Privacy Attacks Against Generative Models

Chenghan Song
2680951(VU) 13071084(UvA)
Vrije University Amsterdam
University of Amsterdam
c.song@student.vu.nl

Abstract

Powered by machine learning and cloud computing technologies, the risk of privacy leakage has increased and draws attention from academia and industry. However, attacks against generative models receive less attention than against discriminative models. This paper investigates the effeteness of different privacy attacks against generative models.

1 Introduction

Over the past few years, generative models have gained remarkable achievements in areas like computer vision and natural language processing and have received close attention from academia and industry. Machine learning exploits data to build and gradually raise the accuracy of models. Massive data containing private and sensitive information provides a rich source of the training dataset as machine learning becomes more and more widely applied in high-stakes applications such as healthcare and finance, making research data privacy more imperative. The diverse machine learning scenarios further increase the risk of privacy leakage: companies like Google and Amazon launch machine learning as a service(MLaaS). The cloud service trains model with user-supplied data and provide a trained model interface for users[1], during which privacy attacks might happen and lead to information leakage. Another scenario is distributed learning, where multiple parties train a model together without sharing data. Even though the data is processed locally, it will still face privacy attacks during the prediction phase or training phase[2].

This paper presents an overview and comparison of four representative privacy attacks: membership inference attack, model reconstruction attack, model inversion attack, and property inference attack. The focus of the target machine learning algorithms is generative models. The research questions of this paper are:

- What is the methodology of each attack against generative models?
- What are the differences among these attacks?
- What impact do data characteristics have on these attacks?

The paper is organized as follows: the first section introduces the background of privacy attacks against generative models. Section 2 classifies and elaborates the four privacy attacks. The final section concludes the paper and discusses the remaining challenges and open questions.

2 Privacy

2.1 Adversarial Target

The adversary's objective in privacy attacks is to compromise the confidentiality of the model, which can be broadly classified into the following groups.

- Determining if a data record is included in the training set of the target model. For instance, if the model is trained using genetic data from cancer patients, the attacker can infer the cancer status of the patients after learning that the data exists in the training set.
- Inferring the sensitive features or the full data sample. For example, if the training set contains genetic data and specific gene sequences, the privacy of the patient will be breached if the attacker possesses relevant background knowledge.
- Reconstructing one or more training samples. For instance, the attacker can reconstruct a face image using a privacy attack, thus linking the name and appearance of the individual and violating privacy.
- Extracting dataset properties that were not explicitly encoded as features or correlated to the learning task.

Machine learning is a process in which the computer attempts to mine large amounts of data for implied patterns. Several taxonomies for machine learning have been developed, one of which is that they can be categorized into generative and discriminative models from the perspective of the probability distribution. While there are a lot of studies about privacy attacks against classification models: membership inference attack[1, 3, 4, 2, 5], model reconstruction attack[6, 7, 8, 9], and property inference attack[10], the attacks against generative models have gotten less attention. There is a growing number of studies focusing on MIAs against generative models. However, little literature has been researched on the model reconstruction attacks of generative models. As to property inference attacks, there is only one very new paper on GANs[11].

Discriminative model can be used to learn the conditional probability distribution of the target Y: $P(Y|X) = X$. The goal is to find the optimal classification between different categories, reflecting the differences between heterogeneous data.

Generative model is a joint probability model $P(X, Y)$ for all variables[12]. Thus, it can generate the distribution of any variable in the model, whereas the discriminative model can only obtain a sampling of the target variable. The discriminative model does not model the distribution of the observed variables, so it is not able to express a more comprehensive relationship between the observed and target variables. Therefore, generative models are more suitable for unsupervised tasks.

With the advent of the big data era, the volume and dimensionality of data have become larger and larger. Sparse and expensive data labeling and bulky noises make deep generative models a research hotspot. There have been several different deep generative models being adopted such as generative adversarial networks(GAN)[13] as shown in figure 1, Variational Autoencoders(VAE)[14], and normalizing flows[15].

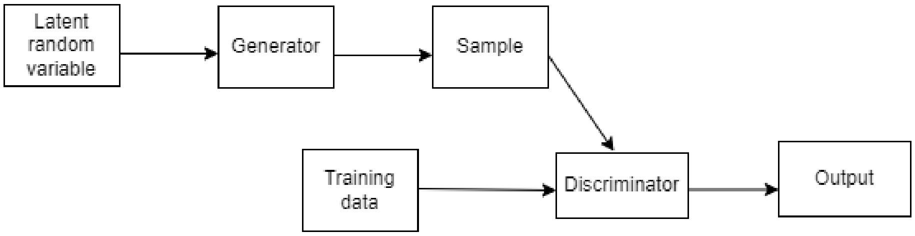

Figure 1: Architecture of Generative Adversarial Network

2.2 Adversarial Scenario

Before discussing privacy attacks against generative models, it is helpful to clarify the privacy leakage scenarios and adversary background. The attack scenario refers to the machine learning process that may cause privacy leakage. On the one hand, the training dataset may contain sensitive information; on the other hand, untrustworthy participants may present during the training or prediction phase. In the following, two main adversarial scenarios will be introduced.

Centralized learning is when conducting machine learning tasks, all the training data is stored in a central server. Model training and prediction are also performed on the central server. From users' perspective, the central server and model visitors are unreliable third parties[16]. While there

Table 1: Membership inference attacks against generative models

Ref.	Target model	Adversarial knowledge	Approach	Baseline	Metric	Dataset
[20]	GANs	White-box Black-box	Prediction confidence	Random guess	ASR	LFW DR CIFAR-10
[21]	GANs VAEs	White-box Black-box	Prediction confidence Monte carlo integration	Logan[20]	ASR, AP	MNIST Fashion-MNIST CIFAR-10
[22]	GANs VAEs	White-box	Reconstruction error	−	AUC	MNIST CelebA ChestX−ray8
[23]	GANs	White-box Black-box	Reconstruction error	Logan[20] Monte Carlo[21]	AUC	MNIST−III CelebA Instagram NewYork
[24]	GANs	Black-box	Prediction confidence	−	AUC	VGGFace2
[25]	GANs	White-box	Prediction confidence	Logan[20]	AUC	FFHQ

are rules(e.g., GDPR) governing data collection, the lack of common standards for data collection, untrustworthy data collectors may still over-collect data and sell user privacy, which is the most straightforward way to cause privacy leakage.

During the model prediction phase, privacy threats arise when untrusted third parties request access to the model. The trained model is provided by deploying it directly on the user side or via API access to the MLaaS platform. In this case, an attacker can conduct membership inference, model inversion, and property inference attacks against it.

Collaborative /federated learning means that multiple data owners jointly learn the model without uploading local data to a central server[17]. The attacker in this situation could be the central server or one of the training participants. In collaborative learning, there is no data collection step. Instead, each party retains its own data and participates independently in model training. Existing research on privacy attacks in federated learning has concentrated mainly on the model training phase.

2.3 Adversarial Knowledge

Adversarial knowledge is the attacker's prior knowledge of the target model, including training data distribution, model structure, and model parameters. Thus, based on adversarial knowledge, privacy attacks can be classified into three groups.

White-Box attacks are those in which the attacker is aware of the target model structure, model parameters, and training data[18]. Attacks in which the attacker does not have knowledge of the model structure and internal parameters are classified as **Black-Box** attacks. In some attack scenarios, the attacker may obtain partial knowledge, which can be called **Gray-Box** attacks. In addition, some specific attacks can be used to enhance the adversarial knowledge prior to initiating the privacy attack. For example, Hayes J rt al.[19] extract the model parameters first before privacy attack, so converting a black-box scenario that is difficult to attack into a white-box scenario.

3 Membership Inference Attack

Since Shokri et al.[1] first presented a membership inference attack(MIA), research mainly focuses on discriminative models. Shokri et al.[1] propose exploiting the performance discrepancy between discriminative models on the training and non-training sets for membership inference attacks under the black-box setting. They train several shadow models that mimic the target model and then infer whether the data record was a member of the training set by measuring the confidence scores of the model output. However, It is challenging to find out whether a generative model is overfitting, making it hard to infer membership against them. To dive into membership inference attacks against generative models, the existing related papers are listed in table 1. From the table 1 we can see that most current experiments are conducted on GANs and VAEs, lacking studies on attacks against other generative models such as normalizing flows.

3.1 Methodology

Membership inference attacks aim to ascertain whether a particular data record exists in the training data set[1]. When training data contains sensitive information like medical records, keeping personal data away from disclosure is of vital importance. However, membership inference attacks compromise this level of secrecy. Apart from direct invasion of privacy, membership inference attacks can also be used to regulate data misconduct and assess privacy protection[20].

3.1.1 Factors contribute to the successful MIAs

Because the majority of the existing publications analyze the membership inference attacks on the basis of practical evaluations, more formalized and strict explanations still need to be studied in the future. Nonetheless, MIA is effective for two main empirically-based reasons.

The first reason is the **overfitting** of the target model. Overfitting implies that the model performs much better on the training set than on the data drawn from the same population[26]. If the target model is closely fit towards the training set or has poor generalization ability, MIA can distinguish the training set and non-training set members.

Another reason is the **diversity of training record**. The general assumption of machine learning is that the testing and training data have a similar distribution. If the training data is not representative (i.e., the distribution of the training data is different from the testing data), the model trained on the training set can not fit the testing data well, making it possible to distinguish training and testing data.

The key factor contributing to the membership inference attacks is overfitting. Additionally, different models exhibit varying degrees of overfitting, resulting in varying vulnerability to MIAs[26]. In generative models, there are no direct confidence values on records in the same classes, leaving less information for implementing attacks[17]. Therefore, inferring data membership against generative models is more challenging than against discriminative models.

3.1.2 Approach

Hayes et al.[20] introduce the first membership inference attack against generative models using generative adversarial networks(GANs). Based on the knowledge that GANs are trained to understand statistical distinctions between the training set and generated data, the discriminator has a higher confidence value output when the target model overfits on the training set[26]. An assumption is made beforehand that the attacker is aware of the training set size *n*. In the white-box setting, the attacker has access to the discriminator D_{target}. The adversary copy D_{target} locally as D_{wb} and input dataset $X = \{x_1, ..., x_{m+n}\}$ into the D_{wb}. Then output probabilities are sorted in descending order, and the top *n* records are considered members of the training set. Even though the attack can achieve 100% accuracy in this setting, the discriminator of GANs is usually dropped and not accessible to the adversary. Thus, it is an unpractical setting. In the black-box setting, the adversary retrains a shadow model equivalent to the victim GAN through the sample data generated by the victim GAN and thus transforms the black-box attack into the white-box case.

Hilprecht et al.[21] specify membership inference attacks against both GANs and VAEs. They propose a reconstruction attack solely for VAEs in the white-box setting where the full model is accessible to the adversary. The loss function of VAEs is utilized to compute the reconstruction error, which will be later compared with a threshold. Records in the training set shall have lower reconstruction errors than non-training data. The second Monte Carlo(MC) attack is designed not only for GANs but also for every generative model from which records can be drawn. They assume that the testing set has the same amount of records as the testing set. The idea behind the MC attack is that if the model overfits, the generator G should yield records close to the training data. The adversary can employ Monte Carlo integration to approximate the probability of data is in the training set[27].

Chen et al.[23] present a generic membership inference attack framework that can be applied to many deep generative models in different settings, ranging from complete white-box to full black-box. The idea is that the generator produces more synthetic output records for training data than non-training data. They first reconstruct the closest synthetic record generated by the generator and then compute the membership probability by measuring the construction error between the reconstructed record and the testing record. Then, a reference GAN is trained with a relevant but disjoint dataset to alleviate the query dependency and calibrate the reconstruction error. If the reconstruction error is lower than the threshold, the record belongs to the training set.

Single MIA VS. Set MIA: The above attacks aim to infer a single record while attacks against a set of records are also worthy of exploration. Hilprecht et al.[21] first propose set MIA with the assumption that test and train records are the same. An attacker is presented with dataset $X_a = \{x_1, ..., x_n\}$ and dataset $X_b = \{x_{n+1}, ..., x_{2n}\}$. Then, just like single MIA, the attacker identifies the n records associated with the highest possibilities. Dataset that contains the most top n records is considered as the training data set.

Liu et al.[22] presented an attack framework that can launch set membership attacks, also known as co-membership attacks, based on a single attack. Their approach that retrains neural networks for different input records differs from the method of Hilprecht et al.[21] that only uses settled outputs of the generator of GANs. The idea behind this approach is that a record is part of the training set when the generator can produce comparable synthetic data.

3.2 Analysis

In this section, the investigation and answer to the second and third research questions are presented:

- What are the differences among these attacks?
- What impact do data characteristics have on these attacks?

3.2.1 Metric

A brief introduction of evaluation metrics used in related papers are shown as follows:

ASR stands for Attack Success Rate, meaning the proportion of successful attacks in all attacks. Hayes et al.[20] and Hilprecht et al.[21] utilize this metric to evaluate MIAs.

AP is short for Attack Precision which is the ratio of training set members correctly classified as members to records classified as members.

AUC is the acronym for Area-under-the-ROC-curve. The attack AUC is highly dependent on the probability ranking, which is more significant if training members are ranked higher than non-training record[26].

3.2.2 Dataset

From the table 1 we can see that a wide variety of datasets are used in membership inference attacks, most of which are image datasets. There are two main dimensions from which we can discuss the impact of data characteristics on membership inference attacks.

Data Size:
The size of the training set has a strong correlation with the degree of overfitting when training GANs[23]. The GAN trained with a smaller training set has a greater capacity for memorizing the training data, making it more susceptible to MIAs. Both Hilprecht et al.[21] and Chen et al.[23] show that the membership inference attacks are sufficiently effective when the size of the training set is small. The job of GANs trained on large training sets moves from memorization towards generalization, making it less venerable to MIAs.

Training Set Selection:
Some image dataset like CelebA contains identity information. Apart from selecting the training dataset randomly, it is also feasible to select individual identities for training[24]. From the experiment results of [24], and [23] we can see that all the GANs are more vulnerable when the training set is selected based on identity.

3.2.3 Comparison

White-box attacks VS. Black-box attacks:
From the literature, we can conclude that the accuracy and efficiency of white-box membership inference attacks are much higher than those of black-box attacks as well as grey-box attacks, meaning that publishing model parameters could increase privacy risk.

Single MIA VS. Set MIA:
According to Hilprecht et al.[21], a single membership inference attack has lower accuracy than a set membership inference attack. Moreover, Set MIA can even achieve pretty high accuracy when the training set of the target model is large.

GANs VS. VAEs:
It is noticeable that all the membership inference attacks are significantly more effective when applied to VAEs than GANs. The possible reason is that VAEs are more prone to overfitting than GANs.

4 Model Reconstruction/ Inversion Attack

Model reconstruction/ inversion attacks are a group of methods that aims to reconstruct part or all of the attributes of the target record based on the model output. Due to the lack of research in model reconstruction/ inversion attacks against generative models, we only present the basic methodology in this section. The intuition behind the attack is that the reconstruction is achievable by following the gradient in a trained network to adjust the weights and acquire the features for all classes in the network[28].

Fredrikson et al.[29] make use of the confidence output of the MLaaS platform as well as the ancillary information of the model in the white-box setting, getting the conclusion that there is a linear relationship between the number of requests one attribute and the possible number of the sensitive attributes of the target. The findings of their reconstruction attack on the face recognition model demonstrate that by integrating imaging technologies, the attack is capable of recovering the data matching to a label in the training set.

5 Property Inference Attack

Property inference attack is used to extract dataset properties that were not explicitly encoded as features or correlated to the learning task. The application and research on property inference attacks against generative models are quite limited. There is only one very new paper[11] studying property inference attacks against GANs.

5.1 Methodology

Recently, Zhou et al.[11] present the first property inference attack against generative models, more specifically GANs in full black-box and grey settings. The idea is that the property of the underlying training set can be reflected in the generated samples of the victim GAN. The workflow of the property inference attack is shown as follows:

- The attacker queries the generator of the target GAN to make synthetic samples.
- The attacker develops a property classifier for the purpose of classifying generated synthetic samples based on the target property.
- The underlying property of the target GAN can be inferred by inspecting the property of generated samples. Thus, the attacker can predict the underlying property of the GAN through the output of the property classifier.

5.2 Analysis

Zhou et al.[11] evaluate their attacks on four datasets: MNST, CelebA, AFAD, and US Census Income. The metric used in the evaluation is *absolute difference* between the inferred property and real property. Results show that both black-box attacks and grey-box attacks achieve good accuracy. We can also conclude that the attacks become more accurate and stable as the number of random samples increases. Additionally, just like membership inference attacks, attacks in the grey-box setting perform better than those in the black-box setting.

6 Conclusion

This paper surveys the literature on privacy attacks against generative models. Privacy of generative models is an emerging area. Among the three attacks, the Membership inference attack is the most possible one. However, during the research, we found that model inversion attacks and property inference attacks against generative models are still in their infancy and need future exploration.

There are several insights during the literature study: first, the effectiveness of privacy attacks highly relies on the adversarial knowledge of the target model. The second is that a smaller training set can result in easier privacy leakage. Finally, we discover that the property inference attack can be utilized to enhance membership inference attack against generative models, which enlightens an open question of what the relations among privacy attacks are?

References

[1] Reza Shokri et al. "Membership inference attacks against machine learning models". In: *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2017, pp. 3–18.

[2] 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)*. IEEE. 2019, pp. 739–753.

[3] Yunhui Long et al. "Understanding membership inferences on well-generalized learning models". In: *arXiv preprint arXiv:1802.04889* (2018).

[4] Ahmed Salem et al. "Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models". In: *arXiv preprint arXiv:1806.01246* (2018).

[5] Liwei Song and Prateek Mittal. "Systematic evaluation of privacy risks of machine learning models". In: *30th {USENIX} Security Symposium ({USENIX} Security 21)*. 2021.

[6] Zecheng He, Tianwei Zhang, and Ruby B Lee. "Model inversion attacks against collaborative inference". In: *Proceedings of the 35th Annual Computer Security Applications Conference*. 2019, pp. 148–162.

[7] Shagufta Mehnaz, Ninghui Li, and Elisa Bertino. "Black-box model inversion attribute inference attacks on classification models". In: *arXiv preprint arXiv:2012.03404* (2020).

[8] Xuejun Zhao et al. "Exploiting Explanations for Model Inversion Attacks". In: *arXiv preprint arXiv:2104.12669* (2021).

[9] Yuheng Zhang et al. "The secret revealer: Generative model-inversion attacks against deep neural networks". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 253–261.

[10] Karan Ganju et al. "Property inference attacks on fully connected neural networks using permutation invariant representations". In: *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*. 2018, pp. 619–633.

[11] Junhao Zhou et al. *Property Inference Attacks Against GANs*. Nov. 2021.

[12] Andrew Y Ng and Michael I Jordan. "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes". In: *Advances in neural information processing systems*. 2002, pp. 841–848.

[13] Ian Goodfellow et al. "Generative adversarial nets". In: *Advances in neural information processing systems* 27 (2014).

[14] Diederik P Kingma and Max Welling. "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114* (2013).

[15] Danilo Rezende and Shakir Mohamed. "Variational inference with normalizing flows". In: *International conference on machine learning*. PMLR. 2015, pp. 1530–1538.

[16] Marija Jegorova et al. "Survey: Leakage and Privacy at Inference Time". In: *arXiv preprint arXiv:2107.01614* (2021).

[17] Emiliano De Cristofaro. "An overview of privacy in machine learning". In: *arXiv preprint arXiv:2005.08679* (2020).

[18] Ximeng Liu et al. "Privacy and security issues in deep learning: a survey". In: *IEEE Access* 9 (2020), pp. 4566–4593.

[19] Jamie Hayes et al. "LOGAN: evaluating privacy leakage of generative models using generative adversarial networks". In: *arXiv preprint arXiv:1705.07663* (2017), pp. 506–519.

[20] Jamie Hayes et al. "Logan: Membership inference attacks against generative models". In: *Proceedings on Privacy Enhancing Technologies (PoPETs)*. Vol. 2019. 1. De Gruyter. 2019, pp. 133–152.

[21] Benjamin Hilprecht, Martin Härterich, and Daniel Bernau. "Monte Carlo and Reconstruction Membership Inference Attacks against Generative Models." In: *Proc. Priv. Enhancing Technol.* 2019.4 (2019), pp. 232–249.

[22] Kin Sum Liu et al. "Performing co-membership attacks against deep generative models". In: *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE. 2019, pp. 459–467.

[23] Dingfan Chen et al. "Gan-leaks: A taxonomy of membership inference attacks against generative models". In: *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*. 2020, pp. 343–362.

[24] Ryan Webster et al. *This Person (Probably) Exists. Identity Membership Attacks Against GAN Generated Faces*. 2021. arXiv: 2107.06018 [cs.CV].

[25] Hailong Hu and Jun Pang. "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. Virtual Event, Republic of Korea: Association for Computing Machinery, 2021, 2387–2389. ISBN: 9781450384544. DOI: 10.1145/3460120.3485338. URL: https://doi.org/10.1145/3460120.3485338.

[26] Hongsheng Hu et al. "Membership Inference Attacks on Machine Learning: A Survey". In: *arXiv preprint arXiv:2103.07853* (2021).

[27] Art B Owen. "Monte Carlo theory, methods and examples". In: (2013).

[28] Bo Liu et al. "When machine learning meets privacy: A survey and outlook". In: *ACM Computing Surveys (CSUR)* 54.2 (2021), pp. 1–36.

[29] 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*. 2015, pp. 1322–1333.