GANMIA: GAN-based Black-box Membership Inference Attack

Yang Bai*†, Degang Chen*, Ting Chen*, Mingyu Fan*
*School of Computer Science and Engineering, University of Electronic Science and Technology of China, China
†No.30 Institute of CETC, China
Email: alicepub@163.com, 1174471760@qq.com, brokendragon@uestc.edu.cn, ff98@163.com

ICC 2021 - IEEE International Conference on Communications | 978-1-7281-7122-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICC42927.2021.9500657

***Abstract*—Membership inference attacks (MIAs) against machine learning systems have drawn tremendous attention from information security researchers. By MIA, an adversary can speculate whether an individual data record is a member of the training set or not. Existing black-box MIA assumes that much information about the training data is available. Specifically, the attacker assumes that (s)he has the ability to query the target model without limitations or can access a sufficient dataset whose distribution is the same as the training data set. However, in a realistic scenario, MIAs usually come up with the limited number and the imbalanced proportion of target training datasets which cause significant challenges for MIAs. To launch an MIA in the realistic scenario, in this paper, we present a *novel* method called GANMIA, which generates synthetic data to augment the training samples of the shadow model for the black-box MIA by a Generative Adversarial Network (GAN). GANMIA firstly augments synthesized samples and then uses the generated samples to train the given shadow model to increase the training efficiency, and additionally improve the MIA's performance. The experimental results show that the accuracy of the black-box MIA increases by 23% with the help of our synthetic data.**

***Index Terms*—Membership Inference Attack (MIA); black-box attack; Generative Adversarial Networks (GANs); data augmentation**

I. Introduction

As machine learning (ML) couples with abilities in model building and Using algorithms that continuously assess and learn from data, more and more learning-based systems for a variety of applications. These training data often contain privacy-sensitive data such as health care records, financial information, individual photos, and so on. For example, Chen et al. [1] exploit ML algorithms for effective prediction of chronic disease based on big data from healthcare communities. Leo et al. [2] review the application of ML in the management of banking risks such as credit risk, market risk, and so on by a large number of financial data. Cevikalp et al. [3] use a set of individual's face images as training examples to build a face recognition system. Within these advances of ML applications, these ML systems face several security and privacy challenges, especially, the adversary interests to inferring some sensitive individual information about training data from an target ML system. These attacks can be classified into two types including the model inversion attacks and the membership inference attack (MIAs). Among the two kinds of attacks, in this paper, we focus on MIAs.

Shokri et al. [4] present the first MIA against ML systems, which tries to infer a given data record whether belongs to the victim model's training dataset by constructing an attack model. After that, many effective attack algorithms for MIAs have been proposed [5]–[8]. These methods mainly explore MIAs with the availability of sufficient prior knowledge about training dataset or unlimited querying access to the victim model. Long et al. [5] and Hayes et al. [6] assume that the adversary knows enough data samples which have the same distribution with the target training datasets. Nasr et al. [8] assume that the adversary knows a dataset that includes all training samples.

However, in a more realistic scenario, the assumptions, such as sufficient information about the training dataset, can not be satisfied. The attacker can only regard the target model as a restricted black box. However, it's hard for the adversary to launch a MIA with limited prior knowledge.

Researchers attempt to overcome this problem. Shokri et al. [4] and Rahimian et al. [9] introduce that the attacker repeatedly querying the target model to generate synthetic samples. Nonetheless, on one hand, the over-frequently accessing to an ML model can be easily discerned; on the other hand, in an ML platform that provides the querying service, i.e., Machine Learning as a Service (MLaaS), frequently querying costs a substantial fee. Thus, how to implement an effective block-box MIA with limited information about the training dataset, such as the limited number of samples which have the same distribution with victim model's training dataset, or limited querying is a significant challenge.

In this paper, we proposed Generative Adversarial Networks (GANs) based black-box MIA named GANMIA to solve this problem. As a typical black-box MIA method trains a shadow model, to transform the attack into a white-box problem [4]. Our work tries to obtain a shadow model that similar to the target model with a limited number of samples which have the same distribution or as a part belonging to the victim model's training data. We exploit Generative modeling to creating artificial instances which used as shadow model's training samples that retain similar characteristics to the original set. Unlike prior works, GANMIA neither assumes the adversary has sufficient sample knowledge, nor has the unlimited querying from the target model. The shadow model can only get a few samples at the beginning of training, and we call these samples 'original data'.Thus we use the GAN-based component to augment samples that have a similar distribution with the original data. so the shadow model will have enough

978-1-7281-7122-7/21/$31.00 ©2021 IEEE
Authorized licensed use limited to: POLO BIBLIOTECARIO DI INGEGNERIA. Downloaded on October 19,2023 at 10:05:56 UTC from IEEE Xplore. Restrictions apply.

samples for model training. The experimental results show that GANMIA can enhance the performance of MIA. The main contributions of this paper are summarized as follows:

- We are the first to explore the realistic black-box MIA scenario, which has more practical assumptions than prior works and gives a challenging statement in this scene, which is relevant for the real-world implementation of MIA.
- We propose GANMIA under the practical black-box assumptions. We introduce the GAN to address the challenges by enriching the training samples for the shadow model.
- Experimental results show that the shadow model trained by the crafted samples from GAN can achieve higher accuracy. Our approach improves the accuracy of black-box MIA under the scenario with limited information about the victim model's training dataset, with the state of the art. Also, we conduct experiments and provide comprehensive analysis for the proposed method.

The rest of the paper is organized as follows. Section II introduces some background knowledge of MIA, GAN. Section III presents the our attack method. Section IV presents the experiments and evaluations of our attack. Section V concludes the paper.

II. Background

In this section, we provide some knowledge about the membership inference attack, Generative Adversarial Networks (GANs), and data augmentation.

A. Membership Inference Attack (MIA)

Membership inference attacks aim to speculate whether an individual data record is a member of the victim ML model's training dataset. Nasr et al. [7] characterize the MIA into two categories: white-box and black-box, based on different attack observations.

In a **white-box** setting, the adversary knows much information of the attacked model including model parameters and the model structure. He et al. [10] propose a white-box attack on ML model. They assume that the attacker as one malicious participant who can use model parameters to recover the input data, without the requirements of knowing training data or query.

The a **black-box** observation means that the attacker does not have any specialized knowledge about the victim model's structure and parameters, but the adversary can querying from the victim model by API services, i.e., the Machine Learning as a Service (MLaaS) platforms. Shokri et al. [4] put forward the black-box MIA that the attacker can build shadow models, which behave similarly to the target model. And then, they use supervised training on the inputs and the corresponding outputs of shadow models to train the attack model. The attack model finally identifies whether a data record belongs to the target model's training dataset or not. In this paper, we focus on the black-box MIA setting.

Shokri et al. [4] develop three assumptions to generate the shadow model's training data, include unlimited querying from the victim model, knowing statistics information about the victim model's training data, or the adversary knows enough noisy version of the target's training dataset. Hayes et al. [6] assume a black-box MIA against the GAN model, in which the attacker knows the size of the training set, but does not know how data-points are split into training and test sets. Salem et al. [11] propose a data transferring attack for MIA. they do an empirical study with several public datasets and find that using a dataset which different from the target model's training data, can achieve effective MIA. But they have not provided the method how to recognize the effective dataset in the black-box observation. The existing black-box MIAs have the common drawback that in the realistic scenario, it's hard to achieve them.

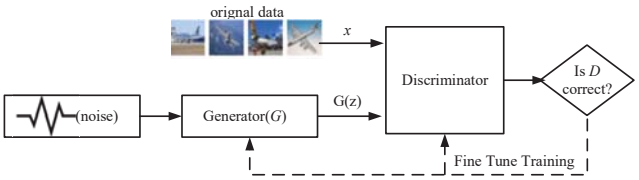

Fig. 1. GAN's architecture.

B. Generative Adversarial Networks (GAN)

GAN, a framework proposed by Goodfellow et al. [12], trains in an adversarial method to generate data mimicking some underlying distribution [6]. As Fig.1 shows, GAN has two components: the generator (G) and the discriminator (D). The generator takes noise as input and generates samples G(z) which have a similar distribution with the training data. The discriminator receives samples from both generator and the original data and then differentiates the two sources. Generator and discriminator play a competitive game where the generator learns to craft more and more realistic samples aiming at misled the discriminator, while the discriminator learns to become more and more accurate in telling apart the two sources [13]. The adversarial networks are trained by optimizing the following loss function of a minimax game:

$$\min_{G}\max_{D} E_{x\sim P_{data}} logD\left(x\right) + E_{z\sim P_z}\left[log\left(1 - D\left(G\left(z\right)\right)\right)\right]$$

G and D denote the generator and the discriminator, respectively. x is the input sample and it has the related output $D(X)$. G get $(z^{(1)}, z^{(2)}, ..., z^{(m)})$.

Antoniou et al. [14] present data augmentation GAN which takes data from a source domain and learns to take one data item and generalize it to generate other within-class data items. Because of its excellent performance, GAN has been widely used for data synthesis. Frid et al. [15] exploit GAN to generate medical images for liver lesion classification. Lim et al. [16] utilize GAN to generate data for unsupervised anomaly detection. Zhang et al. [17] apply GAN in MIA against federated learning. They assume that the adversary is a participant in federated learning who can

Authorized licensed use limited to: POLO BIBLIOTECARIO DI INGEGNERIA. Downloaded on October 19,2023 at 10:05:56 UTC from IEEE Xplore. Restrictions apply.

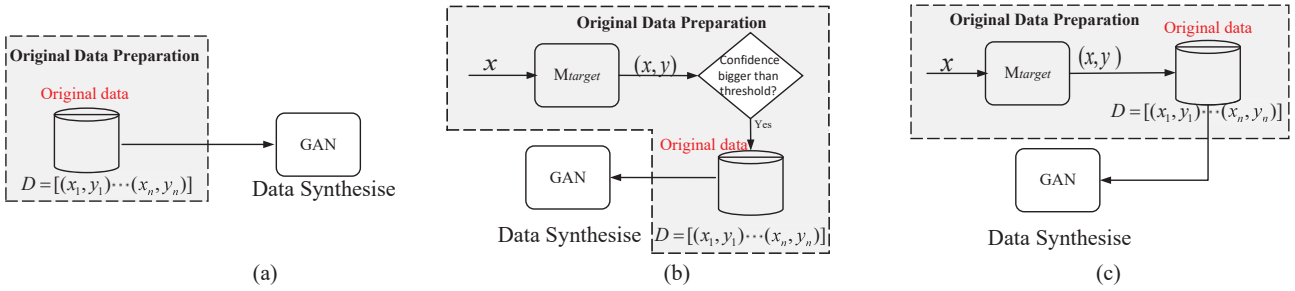

Fig. 2. Three methods of original data Preparation. (a) The adversary has limited number of original data. (b) The adversary only can query the target model. (c) The adversary has some X and can query the target model.

obtain model parameters as the attack scenario. To some extent, Zhang's work is not a black-box scenario. Instead, it has the assumption approximately nearer to white-box one.

III. GAN-based Black-box Membership Inference Attack

The main problem in black-box MIA setting described above is the lack of enough labeled training dataset for the shadow model with initial limited information about it. We introduce the GAN algorithm to enlarge the original data. We start with the GAN-based data synthesis and then describe the MIA method.

A. GAN-based Data Synthesise Method

We leverage GAN-based samples synthesize to enlarge the adversary's knowledge about the victim model's training dataset. The GAN-based samples synthesize includes three phases: assumptions, original data getting method and Data Synthesis.

Assumptions. Our method on GAN-based sample synthesis is considered with three kinds of assumptions. The adversary 1) only knows a few samples which have the statistics attribute with the target model's training dataset; 2) only can make limited querying with the victim model. 3) both know a few samples and can query the target model. In other words, we know a few original data. Compared with prior works, the original data is not sufficient; or we can query the target model infrequently, or even we do not need to query the model.

Original Data Perparation. Accordingly, there have three methods to get original data, as depicted in 2. In the first setting, the samples $X = [(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)]$ with the limited(not more than 2% of the target model's training data size) samples, who has the same underlying distribution with target training data, can directly used as original data. Under the second scenario, the adversary can feed some data x into the target model M_{target}, and judge the confidence of M_{target}'s output, if the output bigger than the threshold, the data record (x, y) can regard as the original data. This method was proposed by Shokri et al. [4]. In the third set, the attacker has a few data x, who has the same underlying distribution with target training data and can query the victim target with limited querying, the adversary can input x into M_{target} to get the original data. There are two methods to label the original data. Firstly, we can query the original records with the target model to output the classification labels. As the original data is limited, the query times should be infrequent. Secondly, if we cannot access the target model, the method proposed by Shi et al. [18] uses the Euclidean distance to measure the similarity between two images and mark the near samples as the same label.

Data Synthesis. There are many GANs models, such as BEGAN [19], VAEGAN [20], DCGANs [21], and so on. We initialize GAN with the Deep Convolutional GAN (DCGAN) according to the architecture proposed by Radford et al. [21], in which the generator and discriminator networks are both deep CNNs. This choice of models is supported by Lucic et al. [22].The generator network takes a vector of 100 random numbers as inputs and output an image of size 64×64×3. The network architecture consists of five fractionally-strided convolutions layers to up-sample the image with a 4×4 kernel size. A fractionally-strided convolutions layer expands the pixels by inserting zeros in between them to get a larger output image. Batch-normalization is applied to each layer of the network, expect for the output layer. ReLU activation functions are applied to all layers expect the output layer which uses a tanh activation functiuon. The discriminator network has a typical CNN architecture that takes the input image of size 64×64×3, and out put one decision: is this image real or fake? The network consist of five convolution layers with a kernel size of 4×4. Strided convolutions of 2×2 are applied to each convolution layer to reduce spatial dimensionality instead of using pooling layers. Batch-normalization is applied to each layer of the network, except for the input and output layers. Leaky ReLU activation functions are applied to all layers except the output layer which uses the Sigmoid function for the likelihood probability (0,1) score of the image. We train the DCGAN to synthesize CIFAR-10 for each lesion category separately as shown in Fig.3. The training process was done iteratively for the generator and the discriminator. We used batches of 128 examples for each subclass and we applied stochastic gradient descent with the Adam optimizer, parameter$\beta^1 = 0.5$. We used a learning rate of 0.0002. We obtain a limited number of the original data from original data preparation phrase and then utilize the label obtaining method

Authorized licensed use limited to: POLO BIBLIOTECARIO DI INGEGNERIA. Downloaded on October 19,2023 at 10:05:56 UTC from IEEE Xplore. Restrictions apply.

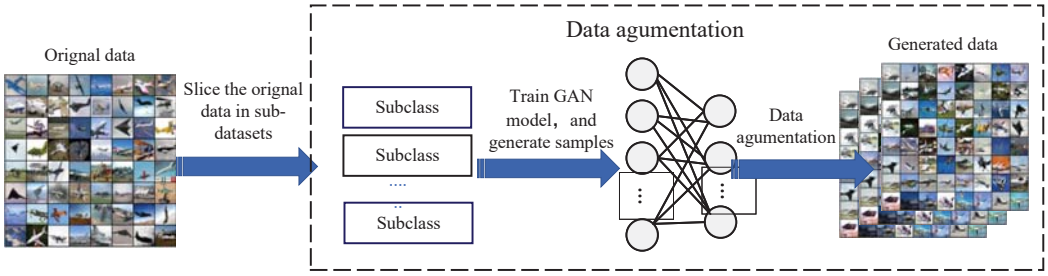

Fig. 3. The GAN-based sample synthesize workflow.

to get the label of samples. After that, we slice the original data in many sub-datasets based on the label. Finally, we use the sub-categories as the input of GAN to generate a specific class sample. These newly generated samples can be marked by the same label of the subcategory. These generated data will be used as input dataset to train the shadow model of MIA.

Algorithm 1 GAN-based Black-box MIA

Input:
Target black-box model M_T,
the initial GAN model M_{gan},
A limited number of orignal data X_o ,
required augmentation sample number N.
Candidate data $D_c=(X^c, Y^c)$.
OutPut: MIA inference result R
1:Get label Y_o for X_o;
2:$M_{gan} \leftarrow \{(X_o, Y_o), N\}$, to generate N samples $D_g=\{(X_i, Y_i)\}_N$.
3:Partion D_g into D_g^{in} and D_g^{out}
4:Train the shadow model M_s with samples D_g^{in}.
5:for i =1; i<=epoch; i++ do
6: for j=1; j<=$|D_g|$, j++ do
7 if((X_j, Y_j) belongs to D_g^{in}
8: (X_j, Y_j) marked as "member";
9: else
10: (X_j, Y_j) marked as "non-member";
11: endif
12: end for
13:end for
14:Train the attack model M_{attack} by (fature(D_g^{in}.) and the member information of shadow model.
15: For k=1; k<=$|D_c|$;k++ do
16: input (X^c_k, Y^c_k) *to* M_{attack}, and obtain the model's out put R_k
17: $R=RUR_k$
18: end for
19: Return R

B. Attack Method

The adversary approach can be divided into five stages, including (1) candidate data preparing, (2)shadow training samples preparation, (3) shadow model structure selection and training, (4) the attack model building, and (5) the membership inference. The whole processes are summarized in Algorithm 1.

Candidate data preparing, the adversary should determine target samples which supposed to be the member of target model's training dataset. We use a sample which has the high confidence value of the target model's prediction as to the candidate. the detailed method was proposed at Shorkri et al.'s work [4].

Shadow training samples preparation. In this step, GAN is exploited to augment the original data to obtain sufficient samples which will be used as shadow training samples. The detail processes introduces in §III-A.

Shadow model structure selection and training. If the adversary knows the target model, they can use a similar type of neural network as the shadow model [23]. MLleaks [11] proposed to an adversary not know the structure of the target model. The proposed a data transferring method [24] for MIA in this setting. We applied convolutional neural network(CNN) to build the shadow model. The CNN network constructed with two convolutional layers and two pooling layers with one hidden layer containing 128 units in the end. After selecting the shadow model, the attacker trains the shadow model. The crafted data will be separated into two parts. One is the training data for the shadow model, the other is used for testing the shadow model.

Attack model building. This stage uses the shadow model to produce data which is marked as members or non-members of the shadow model to build the attack. Such data reflects the behavior of the shadow model on their training and test dataset. Shokri et al. [4] propose to build the attack model as the binary classifier. Our attack model is established with a 64-unit hidden layer and a softmax output layer.

Membership inference. Finally, the attacker feeds the candidate samples into the attack model to infer whether they are the member of target model's training dataset.

IV. Experiments and Evaluation

In this section, we present our experiments and results. We test the attack efficiency of GAN-based black-box MIA and analyze the reason why the GAN-based method is effective.

TABLE I
Comparing the Attack Accuracy on Different Augmented Data Size.

	Original Data	**Our Method**
Attack Accuracy	59.1%	82.1%

Authorized licensed use limited to: POLO BIBLIOTECARIO DI INGEGNERIA. Downloaded on October 19,2023 at 10:05:56 UTC from IEEE Xplore. Restrictions apply.

A. Experiments Setup

Testbed. For the implementation of GAN-based black-box MIA, we use the Pytorch framework. All training processes are performed on an NVIDIA GeForce GTX 1060 GPU.

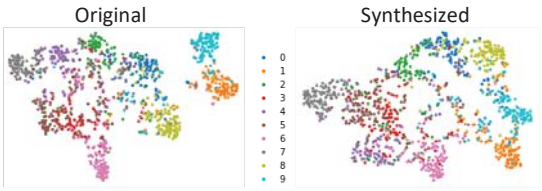

Fig. 4. t-SNE graph of CIFAR-10 original data and synthesized data. the colorful dots denote different classes of CIFAR-10.

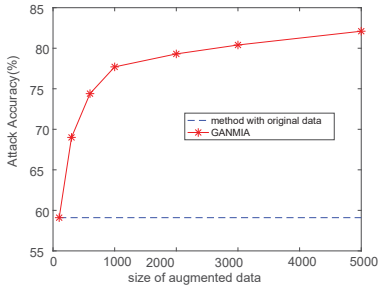

Fig. 5. The GANMIA's attack accuracy under different augmented data size compared with the original one.

Dataset and Setting. We use CIFAR-10 [25] for our experiments, which consists of 50000 color images of size 32×32 for training and 10000 color images of size 32×32 for testing. CIFAR-10 has 10 classes such as "airplane", "dogs","cats". Each round of image augmentation generates the same amount of new images from existing ones. The first batch of augmented images is obtained from the original images. We use a convolution neural network(CNN) to build the target model. The CNN is assembled with two convolution layers and two pooling layers with one hidden layer containing 128 units in the end. We use Tanh as the activation function. and we set the learning rate to 0.001, and the maximum epochs of target training and shadow model training to 50. We choose the CNN as shadow model because the CNN model with a different parameters can mimic lot of other models. The initial number of original images is 100.

Attack Accuracy. The effectiveness of our method is measured by the inference accuracy or the poison accuracy of the attack model, where the inference accuracy is the fraction of data samples in members and non-members that the attack classifier can correctly predict as member or no-member.

B. Performance of GANMIA

Accuracy of GANMIA. We evaluate the performance of our method, by computing the attack accuracy of the MIA. We set the target epoch, the attack epoch, and the shadow model's epoch as the same value 50, and we set the GAN epoch as 150. Then We launch a GANMIA to the black-box target model which trained with 10000 samples. And record the inference accuracy of GANMIA with different generated data sizes includes 100, 300, 600, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 10000. The result shows in the table I that under the restrictive black-box scenario, the attack with original data has attack accuracy approx mite to 59.1%. It's easy to understand this result because the prior knowledge of the MIA attack is insufficient, so the attack accuracy is low. Our method, by augmenting data to enlarge the knowledge number, improves the accuracy of the black MIA by 23% to reach 82.1%. Fig. 5 indicates that the accuracy increases quickly when the augmented data size increases from 0 to about 1000. After that, the increasing speed becomes steady. This result can inspire us to find the lower cost of data augmentation by GAN to get better improvement.

Results analysis. After demonstrating the performance of our GAN-based black-box MIA, we analyze why our method has high accuracy. To this end, we embed the original data and the new synthesized data into 2D Space using t-Distributed Stochastic Neighbor Embedding (t-SNE). The result, as shown in Fig. 4, demonstrates that the two datasets (the original data and the augmented data) which our method is effective, are both tightly clustered together.

C. What Influence GANMIA's Performance

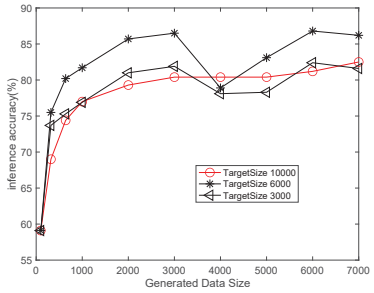

Fig. 6. The GANMIA's attack accuracy under different victim training data size.

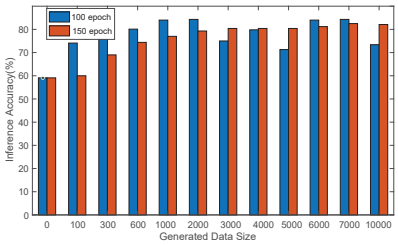

Fig. 7. The inference with different GAN's epoch.

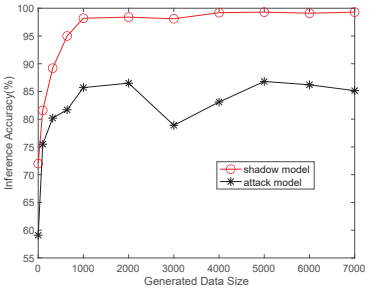

Fig. 8. The accuracy-variations trend between shadow model and attack model.

Authorized licensed use limited to: POLO BIBLIOTECARIO DI INGEGNERIA. Downloaded on October 19,2023 at 10:05:56 UTC from IEEE Xplore. Restrictions apply.

Attack with Different Target Training size. We further explore GANMIA's performance under different target training data sizes, i.e. 3000, 6000, and 10000. Fig. 6 shows the inference accuracy of GANMIA with different target training sizes. We can see clearly that among these three lines, the accuracy under 10000 target training size becomes quicker convergent than others. As Shokri et al. [4] proposed that MIAs have deep relation with the model's overfitting level, we can infer that the larger training size will make the victim model more generalization. As the result shows, the model with the 6000 training size encounters the highest inference accuracy than others.

Accuracy within Different GAN's Training Epoch. We further compare the attack of GANMIA under different GAN epochs includes 100 and 150. The results are shown in Fig. 7. It can be observed that when the generated data size smaller than 3000,the inference accuracy under 100 epoch higher than ones under 150 epoch. However, when the data augmentation scale exceeds 3000, the GANMIA's performance under 150 epoch comes out higher than the 100 epoch ones. It reveals that the relationship between the inference accuracy and the GAN's training epoch is not the simple linearity.

Relationship between Shadow Model Accuracy and Attack Model Accuracy of GANMIA. Fig. 8 shows that the inference accuracy of the GANMIA model and the classifier accuracy shadow model. We assume the experiment under the setting of 6000 target training size and record the accuracy with different generated data sizes. As we expected. the accuracy trends of the attack model gradually improved with the increase of the shadow model's accuracy. In particular, we improve the performance of the shadow model from 72% to 95% when the generation data reach 1000.

V. Conclusion and Future Work

We propose GANMIA, an attack that is effective in the restrictive black scenario with limited information about the training data and limited querying access to the target model. Our method takes advantages of GAN to generate augmented samples of the original data. Experimental results demonstrate that, compared with MIA without data augmentation, GANMIA achieves higher attack accuracy by about 23% with the target model trained on the CIFAR10 dataset. While the recent work about black-box MIA having unlimited information about the training data or can abuse querying the victim model, our method has realistic significance.

As part of our future work, we plan to apply our attack to other privacy-sensitive datasets and compare our method with other schemes for sample augmentations. Additionally, More validation work about the exist mitigation and new defensive method exploring should be expanded in our future work.

References

[1] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities," *Ieee Access*, vol. 5, pp. 8869–8879, 2017.

[2] M. Leo, S. Sharma, and K. Maddulety, "Machine learning in banking risk management: A literature review," *Risks*, vol. 7, no. 1, p. 29, 2019.

[3] H. Cevikalp and B. Triggs, "Face recognition based on image sets," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 2567–2573.

[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)*. IEEE, 2017, pp. 3–18.

[5] Y. Long, V. Bindschaedler, L. Wang, D. Bu, X. Wang, H. Tang, C. A. Gunter, and K. Chen, "Understanding membership inferences on well-generalized learning models," *arXiv preprint arXiv:1802.04889*, 2018.

[6] J. Hayes, L. Melis, G. Danezis, and E. De Cristofaro, "Logan: Evaluating information leakage of generative models using generative adversarial networks."

[7] M. Nasr, R. Shokri, and A. 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.

[8] ——, "Machine learning with membership privacy using adversarial regularization," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, 2018, pp. 634–646.

[9] S. Rahimian, T. Orekondy, and M. Fritz, "Sampling attacks: Amplification of membership inference attacks by repeated queries," *arXiv preprint arXiv:2009.00395*, 2020.

[10] Z. He, T. Zhang, and R. B. Lee, "Model inversion attacks against collaborative inference," in *Proceedings of the 35th Annual Computer Security Applications Conference*, 2019, pp. 148–162.

[11] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M. Backes, "Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models," *arXiv preprint arXiv:1806.01246*, 2018.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[13] D. Chen, N. Yu, Y. Zhang, and M. Fritz, "Gan-leaks: A taxonomy of membership inference attacks against gans," *arXiv preprint arXiv:1909.03935*, 2019.

[14] A. Antoniou, A. Storkey, and H. Edwards, "Data augmentation generative adversarial networks," *arXiv preprint arXiv:1711.04340*, 2017.

[15] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, 2018.

[16] S. K. Lim, Y. Loo, N.-T. Tran, N.-M. Cheung, G. Roig, and Y. Elovici, "Doping: Generative data augmentation for unsupervised anomaly detection with gan," in *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2018, pp. 1122–1127.

[17] J. Zhang, J. Zhang, J. Chen, and S. Yu, "Gan enhanced membership inference: A passive local attack in federated learning," in *ICC 2020-2020 IEEE International Conference on Communications (ICC)*. IEEE, 2020, pp. 1–6.

[18] Y. Shi and Y. Han, "Schmidt: Image augmentation for black-box adversarial attack," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2018, pp. 1–6.

[19] D. Berthelot, T. Schumm, and L. Metz, "Began: Boundary equilibrium generative adversarial networks," *arXiv preprint arXiv:1703.10717*, 2017.

[20] A. Bhattacharyya, B. Schiele, and M. Fritz, "Accurate and diverse sampling of sequences based on a "best of many" sample objective," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8485–8493.

[21] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[22] M. Lucic, K. Kurach, M. Michalski, S. Gelly, and O. Bousquet, "Are gans created equal? a large-scale study," in *Advances in neural information processing systems*, 2018, pp. 700–709.

[23] Z. Li and Y. Zhang, "Label-leaks: Membership inference attack with label," *arXiv preprint arXiv:2007.15528*, 2020.

[24] D. Kirkpatrick and J. Kirkpatrick, *Transferring learning to behavior: Using the four levels to improve performance*. Berrett-Koehler Publishers, 2005.

[25] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

Authorized licensed use limited to: POLO BIBLIOTECARIO DI INGEGNERIA. Downloaded on October 19,2023 at 10:05:56 UTC from IEEE Xplore. Restrictions apply.