Membership Privacy-preserving GAN

Heonseok Ha[1], Uiwon Hwang[1], Jaehee Jang[1], Ho Bae[2,*], Sungroh Yoon[1,3,*]

[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

BMVC 2022

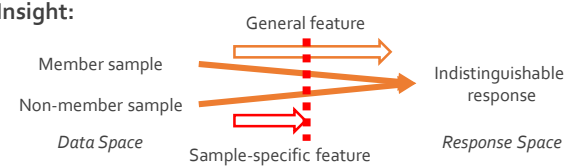
Introduction

Membership inference attack: determines whether a certain sample point was included in the victim model's train dataset.

Problem: Generative adversarial networks (GANs) are vulnerable to MIAs [1].

Goal: training GAN which is robust to MIA

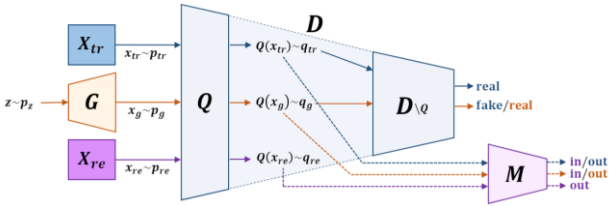
Insight:



Proposed Method: MP-GAN

Propose: membership **p**rivacy-preserving **GAN** (MP-GAN)

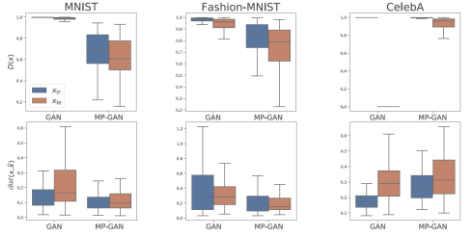
- Three-player game for membership privacy : (G, D) vs. M
- M: membership inference network → member / non-member
- $Q(x)$: response from internal activation of D
- X_{re}: reference dataset (pseudo non-member dataset)



Results: Robustness to MIA

Response → member / non-member

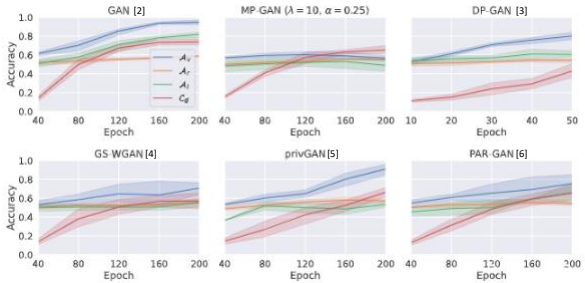
- **Top:** $D(x)$, **Bottom:** dist$(x, \hat{x})$
- **GAN:** distinguishable responses (MIA accuracy ↑)
- **MP-GAN**: indistinguishable responses (MIA accuracy ↓)



Results: Privacy-utility trade-off

Metric: performance of MIAs (attack), and downstream task (utility)

Trade-off: MP-GAN ↓, others ↑



Result: Sample-specific feature

MP-GAN does not learn sample-specific features of the training dataset, such as glasses and wrinkles.

(a) GAN (b) MP-GAN

Conclusion

We have proposed the MP-GAN, a novel defense against MIAs. We have demonstrated that MP-GAN offers a better utility-privacy trade-off than existing membership privacy defense techniques.

References

[1] Chen, Dingfan, et al. "GAN-Leaks: A taxonomy of membership inference attacks against generative models." *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*. 2020.

[2] Goodfellow, Ian, et al. "Generative adversarial networks." *Communications of the ACM* 63.11 (2020): 139-144.

[3] Torkzadehmahani, Reihaneh, Peter Kairouz, and Benedict Paten. "DP-CGAN: Differentially private synthetic data and label generation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2019.

[4] Chen, Dingfan, Tribhuvanesh Orekondy, and Mario Fritz. "GS-WGAN: A gradient-sanitized approach for learning differentially private generators." *Advances in Neural Information Processing Systems* 33 (2020): 12673-12684.

[5] Mukherjee, Sumit, et al. "privGAN: Protecting GANs from membership inference attacks at low cost to utility." *Proc. Priv. Enhancing Technol.* 2021.3 (2021): 142-163.

[6] Chen, Junjie, et al. "PAR-GAN: Improving the generalization of generative adversarial networks against membership inference attacks." *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2021.