Membership Privacy-preserving GAN

Heonseok Ha[1]
heonseok.ha@snu.ac.kr
Uiwon Hwang[1]
uiwon.hwang@snu.ac.kr
Jaehee Jang[1]
hukla@snu.ac.kr
Ho Bae[*,2]
hobae84@ewha.ac.kr
Sungroh Yoon[*,1,3]
sryoon@snu.ac.kr

[1] Department of Electrical and Computer Engineering
Seoul National University
Seoul, South Korea

[2] Department of Cyber Security
Ewha Womans University
Seoul, South Korea

[3] Interdisciplinary Program in Artificial Intelligence
Seoul National University
Seoul, South Korea

[*] Corresponding Authors

Abstract

A membership inference attack (MIA) identifies if an instance was included in the victim model's train dataset. Without an appropriate defense mechanism, MIA can result in serious privacy breaches. Although several methods have been proposed to protect membership privacy in discriminative models, research into generative adversarial networks (GANs), remains insufficient despite their vulnerability to MIAs. In this study, we propose a *membership privacy-preserving GAN* (MP-GAN), which plays an additional adversarial game for membership privacy between an auxiliary membership inference network M and a GAN. M seeks to find out whether an instance belongs to the reference or train dataset, whereas the generator and discriminator of the GAN attempt to deceive M. Our theoretical analysis results demonstrate that the MP-GAN improves membership privacy by not learning sample-specific features. We perform extensive empirical evaluations to show that the MP-GAN can successfully defend against MIAs under advantageous scenarios to the attacker (for example, white-box access to networks and small training dataset size). Furthermore, we demonstrate that the MP-GAN has several advantages over other privacy-preserving GAN training techniques.

1 Introduction

Numerous applications based on deep learning models provide high-quality services owing to recent advancements in machine-learning-as-a-service platforms. However, some studies [12, 16] have demonstrated that such models are susceptible to MIAs. A MIA is a privacy attack that reveals if a specific instance belongs to the victim model's train dataset. Overfitting is one of the causes of MIAs [15, 19] because model responses (such as confidence score and internal activation) to data points differ depending on whether the model overfits those points. Numerous studies have successfully extracted information from discriminative

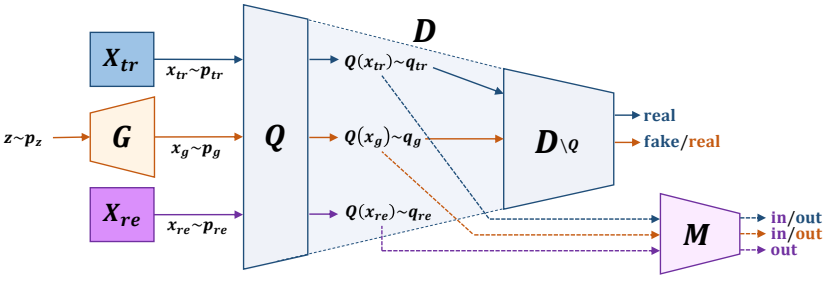


Figure 1: Overview of MP-GAN training phase. The MP-GAN aims to make training and reference data indistinguishable in the response space $\mathcal{Q}$. Here, X_{tr} denotes a training dataset (that is, member dataset), X_{re} denotes a reference dataset that plays the role of a non-member dataset in the training phase, z denotes a noise vector, G denotes a generator, and D denotes a discriminator. The internal activation of D is selected as the MP-GAN's response; that is, the response network Q is a sub-network of D. M represents a membership inference network that predicts the membership of x based on the response $Q(x)$. The parameter **in** denotes the input data, which are a member of X_{tr}, and **out** denotes the input data which are not a member of X_{tr}. The MP-GAN is trained through an adversarial game for membership privacy ((G,D) vs. M).

models in privacy-sensitive applications, such as medical record analysis [1], and federated learning [12], and contrastive learning [3].

Generative adversarial networks (GANs) [6] trained using an adversarial game (AG) between a discriminator (D) and generator (G) are also vulnerable to MIAs [2, 7]. Such attacks are particularly effective when the training dataset is small because D and G can easily overfit each data record [2]. Therefore, robust training methods for D and G against MIAs on small datasets are essential.

Models that preserve differential privacy (DP) [5] are typically robust to MIAs [14]. However, current DP-based GANs [3, 13, 20] exhibit a severe privacy-utility trade-off. This is because the DP reduces the privacy level as the number of data accesses increases. The victim model's privacy level decreases with each training iteration. Therefore, a model that preserves privacy more practically than DP-based models is required.

We propose the **membership privacy-preserving GAN (MP-GAN)**, a novel MIA defense with an AG for membership privacy that directly regularizes both G and D. The MP-GAN combines *a membership inference network* M that determines whether the training dataset includes the input sample, based on the GAN's responses. The MP-GAN is trained using a two-player AG for data generation (G vs. D) and a three-player AG for membership privacy ((G,D) vs. M).

We introduce the *reference data*, which are regarded as non-member data in training G and D, to make M learn the differences between the responses of the non-member and member data. The reference dataset is only used to update the parameters of M, whereas the parameters of M, D and G are updated using the training dataset. M attempts to find out whether the data provided are from the training or reference dataset based on the response from the GAN. D and G are trained to deceive M by providing indistinguishable responses to the provided data. We theoretically demonstrated how the AG for membership privacy improves membership privacy by preventing the MP-GAN from learning sample-specific

features.

Partition-based GANs, such as PAR-GAN [4] and privGAN [10] have been proposed to improve membership privacy by regularizing G using multiple Ds trained on partitioned training datasets. Partition-based GANs' Ds are vulnerable to MIAs because each D can access only $1/N$ of the training dataset, where $N(>1)$ denotes the number of partitions, making it easy for Ds to overfit to the training samples. Although generalized G indirectly regularizes Ds, it has only a marginal effect on membership privacy. Compared with partition-based GANs, the MP-GAN has fewer parameters to learn, and D of the MP-GAN can access more samples. Except for only a small amount (approximately 10%) of the reference data, D of our method can access most (approximately 90%) of the training samples, which allows D to reduce overfitting. Additionally, the MP-GAN directly improves D's membership privacy using the AG for membership privacy.

We empirically demonstrated that the MP-GAN successfully preserves membership privacy. We assume an evaluation scenario that is significantly beneficial to the adversary where 1) the adversary is accessible to model weights; 2) the training dataset's size is sufficiently small for the target model to overfit [2, 8], and 3) the adversary is aware of how many data points in the attack data are member or non-member. However, the MP-GAN significantly reduces the attack success rate against state-of-the-art MIAs. In addition, the MP-GAN outperforms DP-based and partition-based GANs concerning the empirical privacy-utility trade-off.

The contributions of this study and the proposed MP-GAN are as follows:

1. We suggest the MP-GAN, a novel defense framework for GAN against MIAs. The MP-GAN aims to make the responses to member and non-member data indistinguishable through a three-player adversarial game with a G, D and membership inference network M.

2. We discuss the theoretical analyses of the MP-GAN's adversarial games and how the MP-GAN improves membership privacy by not learning sample-specific features. Through experiments, we corroborate that the MP-GAN eliminates sample-specific information from the training data.

3. On benchmark datasets, we demonstrate that the MP-GAN effectively reduces the MIA's performance, even under disadvantageous conditions for the data holder: large target models trained on a small dataset and white-box access. We also compare the proposed method with DP-based and partition-based GANs and show that the MP-GAN has a better empirical privacy-utility trade-off.

2 Background

2.1 MIA against GANs

In the MIA scenario, the attackers attempt to find out whether an instance is present in the victim model's train dataset. This objective is generally achieved by observing how the target model *responds* to certain input queries. For example, the confidence score, internal activation, or gradients of the member data should be distinguishable from those of input queries never seen by the model [12].

The two networks composing GANs, D and G, are susceptible to MIAs against GANs. An attack method against D [7] uses the validity score $D(x)$ as a response. Because $D(x)$

reflects an authenticity of x obtained from the training data distribution, it predicts x as member data if $D(x)$ is above a predetermined threshold. The distance between x and the reconstruction of x (which has the smallest distance from x among the possible synthetic images generated from G) is used as a response in an attack method against G [2]. The concept is that because G reconstructs samples similar that resemble those of X_{tr}, the distance from the reconstruction is larger for the non-member data than that for member data. The method used to calculate these distances is selected based on the MIA scenario; in black-box MIA, the Monte Carlo-based method is used, whereas in white-box MIA, the optimization-based method is used.

2.2 Adversarial regularization of membership privacy

An auxiliary network for membership inference was proposed to build a membership privacy-preserving discriminative model [11]. The target and membership inference models are trained through an *AG* using the given training and reference datasets. The direct application of [11] to the GAN is a two-player *AG* for membership privacy using training and reference samples without considering synthetic data generated by GANs. The MP-GAN performs a three-player *AG* for membership privacy to address the membership of a synthetic sample (Section 3.2). This method significantly improves the synthetic samples' quality (Section 4.5).

3 MP-GAN

For convenience of expression, we consider two disjoint infinite datasets: a member dataset X_{in} and a non-member dataset X_{out}. MIAs determine whether a data point x is included in X_{in} or X_{out} using the responses from the target network. To make a GAN robust to MIA, the GAN's responses to $x_{in} \in X_{in}$ and $x_{out} \in X_{out}$ should be indistinguishable.

We divided X_{in} into two disjoint datasets: training dataset $X_{tr} \subset X_{in}$ and reference dataset $X_{re} \subset X_{in}$. Each dataset served as a member or a non-member during training. The MP-GAN aims to make the GAN's responses to $x_{tr} \in X_{tr}$ and $x_{re} \in X_{re}$ indistinguishable. Two *AG*s are used to train an MP-GAN: *an **a**dversarial **g**ame for **m**embership **p**rivacy* (AG_{mp}) and *an **a**dversarial **g**ame for **gen**eration* (AG_{gen}). AG_{mp} regularizes the GANs to avoid overfitting X_{tr} using X_{re}.

3.1 Adversarial game for generation

Let $\mathcal{X}$ and $\mathcal{Z}$ represent the data and noise spaces, respectively. During AG_{gen} between generator $G : \mathcal{Z} \to \mathcal{X}$ and discriminator $D : \mathcal{X} \to [0,1]$, the MP-GAN learns the distribution of X_{tr} and generates synthetic data x_g from the learned distribution. The objective of AG_{gen} of MP-GAN is the same as that of vanilla GAN [6], as follows:

$$\min_G \max_D U_{gen}(D,G) = \mathbb{E}_{x\sim p_{tr}}[\log D(x))] + \mathbb{E}_{x\sim p_g}[\log(1-D(x))] \tag{1}$$

where U_{gen}, p_{tr}, and p_g denote the utility function of AG_{gen}, the distribution of X_{tr} and x_g, respectively. The optimal point of AG_{gen} is $p_g(x) = p_{tr}(x)$ in $\mathcal{X}$.

3.2 Adversarial game for membership privacy

Based on the response of the GAN to x, MIAs try to determine the membership of the query sample x. Here, we define the general response of the GAN as $Q(x)$, where $Q : \mathcal{X} \to \mathcal{Q}$ denotes the response network and $\mathcal{Q}$ denotes the GAN's response space. As shown in Fig. 1, we set Q as a subnetwork of D and $\mathcal{Q}$ as the output space of Q.

To ensure the membership privacy of the GAN, we introduced an auxiliary binary classification module M (that is, membership inference network $M : \mathcal{Q} \to [0,1]$). Based on the response $Q(x)$, M predicts the membership of x. During AG_{mp}, M predicts whether $x \in X_{tr}$ (that is, $M(Q(x)) = 1$) or not (that is, $M(Q(x)) = 0$), whereas D and G attempt not to disclose the membership information of x through $Q(x)$.

To train M in a supervised learning manner, we introduce a disjoint reference dataset X_{re} (that is, $X_{tr} \cap X_{re} = \emptyset$). This dataset was inspired by existing MIA studies on different models [11, 16]. Because X_{re} is not used to update the parameters of (G,D), M predicts $x_{re} \sim p_{re}$ as a non-member of X_{tr}, where p_{re} denotes the distribution of X_{re}. Responses from X_{tr} and X_{re} are indistinguishable when using AG_{mp}, and Q is trained to improve membership privacy. Consequently, because D is a super-network of Q and $D \setminus Q$ is deterministic, all D's internal responses are therefore indistinguishable, which satisfies membership privacy.

Furthermore, we incorporate the synthetic data x_g generated by G in AG_{mp}; M predicts x_g as a member of X_{tr} (that is, $M(Q(x_g)) = 1$), whereas (D,G) deceives M (that is, $M(Q(x_g)) = 0$). When G learns the data distribution through D, membership privacy-preserving D indirectly makes G membership private without considering the membership of x_g. However, the MP-GAN updates parameters of G in both AG_{mp} and AG_{gen}, allowing for faster convergence. We discovered that x_g is of better quality when G's membership privacy is directly guaranteed by considering the membership of x_g (Section 4.5). The following is a definition of objective of AG_{mp}:

$$\begin{aligned}\min_{D,G}\max_{M} U_{mp}(M,D,G) = \; & \mathbb{E}_{x\sim p_{re}}[\log(1-M(Q(x)))] \\ & + \alpha\mathbb{E}_{x\sim p_{tr}}[\log M(Q(x))] + (1-\alpha)\mathbb{E}_{x\sim p_g}[\log M(Q(x))]\end{aligned} \tag{2}$$

where U_{mp} and $\alpha \in [0,1]$ denote the utility function of AG_{mp}, and a hyper-parameter for the weights of x_{tr} and x_g, respectively. We define q_{tr}, q_{re} and q_g as the distributions of the responses $Q(x_{tr})$, $Q(x_{re})$, and $Q(x_g)$ on $\mathcal{Q}$, respectively.

3.3 Optimization

Because the MP-GAN plays AG_{gen} and AG_{mp} simultaneously, the parameters of the MP-GAN are updated to optimize both U_{gen} and U_{mp}. We weigh U_{mp} by the hyper-parameter λ; subsequently, the parameters are updated to minimize the following losses:

$$\mathcal{L}_D = -U_{gen} + \lambda U_{mp}, \; \mathcal{L}_M = -\lambda U_{mp}, \; \mathcal{L}_G = U_{gen} + \lambda U_{mp}. \tag{3}$$

The training scheme is described in Algorithm S1 of the supplementary material. x_{re} is only used to update the parameters of M. The parameters of Q are updated with D because Q is a subnetwork of D. Updating M does not change the parameters of Q.

3.4 Theoretical analysis of AG_{mp}

The theoretical analysis of AG_{mp} is presented in Theorem 1.

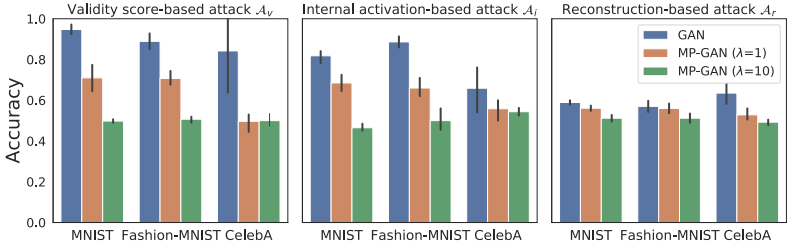


Figure 2: Attack accuracies ($\mathcal{A}_v$, $\mathcal{A}_i$ and $\mathcal{A}_r$) against the vanilla GAN and MP-GAN ($\lambda = 1, 10$). In all cases, the MP-GAN showed lower attack accuracies than vanilla GANs.

Theorem 1. *(Optimal point of AG_{mp}.) For any fixed (D,G), the optimal M to maximize V is $M^*(Q(x)) = \frac{q_\alpha(x)}{q_\alpha(x)+q_{re}(x)}$, where $q_\alpha(x) = \alpha q_{tr}(x) + (1-\alpha)q_g(x)$. The global minimum of $\max_M V(M,D,G)$ is achieved if and only if $q_{re} = q_\alpha$.*

Corollary 1.1. *When AG_{gen} and AG_{mp} attain equilibrium, the following statement holds: $q_{re} = q_{tr} = q_g$.*

Theorem 1 and Corollary 1.1's proofs are described in Section S2 of the supplementary material. Q does not overfit sample-specific features when $q_{re} = q_{tr} = q_g$. Therefore, M is unaware of the dataset from which the query sample was drawn.

Although Corollary 1.1 does not guarantee that the responses of X_{in} and X_{out} are indistinguishable because X_{out} is inaccessible during training, it provides insights into how the MP-GAN makes the responses of GANs indistinguishable between the unseen and seen data.

4 Evaluation Results

4.1 Experimental setup

We compared the MP-GAN with DP-based and partition-based GANs on three benchmark image datasets: MNIST[1], Fashion-MNIST[2], and CelebA[3]. We set the size of the member dataset X_{in} to 220, 200 for the training dataset X_{tr}, and 20 for the reference dataset X_{re}. Note that such a small dataset causes the models to easily overfit, which is the worst-case scenario for defense. For a fair comparison, we trained the other GANs, including DP-based and partition-based GANs, on X_{in} to ensure that the GANs had equal access to information during training. The details of DP-based and partition-based GANs are presented in Section S3 of the supplementary material.

We implemented the GANs with a DCGAN [13] architecture using PyTorch[4]. To train the GANs, we used the WGAN loss for MNIST/Fashion-MNIST and WGAN-GP loss for CelebA. The response network Q is a subnetwork of D, which consists of layers of D from the first to the third hidden layer; hence, the dimension of the response space $\mathcal{Q}$ is identical

[1] http://yann.lecun.com/exdb/mnist/
[2] https://github.com/zalandoresearch/fashion-mnist
[3] http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html
[4] https://pytorch.org/

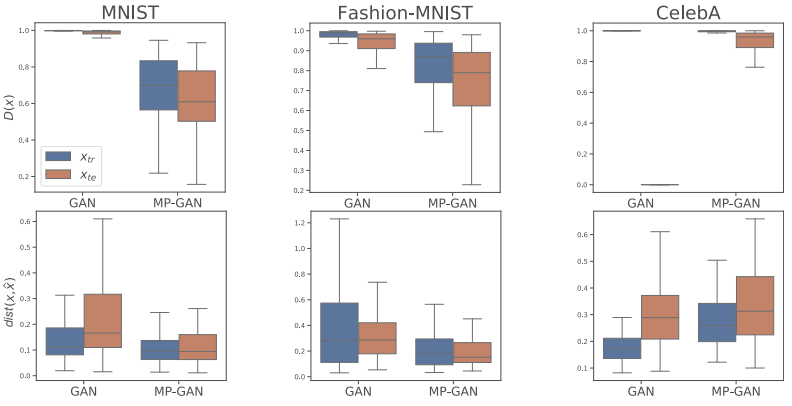


Figure 3: Comparative analysis between the vanilla GAN and MP-GAN in terms of the validity score $D(x)$ **(top)** and the difference between x and its reconstruction $\hat{x}$ **(bottom)**. The target model's robustness against the MIAs increases as the responses from members x_{tr} and non-members x_{te} become more indistinguishable.

to that of the activation of the third hidden layer. Section S4 of the supplementary material presents the details of implementation.

We evaluated the empirical privacy level of the GANs using three MIA attack methods: validity score-based attack $\mathcal{A}_v$, reconstruction-based attack $\mathcal{A}_r$ and internal response-based attack $\mathcal{A}_i$ [12]. $\mathcal{A}_v$ and $\mathcal{A}_r$ are state-of-the-art methods proposed to attack GANs (Section 2.1). $\mathcal{A}_i$ is an attack method proposed against discriminative models, which we applied to GAN for the first time. More details on the attacks are described in Section S5 of the supplementary material. We evaluated the utility of GANs using a downstream task and trained the **downstream classifier** C_d on synthetic samples generated by G and tested them on real samples. Specifically, we trained a 10-class classifier (ResNet18) that predicted the labels of MNIST and Fashion-MNIST for the downstream classification task. Such downstream classifier accuracies can measure the similarities between synthetic samples and real data distributions.

The GANs were trained for 200 epochs on Fashion-MNIST and MNIST, and for 2,000 epochs on CelebA. All experiments were repeated five times.

4.2 Robustness to MIAs

Fig. 2 shows the MIA accuracies of the validity score-based attack $\mathcal{A}_v$, internal response-based attack $\mathcal{A}_i$, and reconstruction-based attack $\mathcal{A}_r$. Every attack method yields an attack accuracy of more than 0.5 against the vanilla GAN, which implies that the MIAs against the vanilla GAN were successful. However, the MIA accuracies for the MP-GAN were lower than those for the vanilla GAN. Furthermore, a random guess accuracy was observed when $\lambda = 10$, where λ controls the weight of AG_{mp} (Section 3.2). These results demonstrate that the MP-GAN is robust against MIAs for the appropriate values of λ. We set $\lambda = 10$ as the default value for subsequent experimental results.

$\mathcal{A}_v$ showed the highest attack accuracy among the three MIAs because D directly accessed the entire training dataset during GAN training and acquired knowledge based on the training dataset. $\mathcal{A}_r$ showed the lowest attack accuracy because G indirectly accessed the

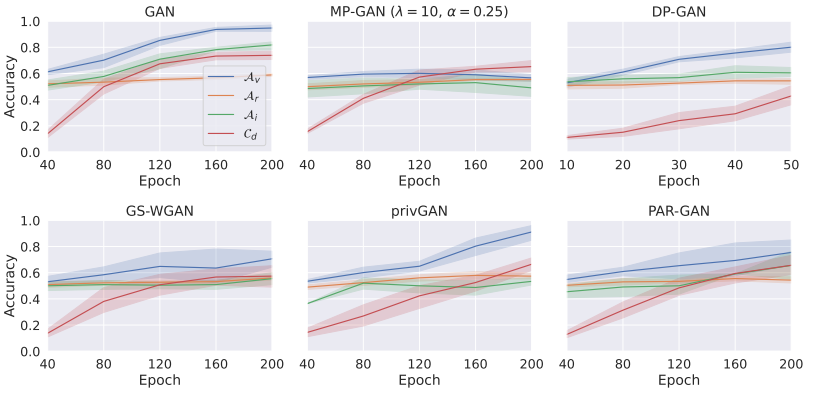


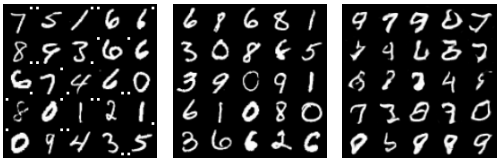
Figure 4: Attack accuracies ($\mathcal{A}_v$, $\mathcal{A}_i$, and $\mathcal{A}_r$) and downstream classification accuracy ($\mathcal{C}_d$) of the vanilla GAN, MP-GAN, DP-GAN [17], GS-WGAN [3], privGAN [10], and PAR-GAN [4] with respect to the training epochs on the MNIST dataset. As the training progresses, the DP-based and partition-based GANs become more vulnerable to $\mathcal{A}_v$, whereas the MP-GAN remains robust against all attacks.

training data through D, making it difficult for G to memorize the data. $\mathcal{A}_i$ and $\mathcal{A}_v$ are comparable because both attacks use D. However, $\mathcal{A}_i$ is partially accessible to the knowledge of D because it extracts this knowledge from the responses to query samples. For $\mathcal{A}_i$ to obtain full knowledge of D, all training samples must be queried, but this only holds true when all training samples are known to $\mathcal{A}_i$.

Fig. 3 presents the box-plot results of the validity scores $D(x)$ (top) and the distance between x and its reconstruction $\hat{x}$ from G (bottom). MIAs manipulate the difference between the non-member and member data to predict the membership of x. Therefore, the model becomes robust to MIAs if trained such that the responses are similar regardless of X_{tr} and X_{te}. According to Fig. 3, compared with the vanilla GAN, the MP-GAN successfully reduces the gaps between $D(x_{tr})$ and $D(x_{te})$ (top), and between $dist(x_{tr},\hat{x}_{tr})$ and $dist(x_{te},\hat{x}_{te})$ (bottom), where x_{te} is an unseen sample of MP-GAN during training.

4.3 Utility-privacy trade-off

We compare the empirical privacy level and utility of the vanilla GAN, MP-GAN, DP-based GANs (DP-GAN [18], GS-WGAN [3]), and partition-based GANs (privGAN [10], PAR-GAN [4]), which varied as the training progressed on MNIST (Fig. 4). As training progresses, the utility of GANs for vanilla, DP-based, and partition-based GANs increases; however, data accumulation in the network makes it vulnerable to attacks. The vulnerability level in the MP-GAN is bounded even though utility increases as learning progresses. When the accuracy of A_v is near 0.6 (at the 40-80th epoch), the classification accuracy is less than 0.4 for all models, except for MP-GAN. When each model reaches the highest classification accuracy (at the 200th epoch), the accuracy of A_v is more than 0.7 for all models, except for MP-GAN. These results suggest that the MP-GAN outperforms DP-based and partition-based GANs in terms of empirical privacy-utility trade-off.



(a) x_{tr} (b) x_{re} (c) $G(z)$

Figure 5: **(a)** x_{tr} (raw sample with 3×3-pixels white square), **(b)** x_{re} (raw sample), and **(c)** $G(z)$ generated by the MP-GAN.

(a) GAN (b) MP-GAN

Figure 6: Sampled reconstructions $\hat{x}_{tr}$ from **(a)** the vanilla GAN and **(b)** the MP-GAN on the CelebA dataset.

4.4 The sample-specific features

We demonstrate that the MP-GAN does not learn sample-specific features to improve membership privacy through experiments on synthetic and authentic face images.

Synthetic image: 3×3-white pixels (that is, sample-specific features) were added to the MNIST training images (Fig. 5(a)). Meanwhile, for X_{re}, the raw images were used (Fig. 5(b)). As a result of training the MP-GAN using the above data settings, a white square hardly appeared in $G(z)$ (Fig. 5(c)). In other words, AG_{mp} regularized G to avoid learning the sample-specific features of X_{tr}. Note that we added white pixels in this experiment only.

Face image: Fig. 6 depicts $\hat{x}_{tr}$ from the vanilla GAN and MP-GAN on the CelebA dataset. Compared with the vanilla GAN, the MP-GAN does not learn the sample-specific features of x_{tr}, such as glasses, wrinkles, and bandanas. Meanwhile, the MP-GAN learns the overall face shape and gender, which are the features that X_{tr} and X_{re} have in common.

4.5 Three-player AG_{mp} vs. two-player AG_{mp}

As described in Section 3.2, α controls how much x_g is involved in AG_{mp}. When $\alpha = 1.0$, AG_{mp} becomes a two-player game between M and D because M does not receive a response from x_g. As shown in Fig. 7, the accuracy of $\mathcal{C}_d$ is the smallest when $\alpha = 1.0$, indicating that G does not sufficiently learn the distribution of X_{tr}. However, the accuracy of $\mathcal{C}_d$ is at its highest when $\alpha = 0.25$ (three-player game). In the two-player AG_{mp}, the parameters of G were updated only in AG_{gen}, but in the three-player AG_{mp}, the parameters of G were updated in both AG_{mp} and AG_{gen}, allowing fast convergence. In the case of $\alpha = 0$, M only considers x_g without x_{tr} in AG_{mp}; thus, it is more vulnerable to MIA than other values of α.

4.6 Dimension of the response space $\mathcal{Q}$

As shown in Fig. 1, the response space $\mathcal{Q}$ is determined as one of the activation spaces of the hidden layers in D. The dimension of $\mathcal{Q}$, determined by that of the selected hidden layer, significantly affects AG_{gen} and AG_{mp}, resulting in different model utility and membership privacy. When $\mathcal{Q}$ is located in the front layer (larger dimension), AG_{gen} and AG_{mp} become challenging to reach equilibrium, and λ determines which AG is affected. For example, when $\lambda = 1$, AG_{mp} was disturbed, resulting in a higher attack accuracies (Fig. 8, left). When $\lambda = 10$, AG_{gen} was affected, resulting in a lower accuracy for $\mathcal{C}_d$ (Fig. 8, right). Conversely, when

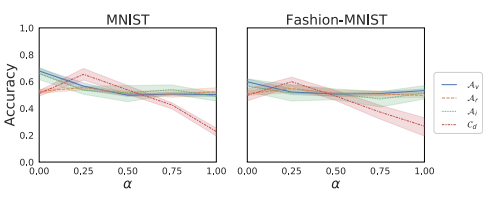


Figure 7: Attack accuracies ($\mathcal{A}_v$, $\mathcal{A}_i$, $\mathcal{A}_r$) and downstream classification accuracy ($\mathcal{C}_d$) of the MP-GAN ($\lambda = 10$) with respect to α.

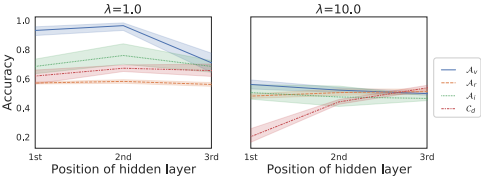


Figure 8: Attack accuracies ($\mathcal{A}_v$, $\mathcal{A}_i$, $\mathcal{A}_r$) and downstream classification accuracy ($\mathcal{C}_d$) of the MP-GAN ($\lambda = 1,\ 10$) with respect to the position of the hidden layer of D.

$\mathcal{Q}$ is located in the later hidden layer (smaller dimension), the accuracy of $\mathcal{C}_d$ increases, and the attack accuracies decrease.

4.7 Additional experimental results

In the supplementary material, we further analyzed the MP-GAN from the following perspectives: the impact of X_{re} (Section S6.1), the impact of $|X_{re}|/|X_{tr}|$ (Section S6.2), the practical dataset size (Section S6.3), the scenario when an attacker can access the layers before Q (Section S6.4), and computational cost (Section S6.5).

5 Conclusion

We have proposed the MP-GAN, a novel defense against membership inference attacks. The MP-GAN performs an additional min-max optimization (that is, an *AG* for membership privacy) during the training phase to prevent it from overfitting the training samples. During the *AG* for membership privacy, G and D are trained to deceive the membership inference network, whose goal is to determine where the input data are sampled from. We have demonstrated that MP-GAN offers a better utility-privacy trade-off than existing membership privacy defense techniques for GANs, even under advantageous scenarios for attackers.

Acknowledgements

This work was supported in part by LG Innotek and in part by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [No.2022-0-00516, No.RS-2022-00150000, NO.2021-0-02068 Derivation of a Differential Privacy Concept Applicable to National Statistics Data While Guaranteeing the Utility of Statistical Analysis, Artificial Intelligence Convergence Innovation Human Resources Development and Artificial Intelligence Innovation Hub (Ewha Womans University)].

References

[1] Michael Backes, Pascal Berrang, Mathias Humbert, and Praveen Manoharan. Membership privacy in microrna-based studies. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 319–330, 2016.

[2] Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. Gan-leaks: A taxonomy of membership inference attacks against gans. *arXiv preprint arXiv:1909.03935*, 2019.

[3] Dingfan Chen, Tribhuvanesh Orekondy, and Mario Fritz. Gs-wgan: A gradient-sanitized approach for learning differentially private generators. *Advances in Neural Information Processing Systems (NeurIPS 2020)*, 2020.

[4] Junjie Chen, Wendy Hui Wang, Hongchang Gao, and Xinghua Shi. Par-gan: Improving the generalization of generative adversarial networks against membership inference attacks. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 127–137, 2021.

[5] Cynthia Dwork. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer, 2008.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[7] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. Logan: Membership inference attacks against generative models. *Proceedings on Privacy Enhancing Technologies*, 2019(1):133–152, 2019.

[8] Klas Leino and Matt Fredrikson. Stolen memories: Leveraging model memorization for calibrated white-box membership inference. *arXiv preprint arXiv:1906.11798*, 2019.

[9] Hongbin Liu, Jinyuan Jia, Wenjie Qu, and Neil Zhenqiang Gong. Encodermi: Membership inference against pre-trained encoders in contrastive learning. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 2081–2095, 2021.

[10] Sumit Mukherjee, Yixi Xu, Anusua Trivedi, Nabajyoti Patowary, and Juan L Ferres. privgan: Protecting gans from membership inference attacks at low cost to utility. *Proceedings on Privacy Enhancing Technologies*, 2021(3):142–163, 2021.

[11] Milad Nasr, Reza Shokri, and Amir Houmansadr. Machine learning with membership privacy using adversarial regularization. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pages 634–646, 2018.

[12] Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 739–753. IEEE, 2019.

[13] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[14] Alexandre Sablayrolles, Matthijs Douze, Yann Ollivier, Cordelia Schmid, and Hervé Jégou. White-box vs black-box: Bayes optimal strategies for membership inference. *arXiv preprint arXiv:1908.11229*, 2019.

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

[16] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2017.

[17] Reihaneh Torkzadehmahani, Peter Kairouz, and Benedict Paten. Dp-cgan: Differentially private synthetic data and label generation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

[18] Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739*, 2018.

[19] 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)*, pages 268–282. IEEE, 2018.

[20] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=S1zk9iRqF7.