Membership Inference Attacks against GANs by Leveraging Over-representation Regions

Hailong Hu
SnT, University of Luxembourg
Esch-sur-Alzette, Luxemburg
hailong.hu@uni.lu

Jun Pang
FSTM & SnT, University of Luxembourg
Esch-sur-Alzette, Luxemburg
jun.pang@uni.lu

ABSTRACT

Generative adversarial networks (GANs) have made unprecedented performance in image synthesis and play a key role in various downstream applications of computer vision. However, GAN models trained on sensitive data also pose a distinct threat to privacy. In this poster, we present a novel over-representation based membership inference attack. Unlike prior attacks against GANs which focus on the overall metrics, such as the attack accuracy, our attack aims to make inference from the high-precision perspective, which allows the adversary to concentrate on inferring a sample as a member confidently. Initial experimental results demonstrate that the adversary can achieve a high precision attack even if the overall attack accuracy is about 50% for a well-trained GAN model. Our work will raise awareness of the importance of precision when GAN owners evaluate the privacy risks of their models.

CCS CONCEPTS

• **Security and privacy**; • **Computing methodologies** → **Machine learning**;

KEYWORDS

Membership Inference Attacks; Generative Adversarial Networks; Human Face Generation; Over-representation

ACM Reference Format:
Hailong Hu and Jun Pang. 2021. Membership Inference Attacks against GANs by Leveraging Over-representation Regions. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS '21), November 15–19, 2021, Virtual Event, Republic of Korea.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3460120.3485338

1 INTRODUCTION

Machine learning, including discriminative models and generative models, has made tremendous progress in a wide range of application domains. In particular, generative adversarial networks (GANs) have made enormous progress in image generation since the seminal work was proposed by Goodfellow et al. [4] in 2014. Since then, GANs have achieved impressive performance in a variety of areas — image synthesis, image-to-image translation, and texture generation, etc. However, deploying these state-of-the-art techniques on applications involving sensitive personal data, such as the human face or healthcare data, has caused severe concerns about privacy [2, 14, 16]. For instance, adversaries can mount a membership inference attack against a machine learning model in order to infer whether a given sample was in the training set, which directly leads to information leakage of the training set [14].

Early studies about membership inference attacks concentrate on discriminative models [13, 14], and overfitting is considered as an important reason causing the leakage of training samples. Furthermore, Yeom et al. [16] formally illustrate the connection between overfitting and privacy risks and show that overfitting is a sufficient condition but not a necessary condition for membership inference attacks. Indeed, there have been several works advocating that for the training set of a machine learning model, there are always some training points that are more vulnerable to membership inference attacks, no matter whether the model is overfitting [1, 2, 10, 11, 15]. For example, Carlini et al. [2] reveal that certain training samples in language models which exhibit no overfitting can be extracted, such as phone numbers and email addresses from the victim model GPT-2. Long et al. [11] also show that there exist vulnerable samples in well-generalized classification models. Additionally, Leino et al. [10] further advocate that even if only one training sample is inferred as a member confidently, then it should be also considered as a privacy violation. Therefore, all these works motivate us to study membership inference attacks against GANs from the perspective of precision, i.e., whether the adversary can infer a sample as a member confidently.

In this paper, we propose a novel membership inference attack against GANs, which focuses on a high-precision inference. The precision refers to the proportion of real members among the samples that are inferred as members. Our attack methodology is based on the over-representation of GAN models: if the proportion of training samples (member samples) in some regions is significantly higher than that in other regions, then it can be abused by the adversary to mount membership inference attacks. Our preliminary results show that a high-precision attack can be achieved for a well-trained GAN.

2 METHODOLOGY

2.1 Problem Formulation

Given a target GAN model G_{target} and a target dataset X_{target}, the goal of membership inference attacks is to infer whether a sample x_t from X_{target} is used to train the target model G_{target}.

Prior works [3, 5] perform membership inference attacks against GANs by comparing how close a sample x_t from X_{target} is to the

CCS '21, November 15–19, 2021, Virtual Event, Republic of Korea

ACM ISBN 978-1-4503-8454-4/21/11.
https://doi.org/10.1145/3460120.3485338

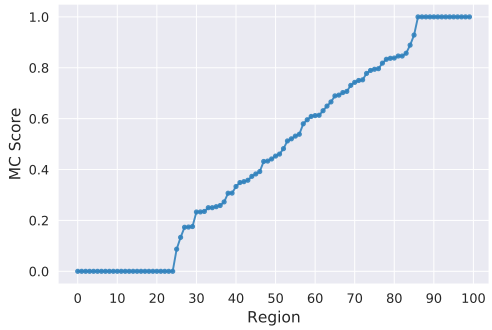

Figure 1: MC scores for target model StyleGAN.

sample generated by the target model G_{target}. Our attack methodology exploits one particular insight: when a generative model learns the distribution of a data set, there exist over-representation regions [12]. If a sample from X_{target} falls in a region where the most training samples of the target model lie, we believe this sample is more likely to be a member sample. Thus, we define a member confidence (MC) score to estimate the probability that each region contains members (see Figure 1). Finally, a sample with MC score higher than a threshold is predicted as a member. The key step of our method is to estimate a MC score which represents a region how frequent training samples occur.

Member Confidence Score. Let m, n be the numbers of member samples and nonmember samples, respectively. Here, member samples also refer to training samples of a target model. Nonmember samples are from the same distribution of training samples, but are not used for training. A region is a set of samples, which can be constructed by clustering algorithms. It is also called a cluster in our work and k is the number of regions. Member and nonmember samples are distributed among these regions, where $m = m_1 + m_2 + \ldots + m_k$ and $n = n_1 + n_2 + \ldots + n_k$. For the i^{th} region, there are m_i member samples and n_i nonmember samples. Therefore, the ratio of member samples in the i^{th} region is defined: $r_{memb}(i) = \frac{m_i}{m}$, and the ratio of nonmember samples in the i^{th} region is: $r_{nonmemb}(i) = \frac{n_i}{n}$, we define the member confidence score in the i^{th} region as:

$$MCScore(i) = \frac{r_{memb}(i)}{r_{memb}(i) + r_{nonmemb}(i)} \tag{1}$$

As an example, we show the MC scores for target model StyleGAN trained on FFHQ in Figure 1. We can observe that different regions indeed show different proportions of member samples, i.e. different MC scores.

2.2 Over-representation based Attack

In this work, we assume that an adversary can have access to the whole GAN model, including the generator and the discriminator. However, the adversary has no knowledge of the training dataset. This attack scenario usually occurs when some research institutions publish their models to the public to avoid directly sharing original data, or model providers grant their models to their customers which utilize their state-of-the-art models to develop their own applications. For attack scenarios that require much less knowledge, we leave it for future work.

Algorithm 1: Over-representation based attack

Input: target data: X_{target}; target model: G, D; the number of clusters: k

1 **def** constructMCScore(G, k):
2 Sample m samples $\tilde{X}_{memb}$ from G;
3 Sample n samples $\tilde{X}_{nonmemb}$ from G;
4 $\tilde{G}, \tilde{D} \leftarrow$ trainSubstituteModel($\tilde{X}_{memb}$);
5 $\Phi_{memb} \leftarrow$ sigmoid($\tilde{D}(\tilde{X}_{memb})$);
6 $\Phi_{nonmemb} \leftarrow$ sigmoid($\tilde{D}(\tilde{X}_{nonmemb})$);
7 *clusters* $\leftarrow$ cluster($\Phi_{memb}, \Phi_{nonmemb}, k$);
8 *MCScore* $\leftarrow$ computeMCScore($\Phi_{memb}, m, \Phi_{nonmemb}, n$,*clusters*)
 ▷ based on Eq. 1, *MCScore*: MC scores of regions ;
9 **return** *MCScore*,*clusters*
10 **def** assignMCScore(X_{target}, *D, MCScore, clusters*):
11 $X_{MCScore}$ = [] ▷ MC score of each sample from X_{target} ;
12 **forall** x_t *of* X_{target} **do**
13 $\Phi_t \leftarrow$ sigmoid($D(x_t)$);
14 $i \leftarrow$ assignCluster(Φ_t, *clusters*) ▷ i: index of cluster;
15 $X_{MCScore}$.append(*MCScore*[i])
16 **return** $X_{MCScore}$
17 **def** predictMemb($x_{MCScore}, \tau$):
18 **return** 1 if $x_{MCScore} \geq \tau$ else 0 ▷ τ: threshold

Our attack consists of three steps, which is also described in Algorithm 1. In the first step, we construct MC scores that represent the degree of over-representation in each region. In order to estimate MC scores, we first train a substitute model by querying the target model. The training process of the substitute model can be regarded as a model extraction attack against the GAN [7], which aims to duplicate the target GAN model including its functionality and implicit data distribution. In this way, data used for training the substitute model is considered as members of the model, and data sampled from the target model but not used for training the substitute model is regarded as nonmembers. Therefore, we can know the members and nonmembers for this substitute model, which is utilized to construct MC scores. Here, we leverage the discriminator's outputs instead of raw samples to split regions (Algorithm 1, line 5-7). we convert these outputs to a fixed interval through the sigmoid function because they have different ranges for different discriminators. Since the discriminator's output is a single value, we do not need to perform clustering, instead directly dividing the range of outputs into k parts, i.e. k regions. Finally, we calculate the MC score on each part through the substitute model (Algorithm 1, line 8).

In the second step, we assign the MC score to each suspect sample from the target dataset X_{target}, based on the distance between each suspect sample and each region (Algorithm 1, line 10-16). In the last step, a sample with MC score higher than a threshold is predicted as a member (Algorithm 1, line 17-18).

Note that, our method is similar to the membership inference attacks against discriminative models [15] where the top-1 confidence scores of a discriminative model are used as the features. However, our attack does not make any assumption about the training set while the attack against discriminative models needs a shadow dataset that is from the same distribution of the training set.

Table 1: Attack performance (SD: standard deviation).

Target Model	Methods	Precision (%)	Precision (%)	Recall (%)	AUCROC/ Accuracy (%)
		Mean (SD)	Maximum	Mean (SD)	Mean (SD)
StyleGAN	Ours	**96.00 (8.00)**	**100.00**	**0.03 (0.007)**	**50.02 (0.003)**
StyleGAN	LOGAN	55.65 (0.15)	55.85	55.64 (0.15)	55.65 (0.15)
PGGAN	Ours	**59.00 (2.04)**	**61.17**	**0.39 (0.20)**	**50.06 (0.04)**
PGGAN	LOGAN	51.83 (0.10)	52.01	51.83 (0.10)	51.83 (0.10)

3 PRELIMINARY RESULTS

Datasets. We perform all of some experiments with the FFHQ dataset [9], which contains 70, 000 human face images. We split the dataset into two parts: a training set for model training (60, 000 images) and a test set that is not used for training. A target set is used to evaluate the performance of membership inference attacks. It consists of the equal number of member samples (randomly selected from the training set) and nonmember samples (randomly selected from the test set). In our experiments, images are resized to 64×64 and the size of a target set is 20, 000.

Target Models. We choose PGGAN [8] and StyleGAN [9] as our target models to be attacked, considering their excellent performance and widespread adoption. In our experiments, target models with the best Fréchet Inception Distance (FID) [6] during the training progress are selected. Specifically, the FID of target model StyleGAN and PGGAN are 5.05 and 6.59, respectively.

Attack Evaluation. We use precision as a key indicator to evaluate the attack performance because it can better capture the severity of the leakage of a training set. The precision of an attack refers to the ratio of real-true member samples in all the positive inferences. We also report recall, accuracy and AUCROC. We compare our attack approach with the prior work LOGAN [5], due to the similar attack scenario. The suggested hyperparameters of LOGAN are used, and for our method, we set the threshold as 99.99th percentile of all MC scores of the target set and the number of clusters is 100. In all experiments, we repeat 5 times to evaluate attack performance.

Results. Table 1 shows a comparison of different attack methods. Overall, our method can achieve much higher mean precision than LOGAN on both target models, even if the accuracy or AUCROC is about 50%. We also report the maximum precision as a reference because it can be considered as a worst-case for target models. Our method can achieve 100% maximum precision in some cases, which indicates these samples predicted as members are all real-true samples. Our attack method achieves high precision at the expense of recall because we only consider samples with higher MC scores. It means that not all training samples can be easily inferred and there only exist some vulnerable samples in a training set. This is consistent with observations made on other machine learning models, i.e., language models or classification models [1, 2, 11].

4 CONCLUSION

In this poster, we have presented a novel membership inference attack against GANs from the perspective of precision. Our method leverages over-representation regions of a GAN model to make inferences. Initial experimental evaluations showed that our method can achieve a high-precision membership inference even though the overall attack accuracy is around 50% for a well-trained model. We hope that our study highlights the necessity that model owners should systematically evaluate the privacy risks when sharing their models, including the worst-case conditions.

As future work, we aim to relax our assumption and generalize our approach to more challenging attack scenarios. In addition, it will be interesting to design possible defense measures against our new attack. In the literature, differential privacy has been shown as a promising approach to defend against privacy attacks. However, an effective differential privacy strategy to train a GAN that produces high-quality images still needs to be developed in the future.

ACKNOWLEDGMENTS

This work is supported by the National Research Fund, Luxembourg (Grant No. 13550291).

REFERENCES

[1] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *Proceedings of USENIX Security Symposium (USENIX Security)*. USENIX Association, 267–284.

[2] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting Training Data from Large Language Models. In *Proceedings of USENIX Security Symposium (USENIX Security)*. USENIX Association, 2633–2650.

[3] Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. 2020. Gan-leaks: A taxonomy of membership inference attacks against generative models. In *Proceedings of ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 343–362.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of Annual Conference on Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2672–2680.

[5] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. 2019. LOGAN: Membership inference attacks against generative models. In *Proceedings on Privacy Enhancing Technologies*, Vol. 2019. Sciendo, 133–152.

[6] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of Annual Conference on Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 6626–6637.

[7] Hailong Hu and Jun Pang. 2021. Model Extraction and Defenses on Generative Adversarial Networks. *arXiv preprint arXiv:2101.02069* (2021).

[8] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive Growing of GANs for Quality, Stability, and Variation. In *Proceedings of International Conference on Learning Representations (ICLR)*.

[9] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 4401–4410.

[10] Klas Leino and Matt Fredrikson. 2020. Stolen memories: Leveraging model memorization for calibrated white-box membership inference. In *Proceedings of USENIX Security Symposium (USENIX Security)*. USENIX Association, 1605–1622.

[11] Yunhui Long, Lei Wang, Diyue Bu, Vincent Bindschaedler, Xiaofeng Wang, Haixu Tang, Carl A Gunter, and Kai Chen. 2020. A Pragmatic Approach to Membership Inferences on Machine Learning Models. In *Proceedings of IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 521–534.

[12] C Meehan, K Chaudhuri, and S Dasgupta. 2020. A non-parametric test to detect data-copying in generative models. In *International Conference on Artificial Intelligence and Statistics*.

[13] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. 2019. ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models. In *Proceedings of Network and Distributed Systems Security Symposium (NDSS)*. Internet Society.

[14] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *Proceedings of IEEE Symposium on Security and Privacy (S&P)*. IEEE, 3–18.

[15] Liwei Song and Prateek Mittal. 2021. Systematic evaluation of privacy risks of machine learning models. In *Proceedings of USENIX Security Symposium (USENIX Security)*. USENIX Association, 2615–2632.

[16] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy risk in machine learning: Analyzing the connection to overfitting. In *Proceedings of IEEE Computer Security Foundations Symposium (CSF)*. IEEE, 268–282.