Beijing Forest Studio
北京理工大学信息系统及安全对抗实验中心

面向生成模型的模型窃取方法

硕士研究生 丁杨

2022年07月17日

内容提要

- 背景简介
- 基本概念
- 算法原理
- 优劣分析
- 应用总结
- 参考文献

背景简介

- 预期收获
  - 1. 了解模型窃取方法的发展历史
  - 2. 理解面向生成模型的模型窃取方法的技术原理
  - 3. 了解生成模型窃取效果的评价指标
  - 4. 了解模型窃取在网络安全领域中的应用

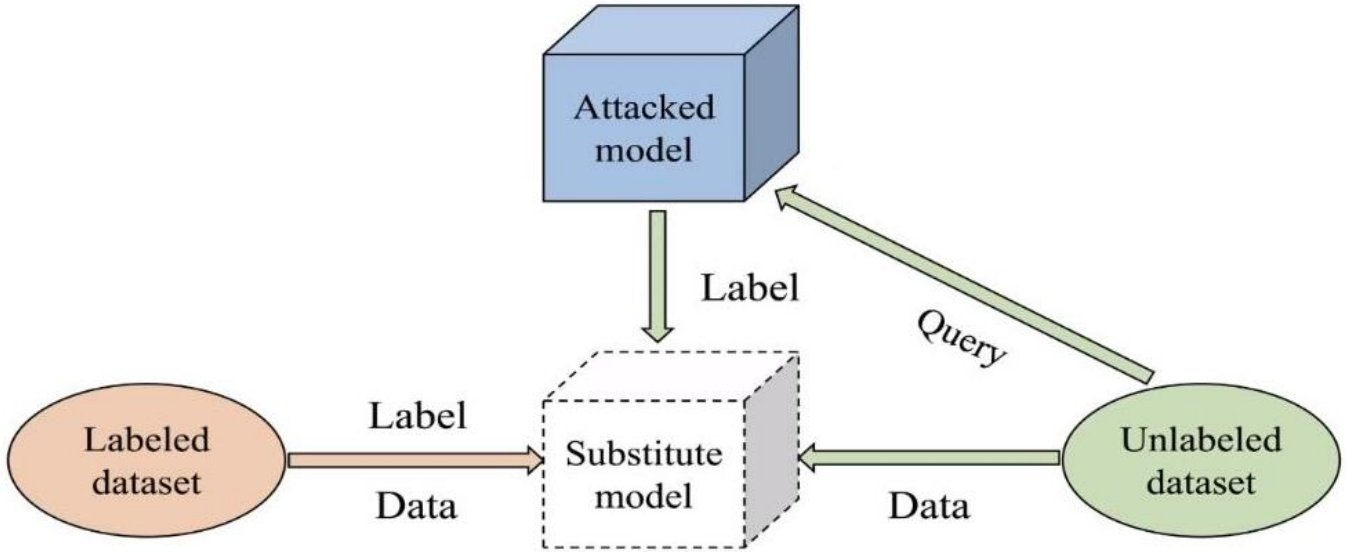
背景简介

- 模型窃取的发展历史
  - 2016年，Tramer 等人通过求解从机器学习模型结构导出的方程来提取模型的参数，但只限于SVM、随机森林等简单的模型
  - 2017年，Papernot 以牺牲替代模型的准确性为代价，通过近似目标模型的决策边界，近似窃取了DNN模型
  - 2018-2019年，Orekondy 等一些科学家先后在已知训练数据、已知训练数据种子样本等条件下，实现了对DNN的窃取
  - 之前的研究都着眼于判别模型，最新的研究表明，生成模型也面临模型窃取风险

基本概念

基本概念

- 模型窃取
  - 概念：一类隐私数据窃取攻击，攻击者通过向目标模型进行查询获取相应结果，窃取目标模型的参数或者对应功能
  - 目的：
    - 免费使用模型功能
    - 进行白盒对抗攻击
  - 方法
    - 构建方程求解参数
    - 基于元模型
    - 构建替代模型



基本概念

- 构建方程求解参数
  - 获取模型的算法、结构等相关信息，然后构建公式方程来根据查询返回结果求解模型参数
  - 考虑逻辑回归模型

$$f(\boldsymbol{x}) = \sigma(\boldsymbol{w} \cdot \boldsymbol{x} + \beta) \qquad \boldsymbol{w} \in R^d,\ \beta \in R$$

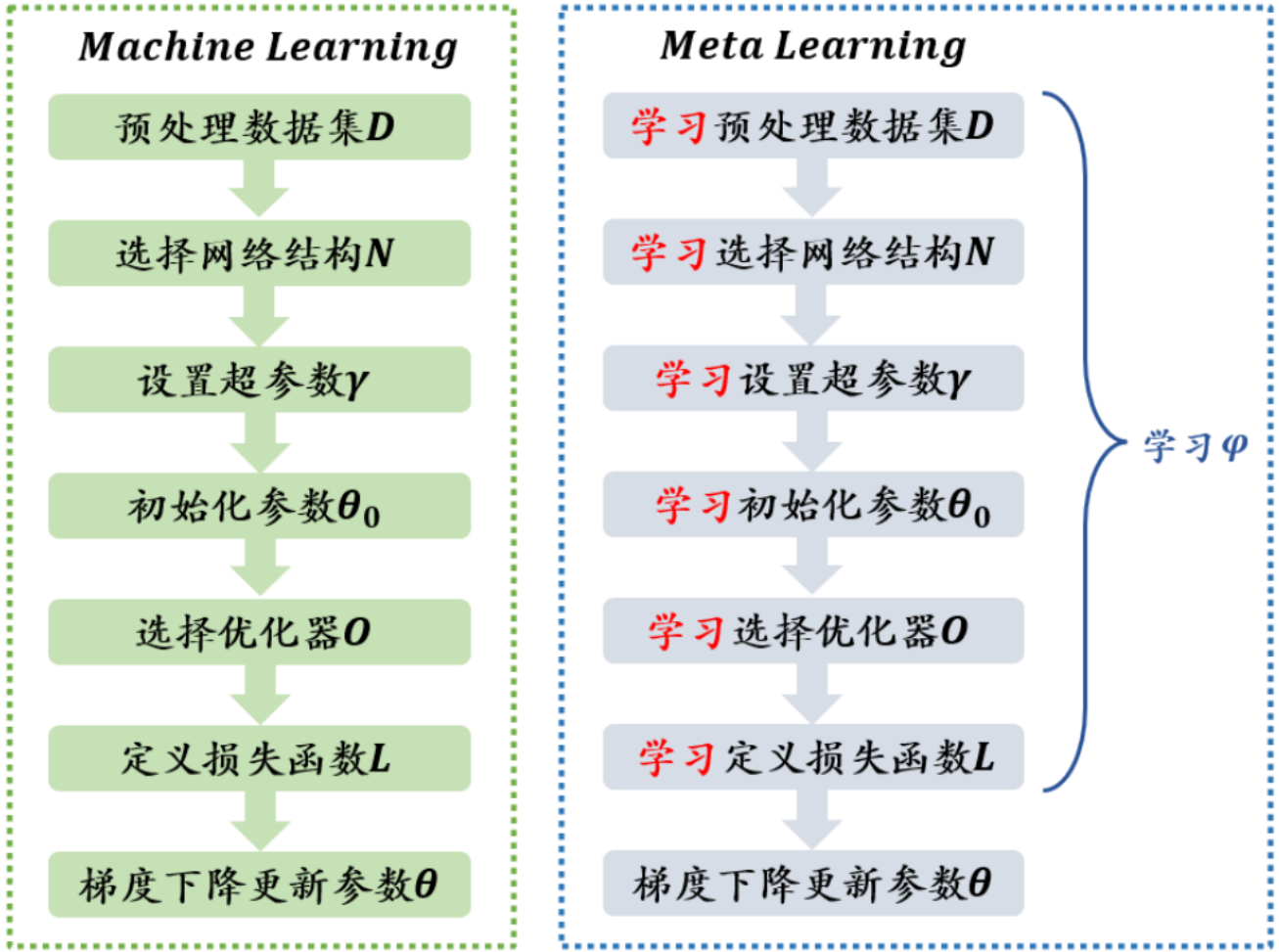
  - 通过$d+1$个线性无关的输入及结果构造公式求解$\boldsymbol{w}$和β

$$\boldsymbol{w} \cdot \boldsymbol{x} + \beta = \sigma^{-1}\big(f(\boldsymbol{x})\big)$$

$$\begin{cases} w_1 \cdot x_1 + \beta = \sigma^{-1}\big(f(x_1)\big) \\ w_2 \cdot x_2 + \beta = \sigma^{-1}\big(f(x_2)\big) \\ \cdots\cdots \\ w_d \cdot x_d + \beta = \sigma^{-1}\big(f(x_d)\big) \end{cases} \Rightarrow \begin{cases} w_1 \\ w_2 \\ \cdots\cdots \\ w_d \\ \beta \end{cases}$$

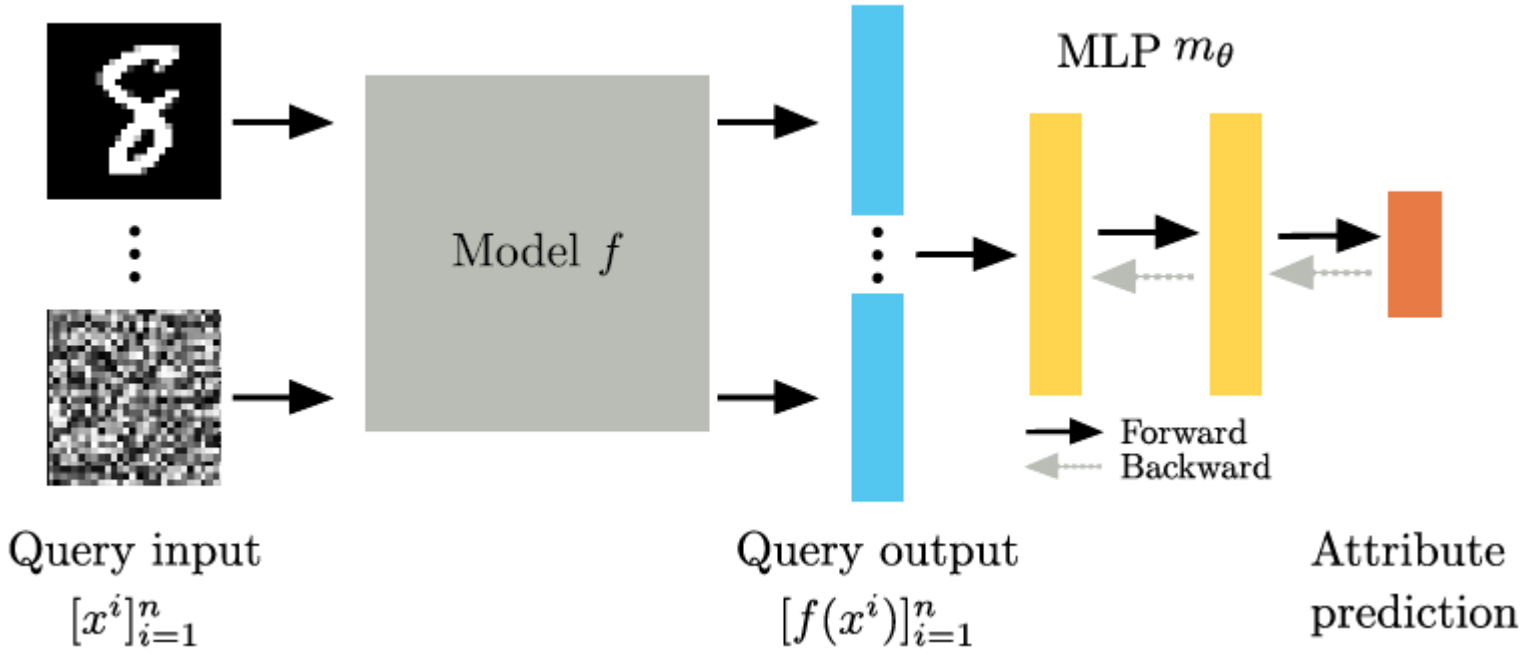
基本概念

- 元学习
  - Meta-Learning，利用以往的知识经验来指导新任务的学习，使网络具备学会学习的能力
  - 应用：优化超参数和神经网络、探索好的网络结构、小样本图像识别



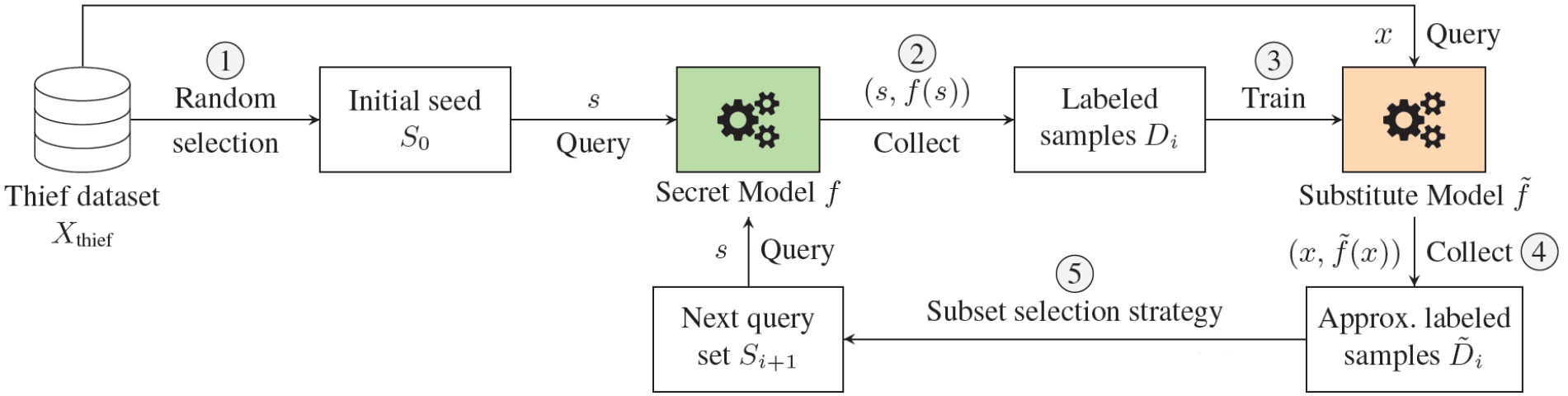
基本概念

- 基于元学习的模型窃取
  - 通过训练一个额外的元模型 $\Phi(\cdot)$ 来预测目标模型的指定属性信息，输入样本是所预测模型在任务数据 x 上的输出结果 $f(x)$，输出的内容 $\Phi(f(x))$ 则是预测目标模型的相关属性



基本概念

- 基于替代模型的模型窃取
  - 在未知目标模型结构的情况下，向目标模型查询数据得到预测结果，并以这些预测结果对查询数据进行标注构建训练数据集，训练一个与目标模型功能相似的替代模型

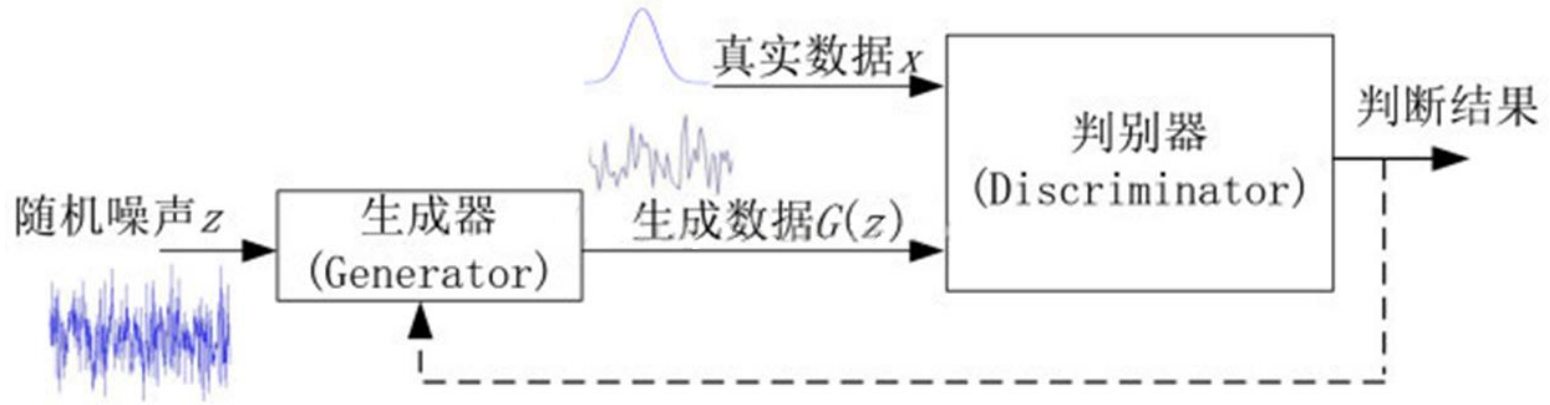


基本概念

- 判别模型
  - 学习数据集中类或标签之间的边界，最终目标是将一类与另一类分开，无法生成新的数据点
  - 学习条件概率分布 P (Y|X)，即在特征 X 出现的情况下标签 Y 出现的概率
  - 逻辑回归、支持向量机、决策树、一般的神经网络
- 生成模型
  - 对生成新数据实例进行建模的一类统计模型
  - 学习联合概率分布 P (X,Y) ，即特征 X 和标签 Y 共同出现的概率，根据贝叶斯公式P(Y|X) = P(X,Y)/P(X)求出条件概率分布作为预测模型
  - 贝叶斯网络、自回归模型、生成对抗网络

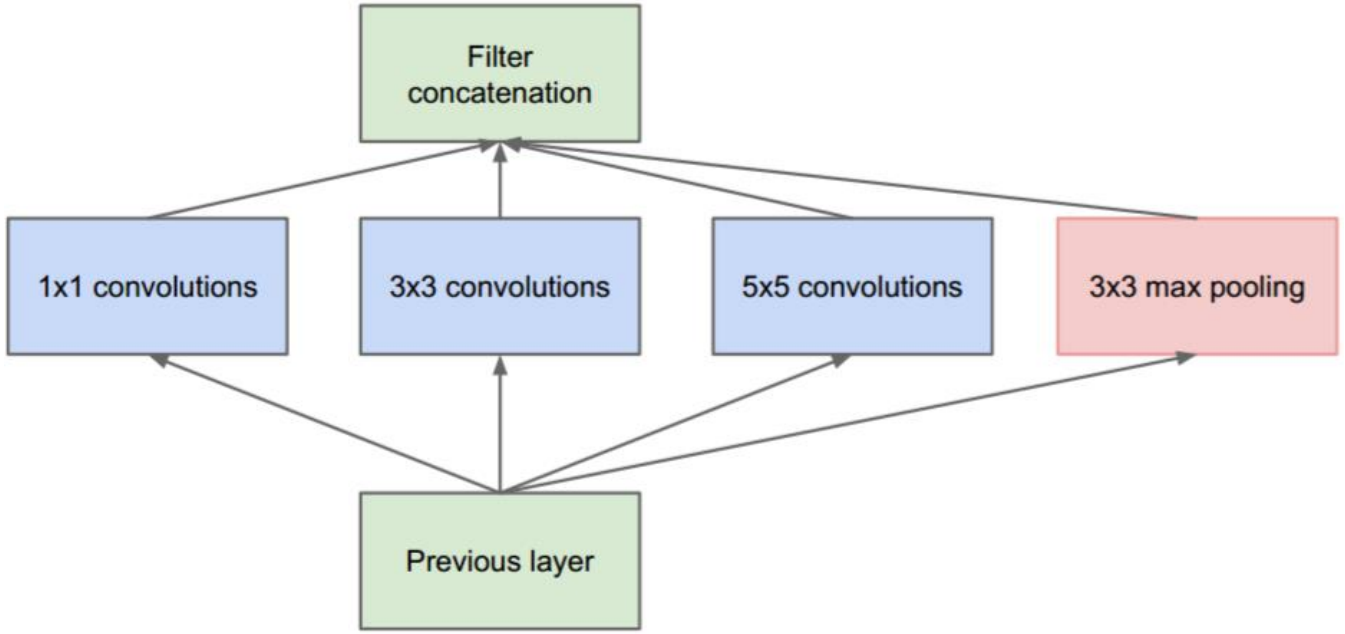
基本概念

- 生成对抗网络（GAN）
  - 通过对抗训练的方式来使得生成网络产生的样本服从真实数据分布
  - 判别网络（D）：尽量准确地判断一个样本是来自于真实数据还是生成网络产生
  - 生成网络（G）：尽量生成判别网络无法区分来源的样本
  - 判别网络和生成网络不断交替训练直至收敛



$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[log(D(x))] + E_{z\sim p_z(z)}[log(1-D(G(z)))]$$

- Inception模型
  - 特征提取的深度网络，基本组成结构有四个：1*1卷积，3*3卷积，5*5卷积，3*3最大池化，最后对四个成分运算结果进行通道上组合
  - 通过多个卷积核提取图像不同尺度的信息，最后进行融合，可以得到图像更好的表征



基本概念

- **Inception Score (IS)**
  - 衡量GAN生成数据的好坏的标准
  - 计算方法，去掉Inception模型最后的pooling层，得到2048维的高层特征

$$\mathbf{IS}(G) = \exp\left(\mathbb{E}_{\mathbf{x}\sim p_g} D_{KL}\left(p(y|\mathbf{x})||p(y)\right)\right)$$

- **Fréchet Inception Distance (FID)**
  - 基于Fréchet distance

$$d^2((m,C),(m_w,C_w)) = ||m-m_w||_2^2 + Trace(C+C_w-2(CC_w)^{1/2})$$

  - m、C表示生成数据的均值和方差，m_w、C_w表示真实数据的均值和方差

基本概念

- **Metropolis-Hasting抽样算法（MH)**
  - 通过目前样本中的最后一个样本来对下一个样本进行抽样
  - 已知：随机变量X服从的分布$p(X|\theta)$与分布参数θ的先验分布$p(\theta)$
  - 1. 随机取得第一个样本θ_1
  - 2. 从$J(\theta|\theta_i)$中抽样，得到θ_{new}，并计算

$$\gamma = \frac{p(\theta_{new}|X)}{p(\theta_i|X)} = \frac{p(X|\theta_{new})p(\theta_{new})}{p(X|\theta_i)p(\theta_i)}$$

  - 如果$\gamma > 1$，则$\theta_{i+1} = \theta_{new}$；否则再取一个$p \sim U(0,1)$的随机数，当$p < \gamma$时$\theta_{i+1} = \theta_{new}$，否则$\theta_{i+1} = \theta_i$
  - 重复第2步，直到获得预设样本数量

算法原理

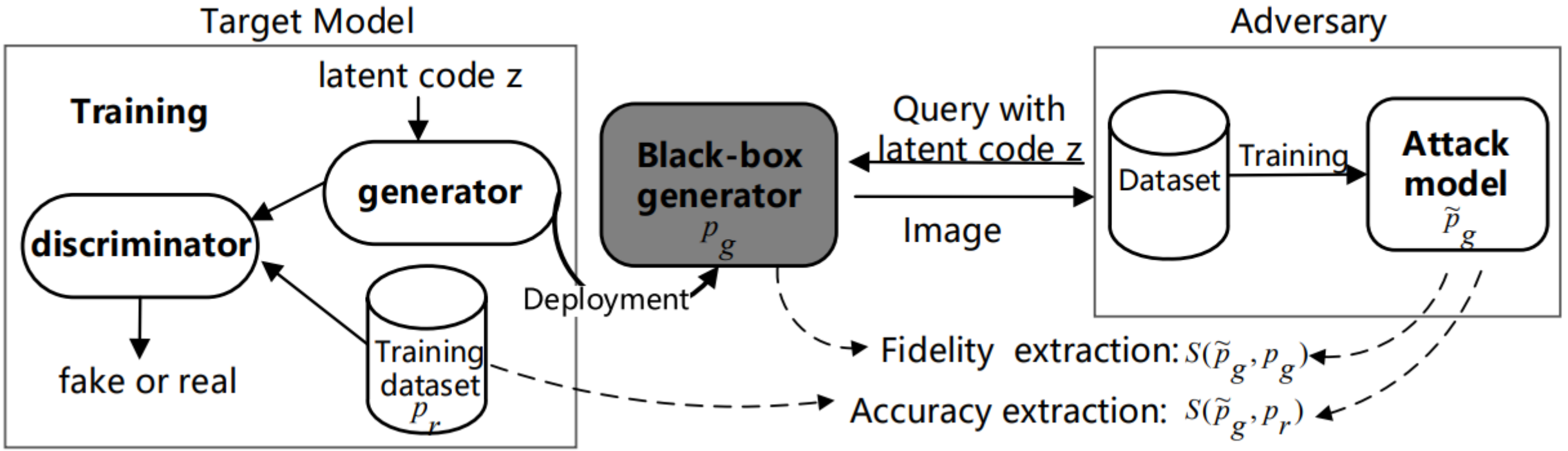
算法原理

T	对生成对抗网络进行模型窃取
I	目标生成对抗网络，潜在向量空间
P	1.从潜在空间中选择潜在向量，通过目标生成对抗网络得到生成数据 2. 利用MH算法从生成数据中采样 3. 使用采样后的数据训练攻击模型
O	与目标模型功能相近的生成对抗网络模型

P	无法获得真实训练数据的分布信息
C	攻击者可以获得目标模型的生成图像、鉴别器
D	获得与真实训练数据分布相近的生成数据
L	ACSAC 2021

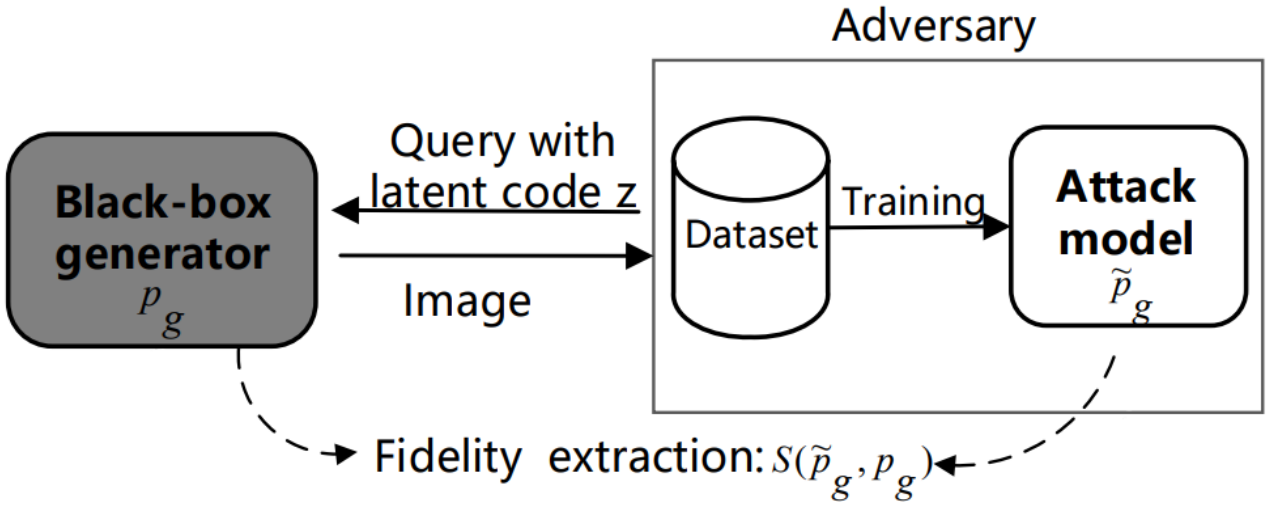
算法原理

- 算法原理图


Target Model
Training
latent code z
generator
discriminator
fake or real
Training dataset p_r
Deployment
Black-box generator p_g
Query with latent code z
Image
Adversary
Dataset
Training
Attack model $\tilde{p}_g$
Fidelity extraction: $S(\tilde{p}_g, p_g)$
Accuracy extraction: $S(\tilde{p}_g, p_r)$

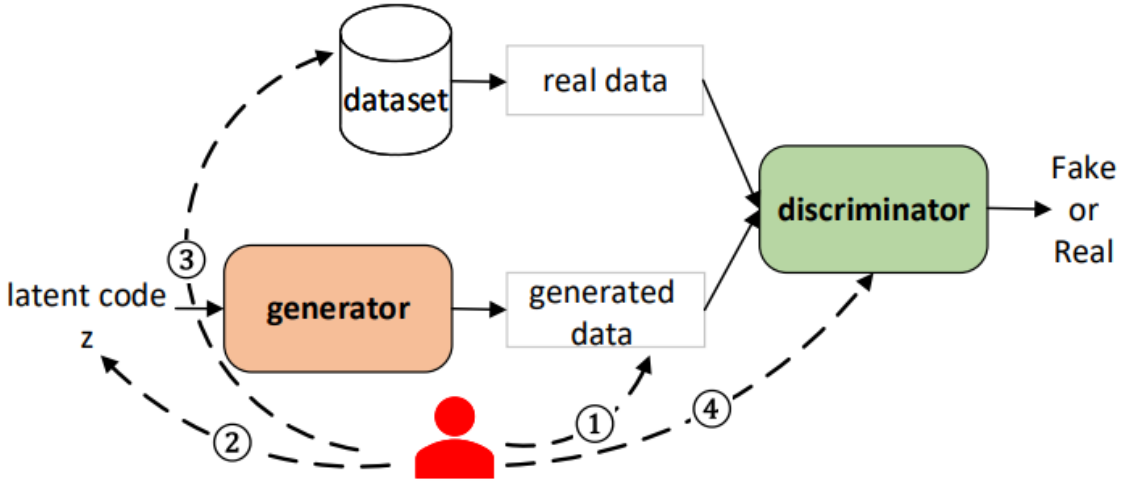
算法原理

- 保真度窃取攻击
  - 目标：窃取目标模型的分布（最小化目标模型与攻击模型生成数据的分布差异）
  - 原理：
    - 从潜在空间中选择潜在向量Z通过目标模型得到生成数据
    - 利用潜在向量Z与生成图像，训练攻击模型
    - 比较同一潜在向量下，目标模型与攻击模型的生成数据分布



实验结果

- 数据集
  - LSUN（卧室、厨房等场景图像数据集）、CelebA（人脸图像数据集）
- 实验条件
  - 攻击者可以获得生成器的生成数据
- 实验设置
  - 目标生成器(PGGAN，SNGAN)
  - 攻击所用生成器（PGGAN，SNGAN）
- 评价方法
  - FID



实验结果

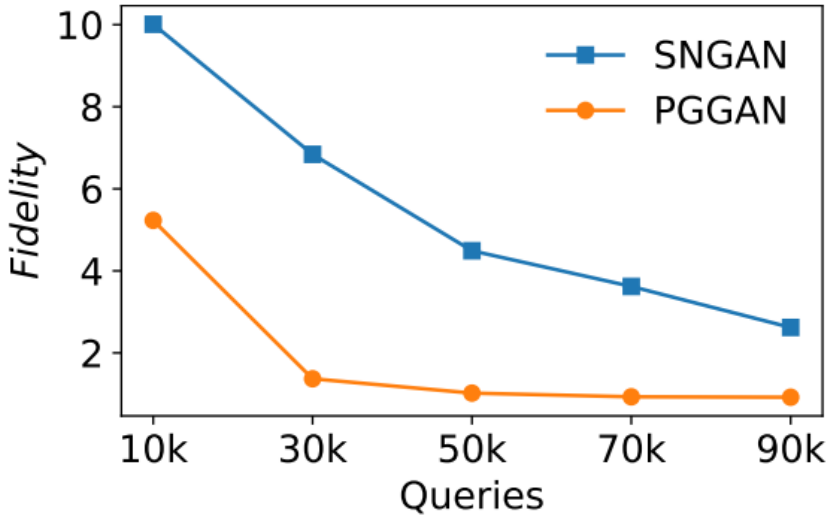
- 实验结果（50K 次查询）
  - 目标模型为 PGGAN 时，使用两种攻击模型都可以获得很好的效果
  - 目标模型为 SNGAN 时，使用 PGGAN 作为攻击模型时 FID 低于使用原模型，这是因为 PGGAN 模型本身结构更加复杂，性能优于 SNGAN，能够更准确的逼近目标模型

Target model	Dataset	FID
SNGAN	LSUN-Church	12.72
SNGAN	CelebA	7.60
PGGAN	LSUN-Church	5.88
PGGAN	CelebA	3.40

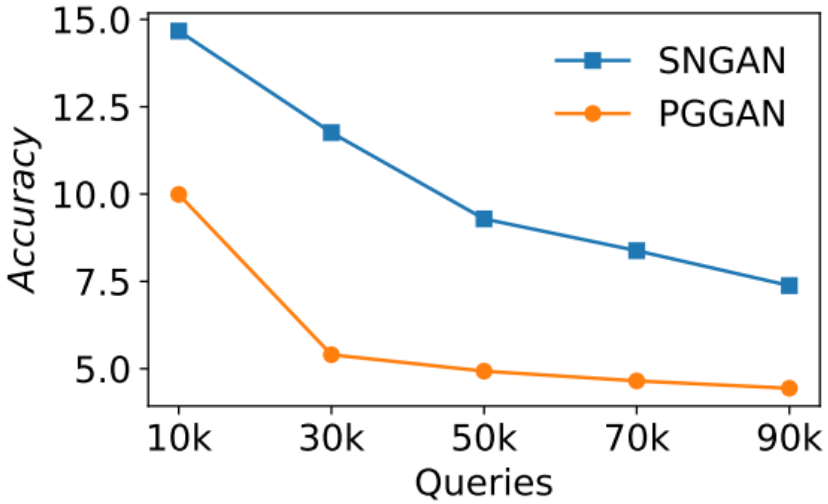
Target model	Attack model	Dataset	*Fidelity* FID($\tilde{p}_g, p_g$)	*Accuracy* FID ($\tilde{p}_g, p_r$)
PGGAN	SNGAN	LSUN-Church	6.11	14.05
	SNGAN	CelebA	4.49	9.29
	PGGAN	LSUN-Church	1.68	8.28
	PGGAN	CelebA	1.02	4.93
SNGAN	SNGAN	LSUN-Church	8.76	30.04
	SNGAN	CelebA	5.34	17.32
	PGGAN	LSUN-Church	2.21	14.56
	PGGAN	CelebA	1.39	9.57

实验结果

- 指标随查询次数的变化情况
  - 随着查询次数的增加而增加，前 50K 次查询时 FID 下降较快
  - 限制查询次数是一种有效的安全防护措施



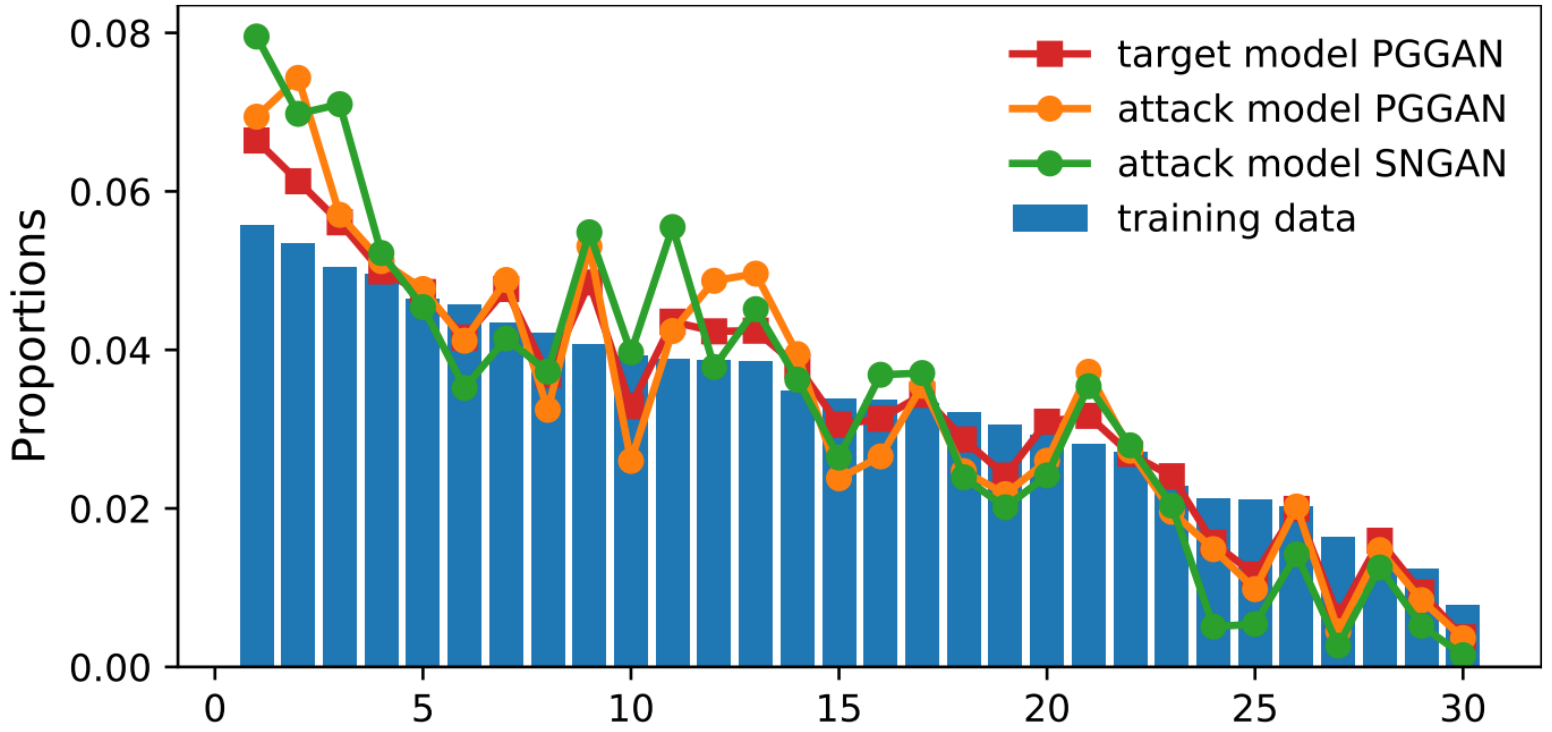
(a) *Fidelity* on CelebA



(b) *Accuracy* on CelebA

实验结果

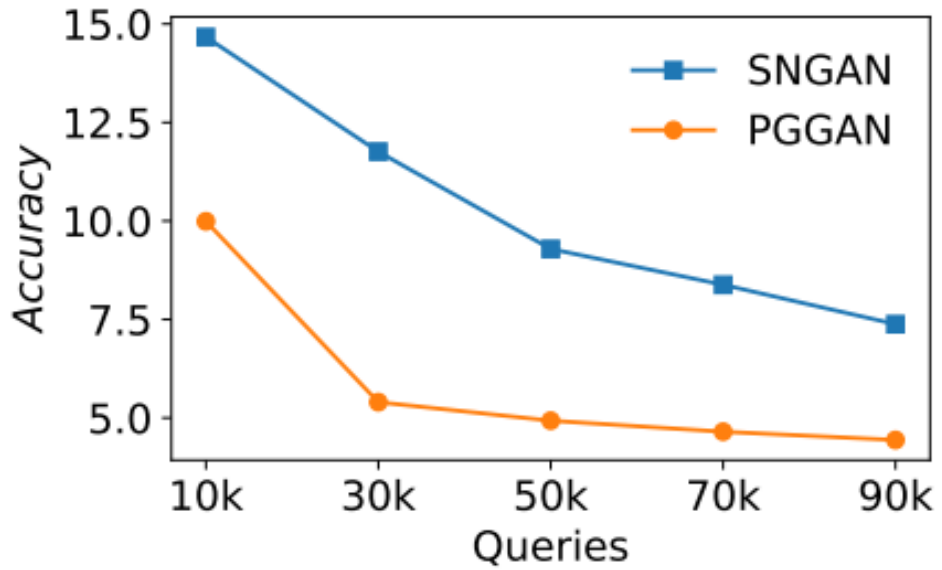
- 什么是模型之间的分布差异?
  - 通过 Inception 模型将数据转化为 2048 维向量，通过 k-means 将这些特征向量聚类成 k 类，计算每个类的比例



算法原理

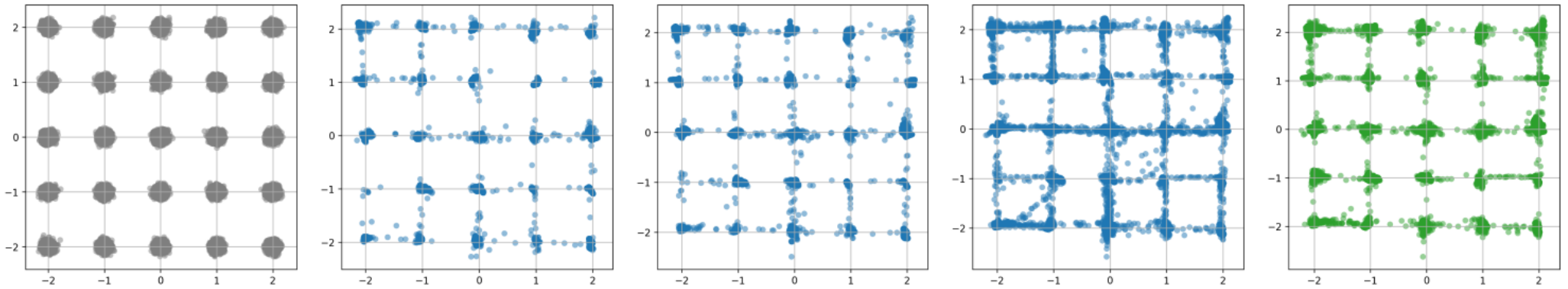
- 准确度窃取攻击
  - 目标：窃取目标模型训练集的分布
  - 难点：缺乏训练数据分布的信息，虽然利用生成模型的分布做近似，但生成模型并未完全学习到真实的训练数据分布，随着查询数量的增加，很快到达极值，后续难以提高
  - 如何获得接近真实数据分布的样本?

Target model	Attack model	Dataset	*Fidelity* FID($\tilde{p}_g, p_g$)	*Accuracy* FID ($\tilde{p}_g, p_r$)
PGGAN	SNGAN	LSUN-Church	6.11	14.05
	SNGAN	CelebA	4.49	9.29
	PGGAN	LSUN-Church	1.68	8.28
	PGGAN	CelebA	1.02	4.93
SNGAN	SNGAN	LSUN-Church	8.76	30.04
	SNGAN	CelebA	5.34	17.32
	PGGAN	LSUN-Church	2.21	14.56
	PGGAN	CelebA	1.39	9.57



算法原理

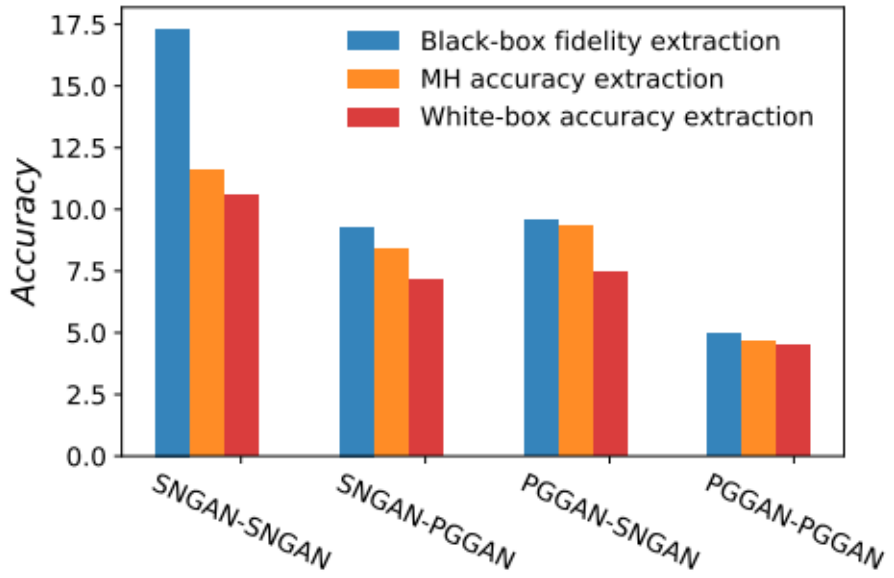
- 准确度窃取攻击
  - 解决方法：获取更多的先验知识——鉴别器信息，利用鉴别器对生成的样本进行子采样（MH采样），使其更接近真实数据分布
  - 图(a)是从二维高斯分布的数据集中抽取的真实数据，(b)-(d)为不同采样条件下生成器生成的数据，采用MH采样方法生成的数据更接近真实数据分布



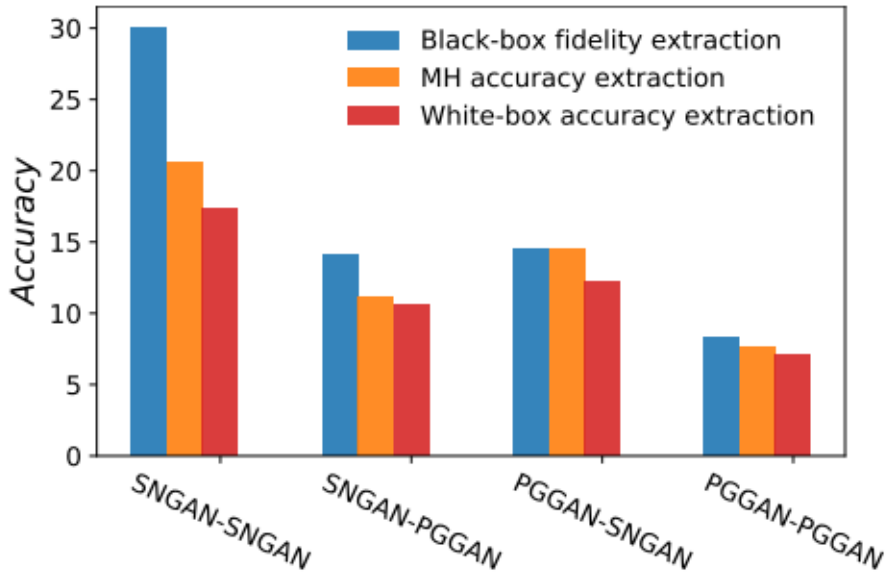
(a) Real Samples 50K　(b) Direct Sampling 5k　(c) Direct Sampling 10K　(d) Direct Sampling 50K　(e) MH Subsampling 50K

实验结果

- CelebA 与 LSUN-Church 数据集上的实验
  - 白盒方法是指在窃取过程中使用了部分真实数据
  - MH 采样方法可以提高实验效果，接近白盒条件下的实验效果



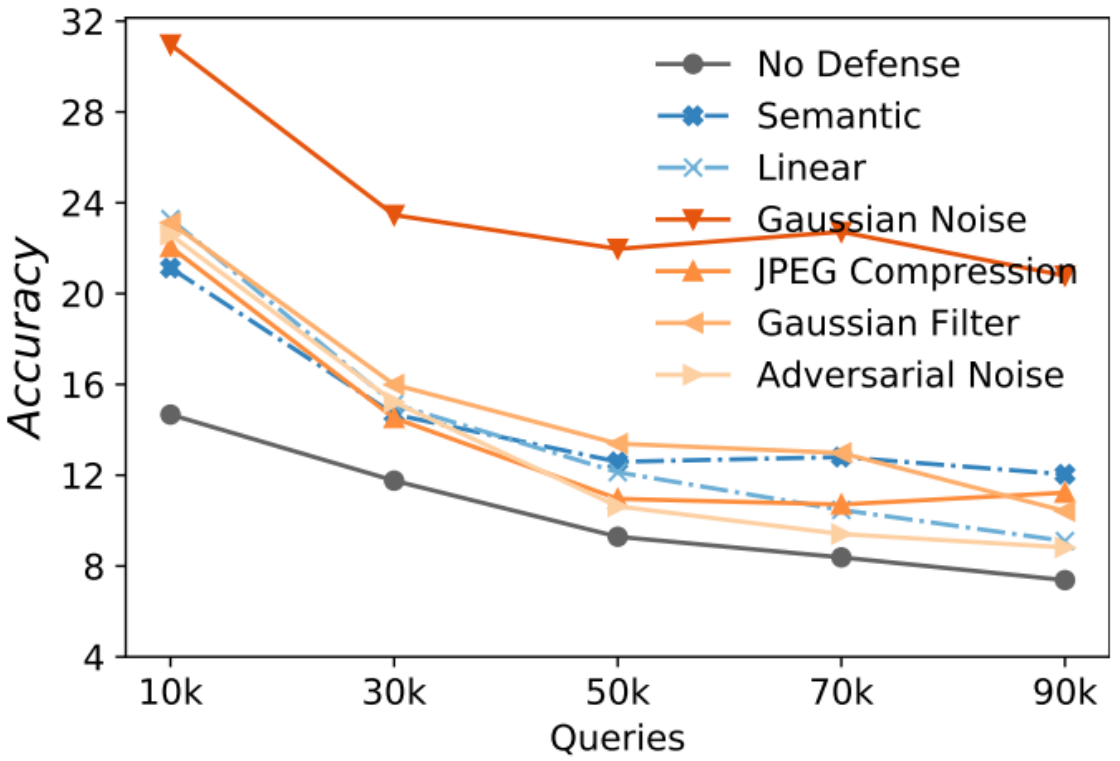
(a) *Accuracy* on CelebA



(b) *Accuracy* on LSUN-Church

算法原理

- 防御方法
  - 面对保真度窃取，限制查询数量
  - 面对准确度窃取，添加基于扰动的防御机制
- 输入扰动
  - 线性插值防御：对输入的n个潜在向量，随机选择两个向量，利用线性插值在其中插入k个向量，重复n/k次，得到n个修改后的潜在向量
- 输出扰动
  - 添加噪声扰动：对输出的生成数据添加噪声



算法原理

算法原理

T	窃取生成式对抗网络模型
I	目标生成对抗网络
P	1.从潜在空间中选择潜在向量，通过目标生成对抗网络得到生成数据 2. 利用潜在向量与生成数据构建向量与生成数据、生成数据与自身分布的目标函数 3. 通过该目标函数训练攻击模型
O	与目标对抗网络功能相似的攻击模型

P	无法获得原始的训练数据及分布信息
C	攻击者可以自由访问目标模型获得其生成数据
D	目标模型训练数据的分布信息无法获得
L	CVPR 2021

算法原理

- 算法原理：
  - 对于目标模型 F_v，将其视为输入 X 与输出 X_s 的函数映射关系与输出 X_s 与分布 S 的函数映射关系

$$F_{\mathcal{V}} = arg\max_{F'_{\mathcal{V}} \in \mathcal{F}} \mathcal{M}_X(X, X_S) \qquad F_{\mathcal{V}} = arg\max_{F'_{\mathcal{V}} \in \mathcal{F}} \mathcal{M}_S(X_S, S)$$

  - 攻击模型 F_A 需要学习这两种映射关系，故其目标函数为：

$$F_{\mathcal{A}} = arg\max_{F'_{\mathcal{A}} \in \mathcal{F}} \mathcal{M}_X(X_{\mathcal{A}}, X_{\mathcal{A}_S})$$

$$F_{\mathcal{A}} = arg\max_{F'_{\mathcal{A}} \in \mathcal{F}} \mathcal{M}_S(X_{\mathcal{A}_S}, S)$$

算法原理

- 算法原理：
  - 实际条件下，无法通过查询获得 $M_S(X_{A_S}, S)$，考虑：

$$F_{\mathcal{A}} = arg\max_{F'_{\mathcal{A}} \in \mathcal{F}} M_{\mathcal{P}}(F_{\mathcal{V}}, F_{\mathcal{A}}, X) \quad s.t.$$

$$if \quad F_{\mathcal{A}}(X) \approx F_{\mathcal{V}}(X) \quad then$$

$$F_{\mathcal{A}} = arg\max_{F'_{\mathcal{A}} \in \mathcal{F}} \mathcal{M}_X(X_{\mathcal{A}}, X_{\mathcal{A}_S}) \quad and$$

$$F_{\mathcal{A}} = arg\max_{F'_{\mathcal{A}} \in \mathcal{F}} \mathcal{M}_S(X_{\mathcal{A}_S}, S)$$

  - 由此获得攻击模型的目标函数进行模型训练

实验结果

- 数据集
  - Monet2Photo、Selfie2anime、LFW、Landscape
- 实验条件
  - 目标模型的训练数据及分布信息未知
  - 攻击者可以通过查询获得目标模型的生成数据
- 实验设置
  - 生成模型使用 CycleGAN
  - 攻击模型使用 Pix2pix（一种CGAN）
- 评价方法
  - FID

实验结果

- 实验结果

Output Distribution	Test Set	Compared Distribution	$\mathcal{V}$ Output		Monet Paintings
$\mathcal{V}$ Output	Monet2Phto Test		N/A	(A)	58.05
	Landscape		N/A		49.59
$\mathcal{A}$ Output	Monet2Phto Test	(B)	42.86	(C)	61.62
	Landscape		15.38		53.99

Output Distribution	Test Set	Compared Distribution	$\mathcal{V}$ Output		Anime Faces
$\mathcal{V}$ Output	Selfie2Anime		N/A	(A)	69.02
	LFW Test		N/A		56.06
$\mathcal{A}$ Output	Selfie2Anime	(B)	138.61	(C)	137.68
	LFW Test		19.67		69.42

优劣分析

- 优势
  - 两种算法都利用训练替代模型的方法，借助查询时目标生成模型的生成数据构建攻击模型，无需目标生成模型的真实训练数据
  - 第一种算法对生成的数据进行采样，使得攻击模型的训练数据与目标模型的训练数据分布更加近似，获得了较好的窃取效果
- 不足
  - 由于查询数据的随机性，生成数据的不可控，需要生成大量数据进行目标生成模型的输入，获得其生成数据
  - 第一种算法需要目标生成模型的鉴别器等信息，现实条件下难以满足

应用总结

应用总结

- 算法的应用领域
  - 模型的安全性检验
  - 模型的功能免费使用
  - 通过模型窃取获取目标模型的参数信息，应用到对抗样本生成、成员推理攻击等下游任务

- 未来的发展
  - 模型窃取方法的可迁移性
  - 目标模型先验知识的依赖性
  - 查询次数限制、输入输出被扰动等防御措施下的模型窃取

参考文献

[1] Hu H, Pang J. Stealing machine learning models: Attacks and countermeasures for generative adversarial networks[C]//Annual Computer Security Applications Conference. 2021: 1-16.

[2] Szyller S, Duddu V, Gröndahl T, et al. Good Artists Copy, Great Artists Steal: Model Extraction Attacks Against Image Translation Generative Adversarial Networks[J]. arXiv preprint arXiv:2104.12623, 2021.

[3] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2818-2826.

[4] Tramèr F, Zhang F, Juels A, et al. Stealing machine learning models via prediction apis[C]//25th USENIX Security Symposium. 2016: 601-618.

谢谢！

大成若缺，其用不弊。大盈若冲，其用不穷。大直若屈。大巧若拙。大辩若讷。静胜躁，寒胜热。清静为天下正。