MEMBERSHIP INFERENCE ATTACKS

Adarsh Jamadandi
Saarland University
adarsh.jam@gmail.com

ABSTRACT

In this report I will be examining two works (Carlini et al., 2021) and (Zarifzadeh et al., 2024) which propose membership inference attack (MIA) based on likelihood ratio tests. The goal is to understand how these methods are similar or different to each other.

1 WHAT IS MEMBERSHIP INFERENCE ATTACK?

Given a target model $f_t(\theta)$ trained by standard gradient based method. An adversary is interested in determining if a certain data point was used in training the target model. This kind of model audit by a malicious party might be undesirable especially when the model has been trained on sensitive data such as medical information. This act of determining if a certain data point was part of the training set is called Membership Inference. The adversary has access to the target model but might not have access to the knowledge of how that model was trained, that is, the learning rate, the number of epochs etc are all unknown. The adversary has access to the samples from the trained data and is allowed to query the model to discern if the data point was included in the training set. In this report, we will closely examine two works (Carlini et al., 2021) and (Zarifzadeh et al., 2024) that propose likelihood ratio test based membership inference attacks. The goal is to understand how similar or different both of these works are.

2 LIRA

2.1 WHY DO WE NEED LIRA?

The main contribution of the paper by Carlini et al. (2021) is a proposal to move away from average-case accuracy metrics and posit true-positive rate at low false-positive rates as a valid metric to evaluate membership inference attacks (MIA). The authors show through a comprehensive set of experiments that methods such as (Shokri et al., 2016; Yeom et al., 2018; Sablayrolles et al., 2019) might over-estimate their effectiveness because their evaluation hinges on average-case accuracy. For example, consider training a ResNet (He et al., 2015) on CIFAR-10 (Krizhevsky, 2009) and interrogating the model by using the LOSS attack proposed in (Yeom et al., 2018). The loss attack essentially says, data points which were part of the training set will have a lower loss

$$\mathcal{A}_{loss}(x, y) = \mathbf{1}[-l(f(x), y) > \tau] \tag{1}$$

After the LOSS attack is carried out, we can evaluate if the attack was successful by measuring the balanced attack accuracy given by

$$\Pr_{x,y,f,b}[\mathcal{A}^{D,f}(x, y) = b] \tag{2}$$

The problem with this evaluation strategy as the authors note is that as the name itself suggests this is a balanced accuracy metric that weighs both false-positives and false-negatives equally which is not required since the adversary is more interested in minimizing the false-positives so the attack causes maximum damage. The authors find that for samples which had the lowest loss values $l(f(x), y)$ for the CIFAR-10 dataset, which according to the evaluation strategy suggest it should be most effective, in reality it is only right 48% of the time and in contrast the higher loss samples had 100% correctness all the time.

2.2 How LiRA works

With extensive experiments, the authors slowly dismantle the existing evaluation strategies and propose a new MIA method that uses hypothesis testing. The idea is, the adversary wants to perform counterfactuals (I use that term loosely here, since we don't actually do any Causal Inference), that is, what would be the outcome if a certain data point was used in the training and wasn't? The hypothesis testing method involves finding two distributions $\mathbb{Q}_{in}(x, y) = \{f \leftarrow \mathcal{T}(D \cup \{(x, y)\} | D \leftarrow \mathbb{D})\}$ and $\mathbb{Q}_{out}(x, y) = \{f \leftarrow \mathcal{T}(D \backslash \{(x, y)\} | D \leftarrow \mathbb{D})\}$, and performing a hypothesis testing to check if the model f was sampled from one distribution or the other given a target sample (x, y). Since evaluating such probabilities is intractable, the authors propose to use the loss on (x, y) to obtain the equation

$$p(l(f(x), y) | \tilde{\mathbb{Q}}_{in/out}(x, y)) \tag{3}$$

where $\tilde{\mathbb{Q}}_{in/out}$ represents the empirical distributions approximated as Gaussian. Since the cross-entropy loss needs to be appropriately handled to transform into a Normal distributions, a logit-scaling transformation is performed to extend the range from $[0, 1]$ to $(-\infty, +\infty)$

$$\phi(f(x)_y) = \log \left(\frac{f(x)_y}{1 - f(x)_y} \right) \tag{4}$$

The authors propose two variants of their attack algorithm an online variant which trains shadow models on the target sample and an offline variant where the shadow models are trained ahead of time and not on the target samples reducing the computational costs significantly. For the offline variant the hypothesis testing is conducting as follows

$$\Lambda = 1 - \Pr[Z > \phi(f(x)_y)] \tag{5}$$

where, $Z \sim \mathcal{N}(\mu_{out}, \sigma^2_{out})$ represents the mean and variance of the model confidences when the target samples' were not included in the shadow model training. The larger the target model's confidence compared to μ_{out}, the more likely the query sample is a member.

3 RMIA

We now turn our attention to another related work by (Zarifzadeh et al., 2024) which also proposes a hypothesis test based membership inference attack. The key difference between this method and the one proposed in (Carlini et al., 2021) is the way the null hypothesis is tested. That is, in LiRA we enumerate the plausible worlds where the target sample was either included or not in the training set, but in this approach dubbed as RMIA (Robust Membership Inference Attack), the enumeration accounts for a more realistic setting where the target data point might have been replaced by a random sample from the data population.

3.1 Designing RMIA

The authors propose a novel hypothesis testing method, wherein the null hypothesis, that is, the target data point x is not a member as a composition of possibilities where the target data point was replaced by a random sample from the data population. The likelihood ratio tests (LR) is conducted for all such possibilities yielding a more robust membership inference attack. That is,

$$\mathtt{LR}_\theta(x, z) = \frac{\Pr(\theta | x)}{\Pr(\theta | z)} \tag{6}$$

This provides a more fine-grained approach to auditing the models, because the LR tests now evaluate the scenario where the probability of observing the model (θ) under the hypothesis when x is in its training set should be larger than the probability of observing the model when, x is replaced by a random sample z.

3.2 How RMIA works?

Similar to the LiRA method, the RMIA also has an online and offline variants. The offline variant which is computationally friends involves computing a $Pr_{OUT}(x)$ giving us the probability of the models where the $x \notin D$. Given $LR_\theta(x, z)$, the membership inference attack is given by

$$\texttt{Score}(x, \theta) = \texttt{Pr}_{z\sim\pi}(LR_\theta(x, z) \geq \gamma) \tag{7}$$

The RMIA proceeds in 2 steps, in the first step random samples $z \sim \pi$ are sampled from the prior distribution on the data population and compute the score given in Equation 7. The samples z are deemed passed if $LR_\theta(x, z) \geq \gamma$ and further if the computed score (Equation 7)$\geq \beta$ then its considered as a member, where $\beta \in [0, 1]$.

4 How do both of these compare against each other?

The online-LiRA mechanism can be seen as an average case of the RMIA attack, because there $z \approx \pi$. The authors show that RMIA performs well even when the number of z samples and the reference models that are trained are restricted. On the other hand, the LiRA attack needs greater number of reference models to be trained to be truly effective and even then lags when the TPR-at-FPR is measured. Both the methods agree on the fact that the MIA should also account for per-example hardness, that is, not all data points have similar effect on the model. For instance, authors in (Baldock et al., 2021) study prediction depth and sample hardness and show that usually easier examples require smaller prediction depth and are learned earlier in the layers while difficult examples need larger prediction depth and more likely memorized. This can have implications for privacy since some *outlier* examples which are computationally hard for the model might be memorized and more vulnerable to MIA. There is also an interesting line of work that talks about some amount of memorization is necessary for good generalization (Feldman & Zhang, 2020; Feldman, 2021; Brown et al., 2021). This would suggest the models might naturally inherit some adversarial brittleness and carefully designed MIAs could exploit such vulnerabilities. We can see that RMIA is a better and more robust membership inference attack because it evaluates the scenarios where the target sample was replaced by a random sample, which provides a much stronger signal to determine membership/non-membership.

References

Robert John Nicholas Baldock, Hartmut Maennel, and Behnam Neyshabur. Deep learning through the lens of example difficulty. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=WWRBHhH158K.

Gavin Brown, Mark Bun, Vitaly Feldman, Adam Smith, and Kunal Talwar. When is memorization of irrelevant training data necessary for high-accuracy learning? In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, STOC '21, pp. 123–132. ACM, June 2021. doi: 10.1145/3406325.3451131. URL http://dx.doi.org/10.1145/3406325.3451131.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. *CoRR*, abs/2112.03570, 2021. URL https://arxiv.org/abs/2112.03570.

Vitaly Feldman. Does learning require memorization? a short tale about a long tail, 2021. URL https://arxiv.org/abs/1906.05271.

Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation, 2020. URL https://arxiv.org/abs/2008.03703.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009. URL https://api.semanticscholar.org/CorpusID:18268744.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Herve Jegou. White-box vs black-box: Bayes optimal strategies for membership inference. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5558–5567. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/sablayrolles19a.html.

Reza Shokri, Marco Stronati, and Vitaly Shmatikov. Membership inference attacks against machine learning models. *CoRR*, abs/1610.05820, 2016. URL http://arxiv.org/abs/1610.05820.

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, 2018. doi: 10.1109/CSF.2018.00027.

Sajjad Zarifzadeh, Philippe Liu, and Reza Shokri. Low-cost high-power membership inference attacks, 2024. URL https://arxiv.org/abs/2312.03262.