Balancing Privacy and Attack Utility: Calibrating Sample Difficulty for Membership Inference Attacks in Transfer Learning

Shuwen Liu[1], Yongfeng Qian[1,*], Yixue Hao[2]
[1]China University of Geosciences, [2] Huazhong University of Science and Technology
{Liu_shuwen,yfqian}@cug.edu.cn
yixuehao@hust.edu.cn

***Abstract*—The growing prominence of transfer learning in domains such as healthcare and finance highlights its efficacy in enhancing machine learning models. However, conventional membership inference attacks (MIA) often struggle to perform well when applied to transfer learning models trained under normal fit. To address this challenge, we propose a novel approach called *PC-MIA*. This approach involves generating multiple poisoned reference models using toxic samples. These poisoned models are then utilized to calibrate the difficulty of samples and reveal their true hardness, thereby enhancing the accuracy of MIA. Through empirical evaluations conducted on real-world datasets and employing diverse model architectures, our approach demonstrates its ability to significantly improve the accuracy of membership inference.**

***Index Terms*—membership inference attack, data poisoning attack, difficulty calibration**

I. INTRODUCTION

The rise of transfer learning and MLaaS has facilitated the emergence of third-party platforms, leveraging pre-trained models for accelerated training and cost reduction. However, this convenience exposes vulnerabilities to malicious actors injecting toxic data, necessitating a balance between data membership disclosure and prediction accuracy maintenance, especially at low poisoning rates. However, this convenience also presents an opportunity for attackers, who can inject toxic data to undermine model accuracy or compromise user privacy. Therefore, balancing data disclosure and model accuracy in low poisoning rates poses a key challenge for obtaining data privacy through integrity compromise.

Difficulty calibration in MIA distinguishes members from non-members but incurs high data costs due to separate training datasets for each reference model. Although it improves true positive rates (TPR) at low false positive rates (FPR), it often diminishes the differentiation between easy-to-predict and hard-to-predict non-members, resulting in lower accuracy for non-members. To address these challenges, we propose *PC-MIA*, a novel method that leverages difficulty calibration via data poisoning attacks to enhance the disclosure of target class membership information. Through clean-label poisoning attacks, we strategically inject toxic samples while carefully managing their quantity to conceal any discrepancies in test accuracy. Furthermore, the poisoned model exhibits enhanced accuracy in identifying non-members, as the attack constrains

*is the corresponding author.

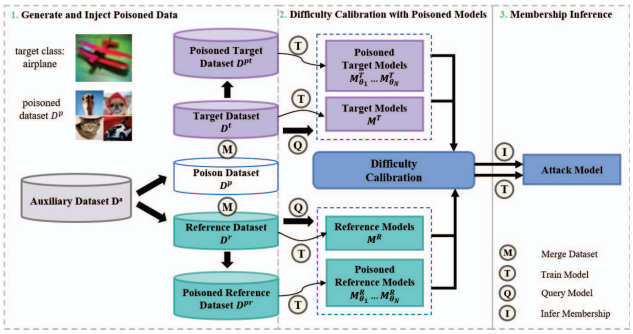

Fig. 1. Workflow overview.

decision-making boundaries for the target class. This facilitates adaptation to data hardness and enables the attack model to accurately infer membership using calibrated loss scores from the reference model.

Our contributions are threefold. 1) Our method significantly improves the accuracy of MIA targeting specific classes, while simultaneously reducing the requirement for auxiliary data to generate reference models. 2) *PC-MIA* ensures that the test accuracy of the poisoned model on the target class remains essentially identical to that of the overall class. This guarantees that the attacker can obtain membership information of the target class without detection, enabling a silent attack. 3) Through evaluations on three datasets and three target architectures, *PC-MIA* demonstrates superior performance in practical settings compared to the original baseline.

II. DESIGN OVERVIEW

Design Overview: The primary objective of *PC-MIA* is to leverage the poisoning model to calibrate the difficulty of target class members, thereby elevating the risk of sensitive membership information leakage from the target class. In this paper, we propose a novel method, *PC-MIA*, to assess sample hardness based on resilience to poisoning, illustrated in Figure 1. We first create a dedicated poisoning dataset D^p using a clean-label poisoning attack [3]. Then, by selecting different-sized subsets from D^p and merging them with datasets used to train a clean reference model M^R, we obtain multiple contaminated poisoning datasets. Finally, we train multiple poisoned reference models $M^R_{\theta_1}...M^R_{\theta_N}$ from these datasets. In the next step, we evaluate all the clean and poisoned reference

2833-292X/24/$31.00 ©2024 IEEE
DOI 10.1109/DSN-S60304.2024.00046

models using dataset D^r, and then aggregate their respective output losses to generate a new, difficulty-calibrated member score. Finally, we obtain the calibrated score of dataset D^r on the reference models, align it with the corresponding member labels, and utilize it as the training dataset for the attack model. Subsequently, we adopt the same strategy to evaluate the hardness of the target dataset D and integrate the calibrated member scores into the trained attack model for membership inference. In the following content, we will provide details of the selected modules.

Generate and Inject Poisoned Data: The objective of this module is to enable poisoned reference and target models to disclose membership information. Previous research has demonstrated that data poisoning attacks can exacerbate privacy breaches, potentially leading to substantial discrepancies in test accuracy between the target class and the overall class. This disparity raises concerns about the detection of poisoning attacks. To mitigate this concern, we discreetly reduced the proportion of poisoned samples to 2%, strategically ensuring that the successful leakage of target class membership information and our attack remain undetected. To compensate for the limited privacy breach resulting from low poisoning rates, we introduce a method for calibrating sample difficulty using poisoning models, which will be further elaborated in the following paragraph.

Difficulty Calibration with Poisoned Models: Traditional difficulty calibration methods usually involve training several reference models on subsets of data with either identical or disjoint distributions. However, these traditional approaches result in an exponential increase in the size of the auxiliary dataset as the number of reference models grows. To address this challenge, we adopt reference models that utilize a shared subset of clean data alongside a pool of poisoned samples. Each reference model selects a distinct number of poisoned samples from this pool, yielding multiple versions capable of evaluating sample difficulty.

III. PRELIMINARY EVALUATIONS

We compare our *PC-MIA* method with baseline attacks [1] and [2], assessing MIA accuracy on target class samples. Difficulty calibration, based on target and reference model sample differences, is conducted at a 2% poisoning rate (*PC-MIA* method 1) for equitable comparison. Subsequently, employing the new difficulty calibration method (*PC-MIA* method 2), we utilize the loss sequence of the samples on both clean and poisoned models as the calibrated hardness value for subsequent accuracy testing. Additionally, we compel malicious third-party platforms to inject toxic data into the target model before publication, thereby diminishing the attacker's efficacy in our approach and further ensuring fairness.

Our two proposed methods outperform the baselines across all datasets and model architectures, as shown in Table I, with an attack accuracy difference of up to 0.25, demonstrating the effectiveness of our *PC-MIA* approach in inducing significant membership privacy leakage. In *PC-MIA* method 2, attack accuracy generally increases with the length of the loss sequence, as shown in Figure 2. Employing three poisoning models for difficulty calibration resulted in an average attack accuracy increase of 0.054. Furthermore, the data overhead of the comparison method [2] increases as the reference model size grows. In contrast, *PC-MIA* only requires the dataset D^r for a clean reference model and a fixed-sized pool of poisoning data D^p as auxiliary datasets. For example, while [2] requires $2n$ auxiliary data, *PC-MIA* only needs 20,900, assuming both the training and test sets of the target model contain 10,000 samples each, and the reference model is of size n.

TABLE I
ATTACK PERFORMANCE.

	Attack Acc			
	Yeom et al. [2]	Watson et al. [1]	*PC-MIA* method 1	*PC-MIA* method 2
CIFAR-10				
Xception	0.6635	0.6135	**0.7720**	0.7560
Resnet50	0.4915	0.5130	0.6165	**0.6195**
VGG16	0.6655	0.6285	**0.7690**	0.7600
STL-10				
Xception	0.6066	0.5983	**0.7466**	0.7400
Resnet50	0.6100	0.4666	0.5366	**0.7166**
VGG16	0.6316	0.5383	**0.7283**	0.7183
CelebA				
Xception	0.6532	0.6304	0.6732	**0.6891**
Resnet50	0.6898	0.6350	0.7192	**0.7271**
VGG16	0.6949	0.6333	0.7469	**0.7589**

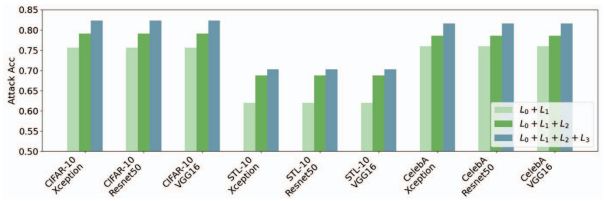

Fig. 2. Attack performance under different loss numbers in the loss sequence. L_0 represents the loss of the sample on the clean model, while L_1, L_2, and L_3 represent the loss of the sample on the poisoning models, each with a different poisoning rate.

IV. CONCLUSION

This study addressed difficulty calibration in MIA and proposed a novel approach using data poisoning attacks to enhance the disclosure of target class membership information. Future research could refine calibration techniques, improve attack model accuracy, and develop robust defense mechanisms.

REFERENCES

[1] L. Watson, C. Guo, G. Cormode, and A. Sablayrolles, "On the Importance of Difficulty Calibration in Membership Inference Attacks," *CoRR*, vol. abs/2111.08440, 2021.

[2] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting," in *2018 IEEE 31st computer security foundations symposium (CSF)*, pp. 268-282, 2018.

[3] Y. Chen, C. Shen, Y. Shen, C. Wang, and Y. Zhang, "Amplifying Membership Exposure via Data Poisoning," *in Advances in Neural Information Processing Systems*, vol. 35, Inc., 2022, pp. 29830-29844.