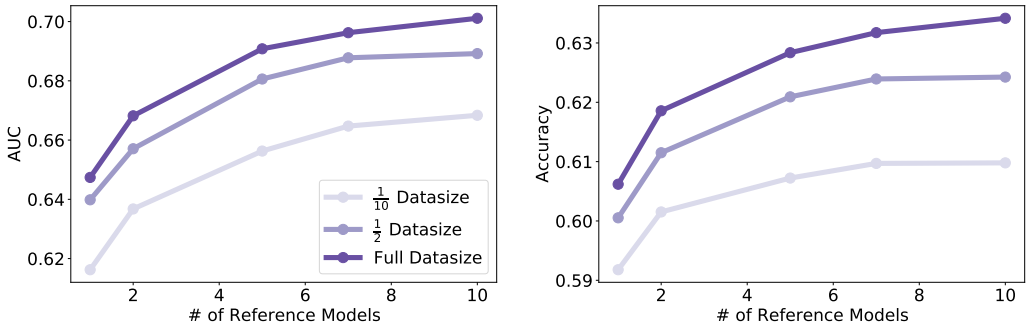


Figure 5: AUC (left) and attack accuracy (right) for the calibrated loss attack with varying number of reference models on CIFAR-10. Using more reference models and training reference models with a higher shadow dataset size strictly improve performance at a cost of more computation.

4.3 Ablation Studies

Training set size. Since vulnerability to membership inference attack is linked to overfitting (Yeom et al., 2018), training set size for the target model is a crucial factor in attack performance. Indeed, Figure 3 shows that for the gap attack and calibrated/uncalibrated loss score attacks, both AUC and accuracy increase when the training set size is reduced. More importantly, difficulty calibration can effectively leverage overfitting to improve the attack's performance across the entire range of training set sizes.

Ratio of member to non-member samples. The experimental setup so far uses an equal number of member and non-member samples. In real world scenarios, the majority of samples are likely non-members, which drastically affects the accuracy and precision-recall trade-off of membership inference attacks. In the following experiment, we simulate this scenario by subsampling the member set for evaluation to decrease the ratio of member to non-member samples.

Figure 4 shows the accuracy and precision-recall curves for different member to non-member ratios. In the left plot, as the number of non-member samples increases, attack accuracy also drastically increases. This is due to the threshold τ for optimal accuracy favoring high recall and low precision (see Figure 2), effectively performing "non-membership inference". In contrast, the right plot shows that precision-recall trade-off worsens when the number of non-member samples increases, which reflects the fact that attacks are less reliable in the real world when most samples are non-members. Nevertheless, for the calibrated attack, it is still possible to achieve a high precision at low values of recall even for low member to non-member ratios.

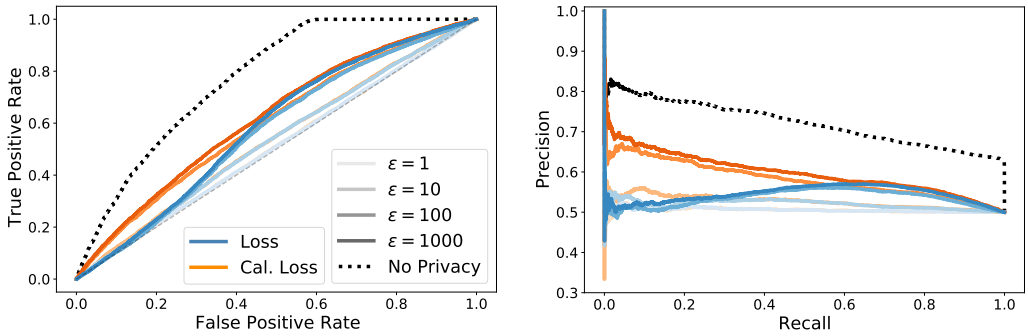


Figure 6: ROC (left) and precision-recall (right) curves for calibrated/uncalibrated loss score attack on CIFAR10 against differentially private models. For large values of DP parameter ϵ, the calibrated attack can still successfully infer member samples with high precision at low recall.

Number of reference models. Results reported in Section 4.2 use a single reference model for difficulty calibration. Figure 5 shows the effect of increasing the number of reference models on AUC and attack accuracy. To train each reference model, we subsample a different subset of the public dataset without replacement to create the shadow dataset. We then apply calibration-via-forgetting to calibrate the loss score, while varying the shadow dataset size. Increasing the number of reference models and increasing the shadow dataset size strictly improve both AUC and accuracy. The difference between 1 and 10 reference models is up to 0.05 for AUC and 0.03 for accuracy.

Differential privacy is by far the most common defense against membership inference attacks, and provably limits the accuracy of these attacks (Yeom et al., 2018) when ϵ is small. With a higher value of ϵ, DP still empirically protects against uncalibrated membership inference attacks *on average*; however, calibrated attacks provide a significant edge against outliers for high values of the privacy parameter ϵ.

Figure 6 shows the ROC and precision-recall curves of calibrated/uncalibrated loss score attacks on CIFAR10 models trained with DP-SGD. As expected, DP clearly reduces the effectiveness of attacks as indicated by the gap between attacks against models with finite ϵ (blue and orange lines) and attacks against an undefended model (dotted line). However, for $\epsilon = 100$ and 1000, calibrated attacks can still attain a non-trivial FPR-TPR and precision-recall trade-off. Only when ϵ gets close to 1 do the AUC and accuracy approach that of uncalibrated attacks, and both attacks converge to the effectiveness of random guessing.

A qualitative inspection of member images from CIFAR10 that are successfully inferred at a high precision of > 0.9 for the DP model at $\epsilon = 1000$, but are not identified for $\epsilon \leq 10$, showed that vulnerable images were often atypical samples from their respective classes. They did not appear to be extreme outliers. A further manual inspection of examples did not suggest any general pattern as to which ones are most vulnerable under differential privacy.

5 Conclusion

We showed that difficulty calibration is a powerful post-processing technique for improving existing membership inference attacks. Most score-based attacks can greatly benefit from calibration, making the attack much more reliable in low false positive rate and high precision regimes. Our study also showed that prior work that focused on evaluating attack accuracy do not reveal the full picture, and detailed analysis using the ROC and precision-recall curves is crucial for designing reliable membership inference attacks.

When given white-box access to the target model, calibration via forgetting can be an efficient alternative to training reference models from scratch. This alternative method also serves as an explicit connection between membership inference and the well-studied phenomenon of catastrophic forgetting in neural networks. We hope that future work can further explore this connection and leverage new findings to improve the state-of-the-art in membership inference attacks.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *CCS*, 2016.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

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*, pp. 267–284, 2019.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. *arXiv preprint arXiv:2012.07805*, 2020.

Christopher A Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. Label-only membership inference attacks. In *International Conference on Machine Learning*, pp. 1964–1974. PMLR, 2021.

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*, pp. 248–255. Ieee, 2009.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pp. 1322–1333, 2015.

Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In *23rd USENIX Security Symposium*, pp. 17–32, 2014.

Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013.

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*, pp. 770–778, 2016.

Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V Pearson, Dietrich A Stephan, Stanley F Nelson, and David W Craig. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS genetics*, 4(8):e1000167, 2008.

Matthew Jagielski, Jonathan Ullman, and Alina Oprea. Auditing differentially private machine learning: How private is private SGD? *arXiv preprint arXiv:2006.07709*, 2020.

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*, pp. 259–274, 2019.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Klas Leino and Matt Fredrikson. Stolen memories: Leveraging model memorization for calibrated white-box membership inference. In *29th USENIX Security Symposium*, pp. 1605–1622, 2020.

Zheng Li and Yang Zhang. Label-leaks: Membership inference attack with label. *arXiv e-prints*, pp. arXiv–2007, 2020.

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.

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pp. 263–275. IEEE, 2017.

Ilya Mironov, Kunal Talwar, and Li Zhang. Rényi differential privacy of the sampled gaussian mechanism. *arXiv preprint arXiv:1908.10530*, 2019.

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)*, pp. 739–753. IEEE, 2019.

Milad Nasr, Shuang Song, Abhradeep Thakurta, Nicolas Papernot, and Nicholas Carlini. Adversary instantiation: Lower bounds for differentially private machine learning. *arXiv preprint arXiv:2101.04535*, 2021.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Shahbaz Rezaei and Xin Liu. On the difficulty of membership inference attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7892–7900, 2021.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Hervé Jégou. White-box vs black-box: Bayes optimal strategies for membership inference. In *International Conference on Machine Learning*, pp. 5558–5567. PMLR, 2019.

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.

Issei Sato and Hiroshi Nakagawa. Approximation analysis of stochastic gradient Langevin dynamics by using Fokker-Planck equation and Ito process. In *International Conference on Machine Learning*, pp. 982–990. PMLR, 2014.

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)*, pp. 3–18. IEEE, 2017.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. An empirical study of example forgetting during deep neural network learning. *arXiv preprint arXiv:1812.05159*, 2018.

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)*, pp. 268–282. IEEE, 2018.

Elias A Zerhouni and Elizabeth G Nabel. Protecting aggregate genomic data. *Science*, 2008.