Unsupervised Membership Inference Attacks Against Machine Learning Models

Yuefeng Peng
School of Cyber Science and Engineering
Wuhan University
yuefengpeng@whu.edu.cn

Bo Zhao
School of Cyber Science and Engineering
Wuhan University
zhaobo@whu.edu.cn

Hui Liu
School of Cyber Science and Engineering
Wuhan University
liuh824@whu.edu.cn

Yang An
School of Computer Science
Wuhan University
yangan@whu.edu.cn

Abstract

As a form of privacy leakage for machine learning (ML), membership inference (MI) attacks aim to infer whether given data samples have been used to train a target ML model. Existing state-of-the-art MI attacks in black-box settings adopt a so-called shadow model to perform transfer attacks. Such attacks achieve high inference accuracy but have many adversarial assumptions, such as having a dataset from the same distribution as the target model's training data and knowledge of the target model structure. We propose a novel MI attack, called UMIA, which probes the target model in an unsupervised way without any shadow model. We relax all the adversarial assumptions above, demonstrating that MI attacks are applicable without any knowledge about the target model and its training set. We empirically show that, with far fewer adversarial assumptions and computational resources, UMIA can perform on bar with the state-of-the-art supervised MI attack.

1 Introduction

Recent researches have shown ML models memorize sensitive information of training data [1, 2, 3], making them susceptible to membership inference (MI) attacks [4, 5, 6]. In MI attacks, the adversary's goal is to determine whether given data points were used to train the target model. Since ML models are usually trained on sensitive data such as medical records [7, 8] and facial images [9, 10], the success of MI can lead to severe consequences. For example, the existence of a patient's medical record in a hospital's analytical training set reveals that the patient was once a patient there.

MI attack can be regarded as a binary classification problem: classify the given data points into members and non-members using a binary classifier. Shokri et al. [4] propose the first MI against ML models, exploiting the differences in the target model's prediction vectors on the members versus the non-members. The main challenge for such attacks is that the adversary needs to collect enough samples labeled as either members or non-members to predict membership for a new data sample with unknown membership. Specifically, they generate a dataset based on data samples with known

Preprint. Under review.

Table 1: **A comparison to prior works.** ✓means the information is required by the adversary, - otherwise.

source	No.shadow models	Target model structure	Target model's training data distribution	Label knowledge	Inference accuracy
[4]	multiple	✓	✓	✓	high
[6]	1	-	✓	-	high
[6]	1	-	-	-	unstable
[6]	-	-	-	-	low
[5]	-	-	-	✓	low
Ours	-	-	-	-	high

membership to train a ML model, referred to as an attack model, and then use the attack model to classify a new data sample with unknown membership. However, when only given black-box access (i.e., having access to the output probability distribution), the adversary cannot derive enough samples with known membership of the target model's training set. To solve this problem, Shokri et al. [4] present the shadow training technique, which creates multiple so-called shadow models to mimic the behavior of the target model. Since the adversary trains the shadow models, they know each data sample's membership of the shadow models' training set and can thus construct a dataset to train an attack model to perform transfer attacks. However, as shown by Salem et al. [6] and confirmed by our experiments, when the shadow model is drastically different from the target model, the attack performance is not promising. Creating shadow models of high quality, on the other hand, have many requirements such as knowledge of the target model's structure and training set.

Researchers also propose some attacks without any shadow model, using simple decision rules instead of binary ML classifiers to predict membership [6, 5]. However, these methods cannot make an effective inference due to the lack of samples labeled with ground truth membership information. As shown in Table 1, existing attacks either make too many assumptions on the adversary or do not perform that well.

In this paper, we introduce a novel MI attack, called UMIA, which achieves high inference accuracy without shadow models. Specifically, given a batch of samples with unknown membership, UMIA first extracts membership semantics via temperature scaling [11, 12], and then uses clustering algorithms to divide these samples into members and non-members. We compare UMIA with the state-of-the-art attacks. We empirically show that with far fewer adversarial assumptions and computational resource, UMIA achieves similar inference accuracy to the attack with shadow models [6]. Our code is available online [1].

2 Threat Model

Our attack is a batch attack. The threat model assumes an adversary trying to infer whether each data point in a given dataset D', called the target dataset, belongs to the target model h's training set D. Specifically, the adversary has access to a dataset D' which partially overlaps with the h's training set D. However, the adversary does not know which and how many data points are in $D' \cap D$. The adversary's goal is to infer which data points are in $D' \cap D$. We assume the adversary only has black-box access to h, i.e., the adversary can only query the target model h with samples to obtain the prediction vectors of output classes. Under strong adversarial assumptions, prior works have proposed high-performance MI attacks. In this paper, we aim to show that, in the absence of these conditions, the adversary can still achieve similar MI accuracy via the unsupervised method. As prior works did, we also assume black-box access to h. The difference is that we do not make any assumptions related to h and its training data.

3 Methodology

3.1 Attack Framework

UMIA exploits the fact that output probability distributions of the target model h may vary between members and non-members. Fig. 1 illustrates the framework of UMIA. Specifically, given a batch of

[1] https://github.com/elpx16443/umia

target data samples, UMIA first obtains the outputs of them by querying h, then processes the output probabilities to extract membership information, and finally uses clustering to divide these processed output probabilities into two clusters, which represent members and non-members respectively. We introduce the details below.

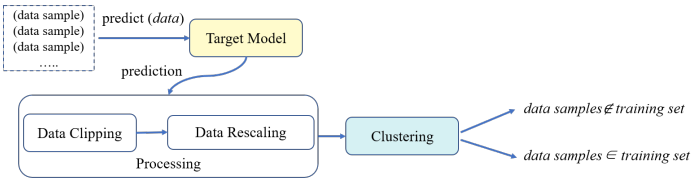

Figure 1: The framework of UMIA.

3.2 Processing

As mentioned above, given any data sample $\mathbf{x}$, UMIA first extract membership semantics from its output probability distribution $h(\mathbf{x})$ given by h. As shown by Salem et al. [6] and confirmed by our experiments, the class types (e.g., a bird or a car in CIFAR-10), are not important for MI, but the ranking of scores in the output probabilities matters. Moreover, using full output probability is unnecessary, and using the biggest scores in the output probability suffices. Based on this idea, after get $h(\mathbf{x})$ for each sample $\mathbf{x}$, we select top-k values in $h(\mathbf{x})$, denoted as $h_k(\mathbf{x})$, removing noisy ones with small values.

Then, we rescale $h_k(\mathbf{x})$ using temperature scaling [11, 13, 14]. The high-level idea is as follows. ML models tend to assign a high probability score to the correct class and assign small values to other classes. For example, for the MNIST dataset, one correctly classified picture of a 2 may be given a probability over 0.99 of being a 2 and a probability less than 0.01 of being others. Despite containing useful membership information, the small values in the output probability have very little influence on the following clustering stage. To address the problem, we rescale the output probability to improve their influence in MI. Note that temperature scaling is also used in knowledge distillation for a similar purpose [11, 15]. The rescaled output probability $R(x;T)$ is computed according to the following equation:

$$R_i(h_k(\mathbf{x});T) = \frac{exp(log(h_k^i(\mathbf{x}))/T)}{\sum_j exp(log(h_k^j(\mathbf{x}))/T))} \tag{1}$$

Here, $h_k^i(\mathbf{x})$ is the ith score in $h_k(\mathbf{x})$, and T is a temperature scaling parameter.

3.3 Clustering

After rescaling, we apply the K-means clustering algorithm to divide these processed output probabilities into two clusters. Samples in the cluster with higher average confidence scores are labeled as members, and others are labeled as non-members, based on our observation that ML models tend to output higher confidence scores for data points on which they trained.

4 Evaluation

Experimental setup. We use seven datasets for evaluation: MNIST [2], CIFAR-10 [16], CIFAR-100 [16], Purchase [4], Location[4], Texas [4] and UCI Adult [3]. For each dataset, we use its corresponding model architecture that is consistent with prior works [4, 6]. We compare UMIA against the attack by Salme et al. [6]., following the original configuration of the authors' code [4].

Results. We first assume the adversary has full knowledge of the target model structure and same data distribution. Such setting allows the adversary to train a high-quality shadow model but does not benefit UMIA because it does not need any shadow model. As depeicted in Fig.2, although the most

[2]http://yann.lecun.com/exdb/mnist/

[3]http://archive.ics.uci.edu/ml/datasets/Adult

[4]https://github.com/AhmedSalem2/ML-Leaks

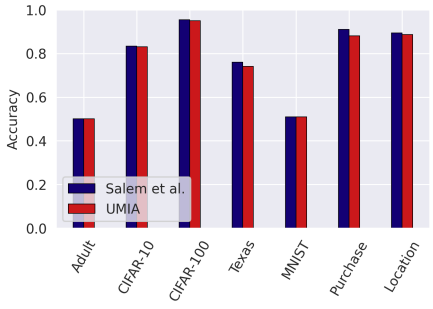

Figure 2: Comparison of UMIA with attacks with shadow models.

Table 2: Accuracy of MI attacks on Purchase dataset.

Classifier	Salem et al.			UMIA
	Neural Network	Logistic Regression	Random Forests	
Neural Network	0.91	0.80	0.22 - 0.67	0.88
Logistic Regression	0.55	0.89	0.53 - 0.80	0.88
Random Forests	0.5	0.53 - 0.62	1.0	1.0

favorable setting is adopted for the attack by Salem el at., UMIA still achieves comparable inference accuracy.

Then, we assume the adversary has no knowledge of the target models including their types. In this setting, for the attack with shadow models, we use different ML classifiers such as Random Forests as the shadow model and attack the target that is the same or different from the shadow, such as a neural network. Table 2 depicts the results. We observe that only when the shadow model's structure is the same as its corresponding target model can the attack with shadow models achieve high performance. On the other hand, UMIA provides high inference accuracy because it performs MI by directly probing the target.

Finally, we demonstrate the attack efficiency of UMIA. Since MI attacks are widely used to evaluate the privacy leakage of the training sets of ML models [17, 18] and modern ML models are usually trained on big datasets, the attack efficiency is important for a good MI. The experimental result is shown in Table 3. The average running time of UMIA is about 320 times faster than the attack with shadow models. The experimental result demonstrates that UMIA is lightweight and efficient.

5 Conclusion

Despite the great success achieved, ML models are vulnerable to MI attacks, which raises severe privacy risks. The existing state-of-the-art MI attacks make many assumptions on the adversary, such as having knowledge of the target model's structure and training set. we relax these assumptions by introducing unsupervised MI attacks, UMIA, leveraging temperature scaling and unsupervised ML algorithms. Our evaluation on various datasets shows that UMIA has a comparable performance with the state-of-the-art attacks [6]. We demonstrate that the adversary can perform high accuracy MI attacks at low cost in a broader range of scenarios.

Table 3: Running time (in seconds) of MI attacks on various datasets.

Dataset	Salem et al.			UMIA Total Seconds
	Training Shadow Models	Training Attack Models	Total Seconds	
CIFAR-10	23.14	24.94	48.08	**0.13**
CIFAR-100	23.65	24.62	48.27	**0.09**
Purchase	2.87	25.11	27.98	**0.10**
Location	1.05	3.49	4.54	**0.04**
MNIST	17.50	24.61	42.11	**0.14**
Texas	35.73	25.29	61.02	**0.18**
Adult	3.76	24.23	27.99	**0.13**

References

[1] 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.

[2] 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 (CCS 17)*, page 587–601, 2017.

[3] 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.

[4] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18, 2017.

[5] 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)*, pages 268–282, 2018.

[6] Ahmed Salem, Yang Zhang, Mathias Humbert, Mario Fritz, and Michael Backes. Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. In *Network and Distributed Systems Security (NDSS) Symposium*, 2019.

[7] Bradley J. Erickson, Panagiotis Korfiatis, Zeynettin Akkus, and Timothy L. Kline. Machine learning for medical imaging. *RadioGraphics*, 37(2):505–515, 2017.

[8] Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, and Dimitrios I. Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13:8 – 17, 2015.

[9] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1701–1708, 2014.

[10] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[11] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[12] Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017.

[13] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 06–11 Aug 2017.

[14] Meelis Kull, Miquel Perello Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[15] Shangchen Du, Shan You, Xiaojie Li, Jianlong Wu, Fei Wang, Chen Qian, and Changshui Zhang. Agree to disagree: Adaptive ensemble knowledge distillation in gradient space. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12345–12355. Curran Associates, Inc., 2020.

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

[17] 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)*, pages 739–753, 2019.

[18] Liwei Song and Prateek Mittal. Systematic evaluation of privacy risks of machine learning models. In *30th USENIX Security Symposium (USENIX Security 21)*, 2021.