GAN-Leaks: A Taxonomy of Membership Inference Attacks against Generative Models

Dingfan Chen1, Ning Yu2,3, Yang Zhang1, Mario Fritz1

1CISPA Helmholtz Center for Information Security, Germany
2Max Planck Institute for Informatics, Germany
3University of Maryland, College Park

Motivation

• Generative adversarial Networks (GANs) have been largely used on privacy sensitive datasets, e.g., face images and medical records
• However, existing works mainly focus on attacks against discriminative models and the privacy risk of generative models have not yet been investigated systematically
• Our work: Membership Inference Attack against GANs (whether a query sample has been used to train a GAN model?)
• Crucial to understand and control privacy leakage; provides insights for privacy-preserving data sharing

Contributions

• Taxonomy
  • Categorize attack scenarios against generative models
  • Benchmark future research
• Novel attack models
  • Generic; easy-to-implement; effective; theoretically grounded
• Extensive evaluation
  • 3 datasets with diverse data modalities, 5 victim models, 4 attack scenarios …

Taxonomy

• What information does the attacker know?
  • White-box black-box
  • Which GAN components are accessible?
    - (2) latent code; (Gen: Generator; Dis: Discriminator)
  • Partial black-box generator
  • White-box generator
• Different types of access:
  - (1) Full black-box generator
  - (2) Partial black-box generator
  - (3) White-box generator
  - (4) Accessible discriminator (full model)

Generic Attack Model

Attacker finds the best reconstruction of a query sample given different types of access to the victim generator.

• Insight: Smaller reconstruction error for training data.

• Generic Model:
  Optimization problem
  \[ \mathcal{R}(x|G_i) = G_i(z^*) \]
  \[ z^* = \arg\min_z L(x, G_i(z)) \]

• Objective:
  \[ L(x, G_i(z)) = \lambda_1 L_2(x, G_i(z)) + \lambda_2 L_{rimp}(x, G_i(z)) + \lambda_3 L_{reg}(z) \]
  \[ L_{reg}(z) = \|z - G_i(z)\|_2^2 \]
  \[ L_{rimp}(x, G_i(z)) = \|G_i(z) - x\|_2^2 - \|z\|_2^2 \]

• Different types of access:
  1. Full black-box generator
  2. Partial black-box generator
  3. White-box generator
  4. Accessible discriminator (full model)

Attack Calibration

• Problem: the reconstruction error is query-dependent (‘hard’ samples, underrepresented samples)

• Solution: Attack Calibration
  \[ L_{cal}(x, R_i(x_0)) = L(x, R_i(x_0)) - L(x, R_i(x)) \]

• • Train a reference model with:
  1. relevant but disjoint dataset
  2. irrelevant network architecture to victim model
• Theory: near-optimal under a Bayesian perspective

Experiment Results

• 3 Datasets:
  - CelebA (face), MIMIC III (medical), Instagram (location)
• 5 GAN Models:
  - PGGAN, WGAN-GP, DCGAN, VAEGAN, MedGAN
• 2 Baselines:
  - LOGAN1, MC2

• Results:
  - Attack (1) CelebA
    - DP-SGD

Summary

• A simple learning-free attack model works sufficiently well
• Attack performance highly depends on:
  1. The size of the dataset
  2. Model structure
  3. Amount of knowledge about the victim model
• Differential privacy defense is effective against real-world MI attack but compromises utility and efficiency
• Code and models are available on Github: https://github.com/DingfanChen/GAN-Leaks