Beijing Forest Studio

北京理工大学信息系统及安全对抗实验中心

深度神经网络模型窃取检测方法

硕士研究生 张辰龙

2023年03月05日

内容提要

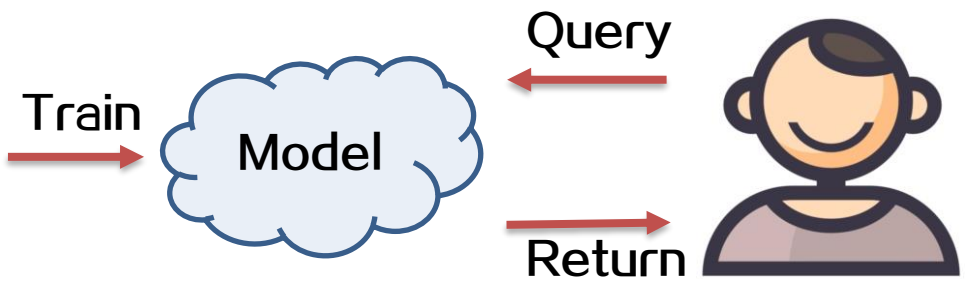
- 背景简介
- 基本概念
- 算法原理
- 优劣分析
- 应用总结
- 参考文献

背景简介

- 预期收获
  - 了解深度神经网络模型窃取基本概念
  - 理解深度神经网络模型窃取检测的算法原理及其理论问题
  - 理解深度神经网络模型窃取检测的发展过程
  - 了解深度神经网络模型窃取检测在网络安全领域中的应用

背景简介

- 云端模型
  - 深度神经网络技术发展迅速，在图像识别、自动驾驶等领域发挥重要作用
  - 深度神经网络模型训练过程繁琐、花销昂贵
  - 微软、谷歌等大型公司将模型部署在云端服务器，仅向用户提供预测接口
  - 研究表明，预测接口仍能泄露模型的大量信息



云端模型并不安全！

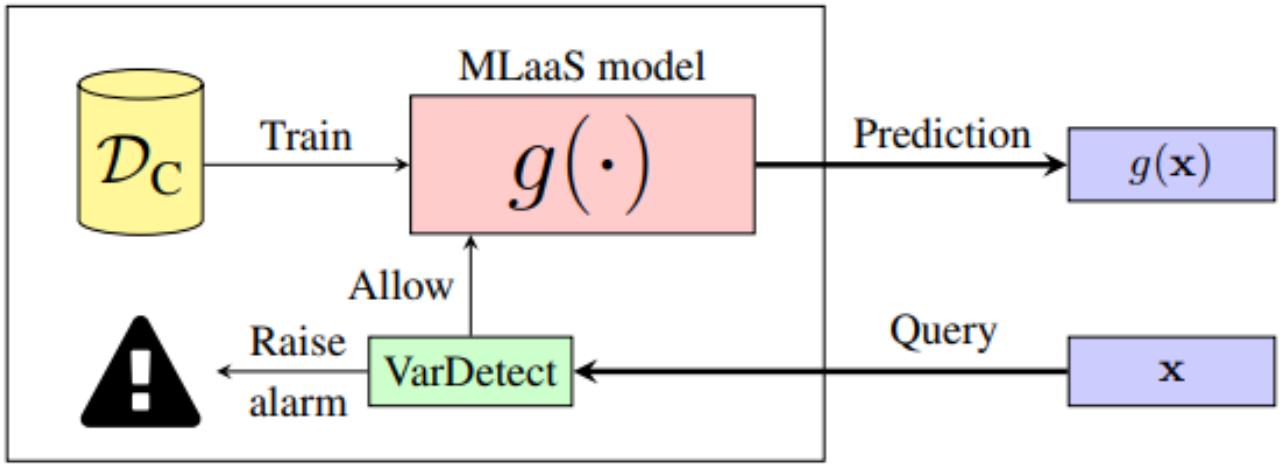
基本概念

- 模型窃取检测中的关键词
  - 守方
    - 目标模型：部署在云端的模型
    - 查询样本：通过预测接口发送给目标模型的样本
    - 检测器：识别攻击者和正常用户所用查询样本的区别，检测攻击者
    - 正常数据集：训练目标模型所用的训练集
    - 正常样本：取自正常数据集的样本
  - 攻方
    - 替代模型：攻击者获得的与目标模型功能相似的模型
    - 攻击者：构造特殊的查询样本，窃取目标模型信息，训练替代模型
    - 攻击者数据集：攻击者构造的特殊的查询样本的集合
    - 攻击样本：取自攻击者数据集的样本

反者道之动，弱者道之用！

基本概念

- 模型窃取
  - 概念：攻击者利用目标模型查询接口泄露的信息，窃取目标模型的参数或功能
  - 流程：攻击者构造无标签的“攻击者数据集”，利用目标模型的查询接口对数据集添加标签，利用带标签的“攻击者数据集”训练替代模型
  - 目的：
    - 免费使用模型功能
    - 进行白盒对抗攻击
  - 危害：
    - 损害模型拥有者的商业利益
    - 侵犯模型的隐私信息



保护模型安全刻不容缓！

基本概念

- 模型窃取分类
  - 生成式：
    - 依据预测结果构造生成模型，常见有VAE、GAN等
    - 利用生成模型产生非真实数据，数据具有随机性

  - 半生成式：
    - 攻击者拥有少量正常数据集数据
    - 利用已有数据产生半真实数据，常用对抗样本类方法

  - 非生成式：
    - 从公共数据集上取得真实数据，与正常样本分布不同

- 真实样本
  -

模型窃取使用的攻击样本特征不统一，检测难度大！

算法原理——PRADA

【IEEE S&P】
PRADA: protecting against DNN model stealing attacks

算法原理——PRADA

T	检测模型窃取行为
I	一组连续的面向目标模型的查询样本
P	1. 保存所有的查询样本 2. 对新输入的样本，计算该样本与历史样本间的最小距离 3. 查询样本达到一定数量时计算所有最小距离的分布特征
O	一组连续的查询样本是否为模型窃取行为

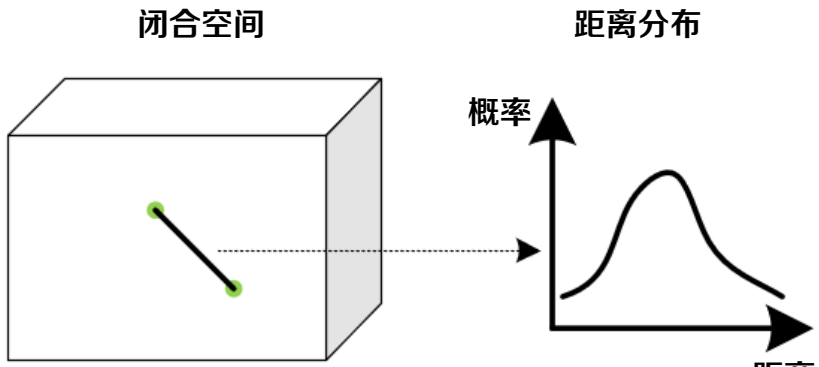
P	无法准确区分半生成式攻击样本与正常样本
C	攻击样本与正常样本均连续，不存在两类样本穿插查询
D	分析攻击样本序列与正常样本序列的分布差异
L	IEEE S&P 2019

算法原理——PRADA

- 算法原理
  - 算法思想

Shapiro-wilk检测是一类正态性符合度检测，输出值在0-1之间，越接近1，正态性符合度越好

    - 在闭合空间中随机取得两点之间的距离符合正态分布
    - 正常样本随机取自闭合空间
    - 半生成式攻击样本间距小，存在分布差异
  - 算法步骤
    - 计算新输入的样本与历史样本之间的最小距离d
    - 若d符合范围要求，将d存储至最小距离集合D
    - 集合D中元素数量超过某一数目时，使用Shapiro-Wilk算法计算D中元素的正态分布符合度
    - 设定检测阈值t，Shapiro-Wilk计算值小于阈值t时，判定检测出模型窃取攻击

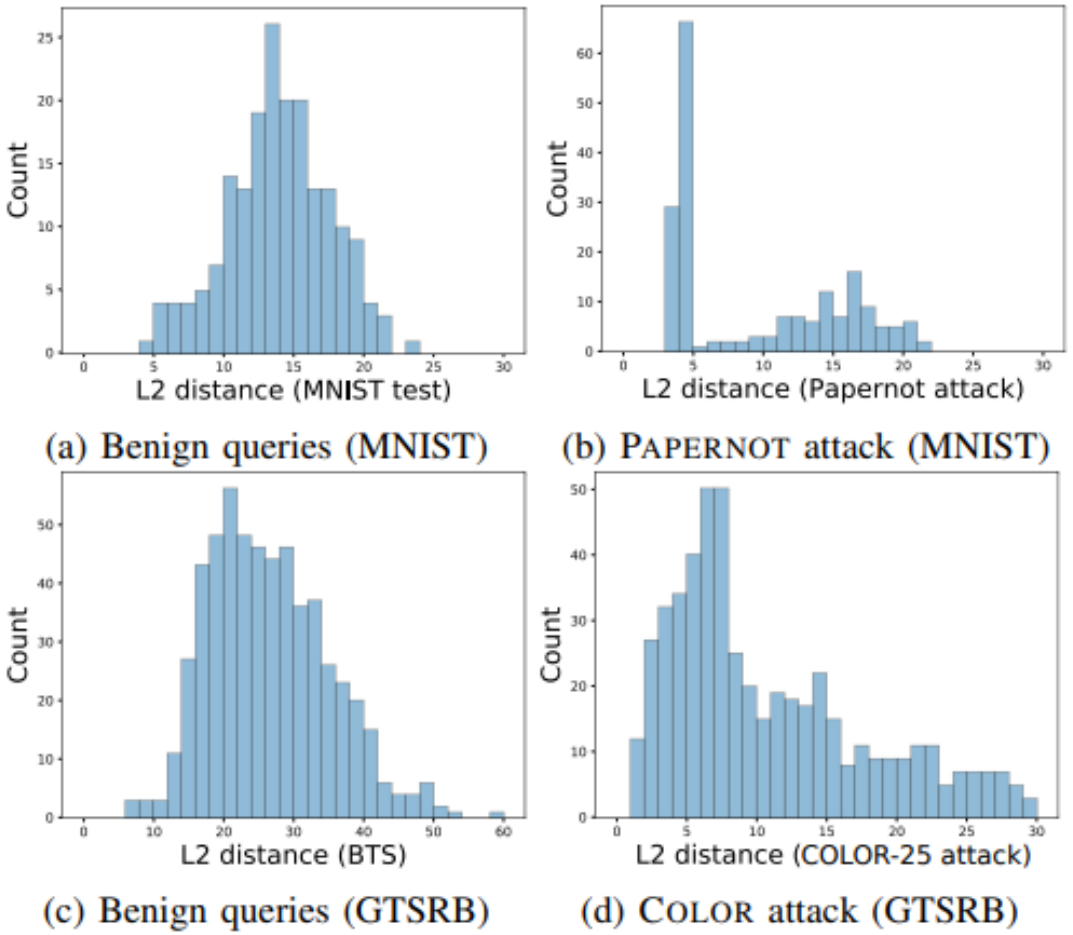


算法思想图


半生成样本对比图

实验结果——PRADA

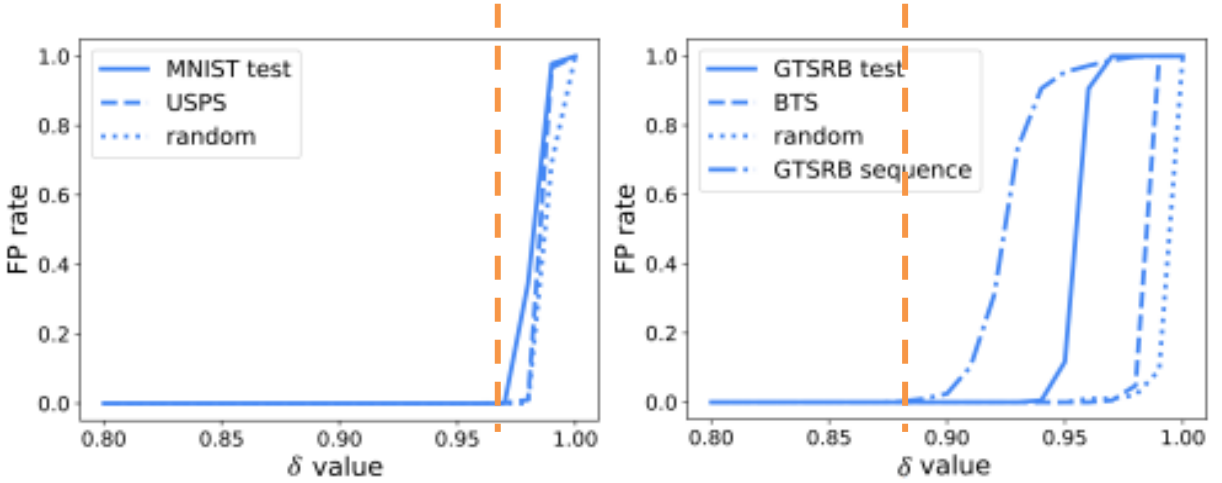
- 实验结果
  - 目标模型：MNIST与GTSRB模型
  - 正常行为：随机采用目标模型测试集样本查询和采用与目标模型相关的数据集查询
  - 攻击行为：PaperNot attack与COLOR attack
  - 实验结果：正常行为距离集合中距离的分布直方图呈正态分布，攻击行为距离集合中的距离值集中在较小距离处



在样本距离分布上攻击行为与正常行为差异显著！

实验结果——PRADA

- 实验结果
  - 检测阈值t：通过实验确定检测阈值t，使检测阈值有效区分正常行为与模型窃取
  - 检测指标speed：检测到模型窃取时攻击者的样本查询数量
  - 检测结果
    - 给定检测阈值t，检测器能够有效区分正常用户和攻击者，FPR < 0.6%
    - 对半生成式攻击（PAP）检测速度较快，在500次查询内完成检测



检测阈值对正常行为的影响

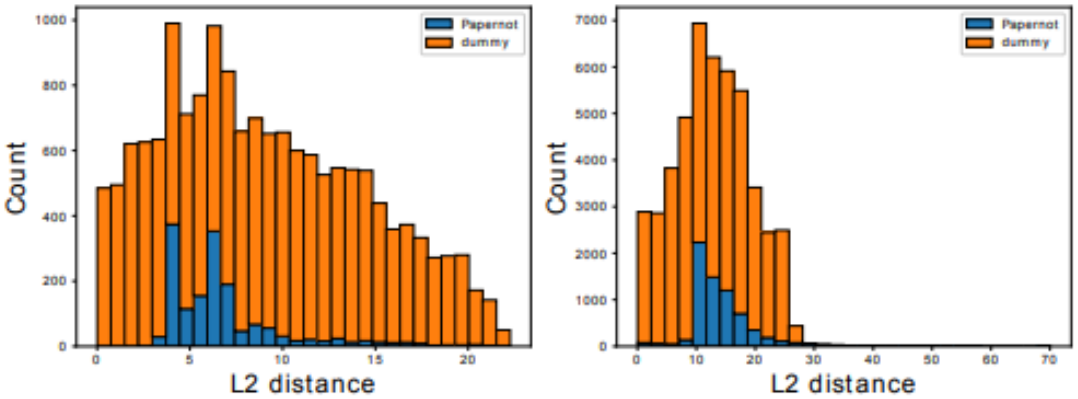
Model (δ value)	FPR	Queries made until detection			
		TRAMER	PAP.	T-RND	COLOR
MNIST (0.95)	0.0%	5,560	120	140	-
MNIST (0.96)	0.0%	5,560	120	130	-
GTSRB (0.87)	0.0%	5,020	430	missed	550
GTSRB (0.90)	0.6%	5,020	430	missed	480
GTSRB (0.94)	0.1%*	5,020	430	440	440

对不同攻击的检测结果

实验分析——PRADA

- 实验分析
  - 理论问题
    - 样本间距离的计算忽视了样本变换的影响
    - 理想化的假设攻击行为中的所有样本均为攻击样本，忽视了穿插正常样本的情况
    - 样本间距离特征没有包括所有类型的攻击
  - 应用问题
    - 攻击者查询前对样本进行一定的变换（旋转、翻转等）可以绕过检测
    - 会被伪查询绕过（攻击者向恶意样本中穿插正常样本以维护距离分布）
    - 会被非生成式攻击绕过（攻击者从公共数据集取得的真实数据）

样本变换对距离计算的影响



(a) MNIST model (b) GTSRB model

Attack	Model: PAPERNOT	MNIST ($\delta = 0.96$): *T-RND*	MNIST ($\delta = 0.96$): TRAMER
Original queries	1,600	1,600	10,000
Additional queries	14,274	4,764	79,980
Overhead	+890%	+300%	+800%

绕过检测器会付出更大的代价！

算法原理——OOD

【EDSMLS】
Extraction of Complex DNN Models:
Real Threat or Boogeyman?

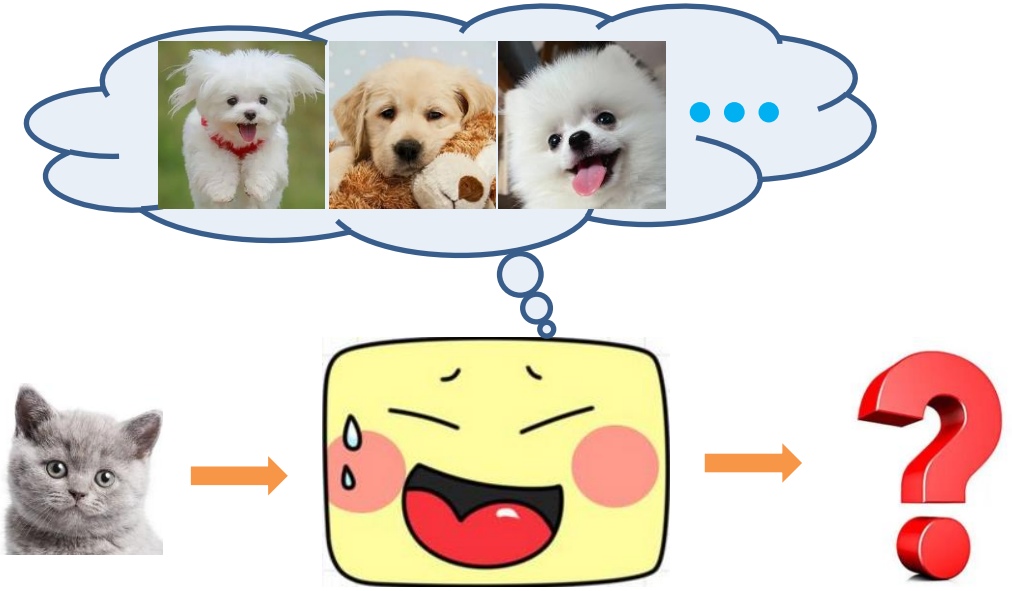
算法原理——OOD（Out Of Distribution）

T	检测模型窃取行为
I	一组面向目标模型的查询样本
P	1. 使用正常样本与大量公共数据集样本训练二分类器 2. 对新输入的样本，采用二分类器判断是否为分布外样本 3. 使用计数器记录分布外样本数量
O	一组查询样本是否存在模型窃取行为

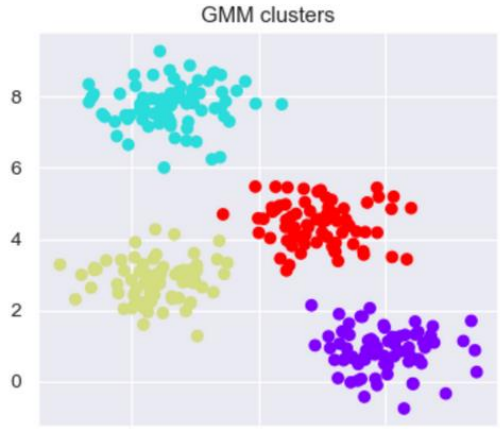
P	无法准确区分非生成式攻击样本与正常样本
C	攻击样本从公共数据集中选取（非生成式攻击）
D	分析攻击样本与正常样本的样本个体差异
L	EDSMLS 2020

算法原理——OOD

- 算法原理
  - 算法思想
    - 深度神经网络模型基于Closed world假设进行训练，在面对Open world样本时会以高置信度给出错误的预测结果
    - 非生成式攻击利用Open world样本获取预测信息
    - 两类样本分布不同
  - 算法步骤
    - 构建“分布内样本”数据集（目标模型训练集）
    - 构建“分布外样本”数据集（公共数据集，如ImageNet数据集）
    - 训练二分类模型



训练有素的狗狗分类模型



样本分布的二维直观图

实验结果——OOD

- 实验结果
  - 目标模型：Caltech、CUBS等模型
  - 正常行为：随机采用目标模型测试集样本查询
  - 攻击行为：使用公共数据集样本查询
  - 实验结果：训练所得二分类模型能够以较高的准确率（> 90%）检测出分布内样本（TNR）和分布外样本（TPR），且优于一些基本的OOD检测算法
  - 问题：算法在Caltech数据集上的检测准确率仅有60%左右

$\mathcal{A}$'s transfer set	In-dist. dataset	Ours		Baseline/ODIN/Mahalanobis	
		TPR	TNR	TPR (at TNR Ours)	TPR (at TNR 95%)
ImageNet	Caltech	63%	56%	87%/**88%**/59%	13%/11%/5%
	CUBS	**93%**	**93%**	48%/54%/19%	39%/43%/12%
	Diabetic5	**99%**	**99%**	1% /25%/98%	5%/49%/**99%**
	GTSRB	**99%**	**99%**	42%/56%/71%	77%/94%/89%
	CIFAR10	**96%**	**96%**	28%/54%/89%	33%/60%/91%
OpenImages	Caltech	61%	59%	**83%**/**83%**/6%	11%/11%/6%
	CUBS	**93%**	**93%**	47%/50%/14%	37%/44%/14%
	Diabetic5	**99%**	**99%**	1%/21%/**99%**	4%/44%/**99%**
	GTSRB	**99%**	**99%**	44%/64%/75%	76%/93%/87%
	CIFAR10	**96%**	**96%**	27%/56%/92%	33%/62%/95%

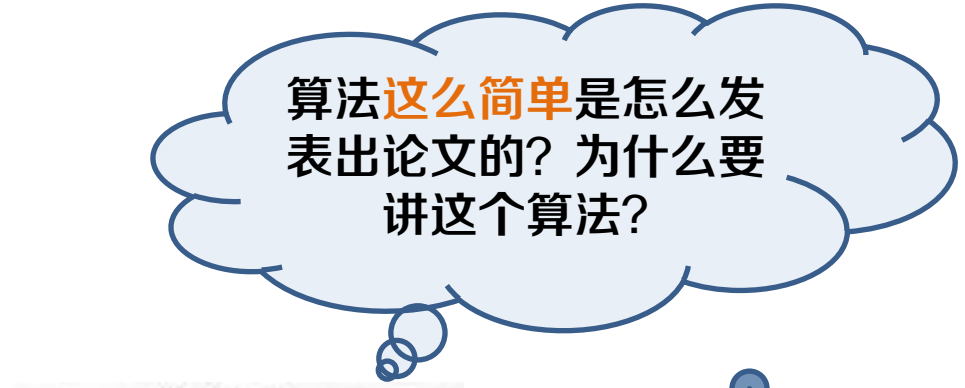


Caltech数据集中的一些样本

OOD检测在两个分布相似的数据集上检测效果不好！

实验结果——OOD

- 实验结果
  - 目标模型：Caltech、CUBS等模型
  - 正常行为：随机采用目标模型测试集样本查询
  - 攻击行为：使用公共数据集样本查询
  - 实验结果：训练所得二分类模型能够以较高的准确率（> 90%）检测出分布内样本（TNR）和分布外样本（TPR），且优于一些基本的OOD检测算法
  - 问题：算法在Caltech数据集上的检测准确率仅有60%左右



Answer1

检测算法只是这篇论文的一章

Answer2

为引出其他领域的OOD检测算法，体现非生成式攻击检测的困难

OOD检测在两个分布相似的数据集上检测效果不好！

实验分析——OOD

- 实验分析
  - 理论问题
    - 算法对攻击样本有明确的定义，使用ImageNet训练二分类模型，并用于检测ImageNet，不现实
    - 算法忽视了正常样本和攻击样本的体量严重不均衡，在非生成式攻击中，攻击样本的数量远大于正常样本

形而上者谓之道
形而下者谓之器

- 常用OOD方法（对抗样本领域）
  - Softmax-based算法：依据模型预测输出的最大的softmax概率判断
  - Generative model类算法：采用AE或VAE对样本重构，根据重构损失判断
- 总结
  - 对非生成式攻击的检测难度过大，后续没有算法针对性的检测
  - 非生成式攻击数据集与正常数据集差距过大时，攻击效率不高

算法原理——SEAT

【ACM CCS】
SEAT: Similarity Encoder by Adversarial Training for Detecting Model Extraction Attack Queries

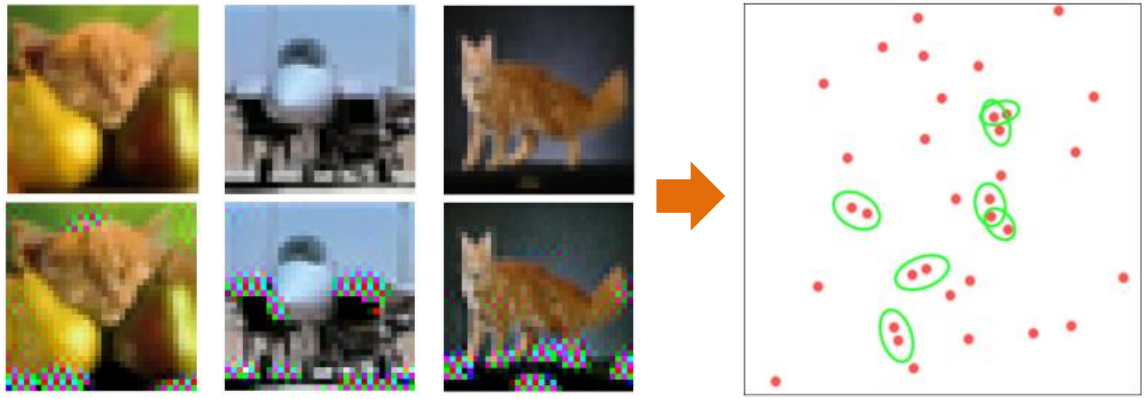
算法原理——SEAT

T	检测模型窃取行为
I	一组面向目标模型的查询样本
P	1. 利用正常数据集训练Similarity Encoder 2. 记录所有查询样本的SE编码后的数据 3. 当有新查询时，首先对其SE编码，其次在历史编码中寻找与其距离相近的样本，构成“样本对” 4. 构造计数器，记录“样本对”数量
O	一组查询样本是否存在模型窃取行为

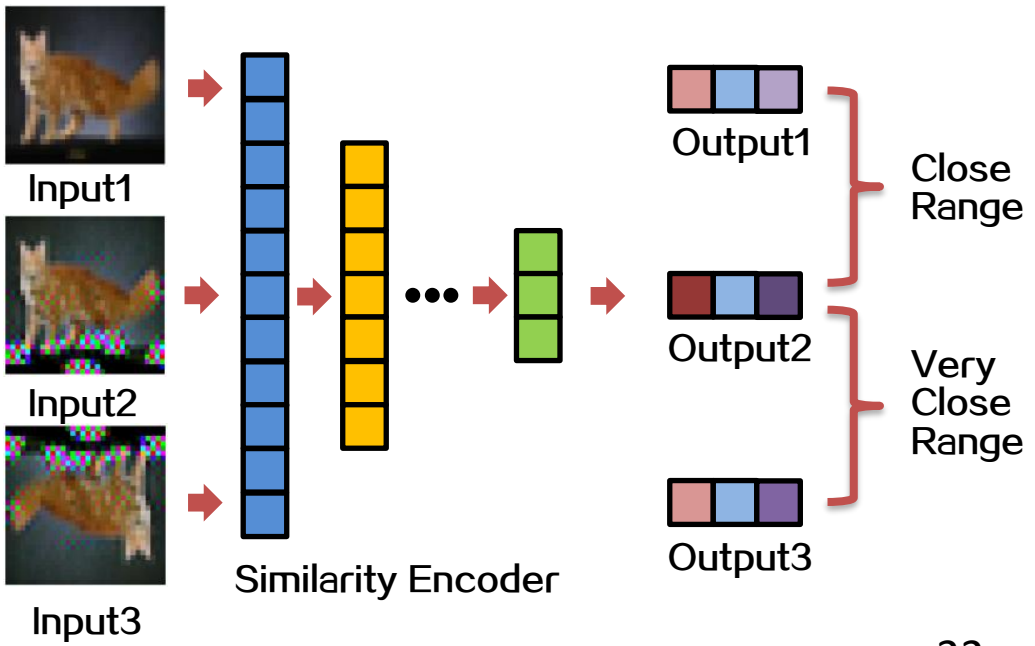
P	无法准确区分半生成式攻击样本与正常样本
C	攻击者拥有少量正常数据集样本
D	分析攻击样本序列与正常样本序列的样本间关联差异
L	ACM CCS 2021【CCF A类】

算法原理——SEAT

- 算法原理
  - 算法思想
    - 半生成式攻击样本间关联密切
    - 攻击期间存在大量距离相近的样本对
  - 算法步骤
    - 使用目标模型的训练集样本训练Similarity Encoder，使视觉相似的样本编码后码间距离相近（解决了PRADA的理论问题1）
    - 保存已查询样本的SE编码值
    - 计算新输入样本编码值与历史样本编码值的L2距离，将样本与其最近且距离小于阈值的样本配对
    - 计算查询样本中码间距离相近的样本对数



半生成式攻击生成的样本和原样本很相似　　SE编码空间



实验结果——SEAT

- 实验结果
  - 评价指标

巧妙的避开了多用户联合攻击的问题

    - Accounts：检测到攻击时，对该账户进行封禁，攻击者必须更换账户，记录攻击者完成攻击所需要的账户数
    - Acc：攻击者完成攻击后，替代模型的准确率
  - 实验结果
    - 在半生成式模型窃取中，攻击者使用各种数据集作为已有样本均会被封禁至少29次
    - 检测器对于使用各类查询样本的正常用户检测的错误率均小于0.05%

检测效率高，对正常用户误判少

Seed Set	# Accounts	Ex. Acc.
CIFAR10	41	75%
TinyImageNet	29	62%
ImageNet1k	35	58%
CIFAR100	43	61%
SVHN	48	33%
CINIC10	30	70%
Indoor67	29	55%
CUBS200	41	34%
Caltech256	65	61%

Query Set	FPR
CIFAR10	0.012%
TinyImageNet	0.007%
ImageNet1k	0.010%
CIFAR100	0.013%
SVHN	0.026%
CINIC10	0.009%
Indoor67	0.006%
CUBS200	0.017%
Caltech256	0.027%
KaggleFrames	0.050%
GTSRB	0.010%
LFW	0.012%
VGG-Flower17	0.037%

实验结果——SEAT

- 实验结果
  - Adaptive Attack（白盒攻击）
    - Query Filtering：模拟构建目标模型的Similarity Encoder，查询前先判断该样本与历史样本是否相距过近
    - Query Blinding：采用样本变换函数，使查询样本的变化尽可能大，而目标模型的预测结果尽可能不变
  - 实验结果
    - 算法能够有效应对Adaptive Attack
    - 正常攻击下PRADA效果最好（封禁最多账户）
    - Adaptive Attack攻击下，SEAT效果最好

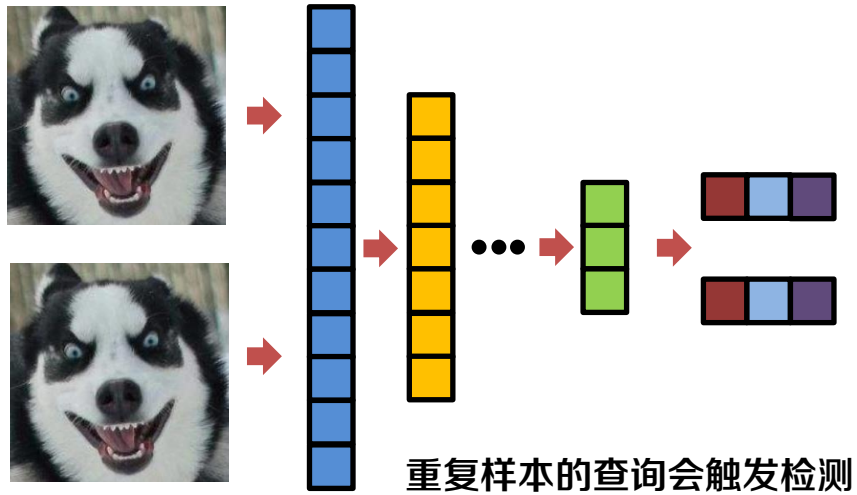
Adaptive Schemes	Strategies	# Accounts	Ex. Acc.
Non-adaptive	N/A	41	75%
Query Filtering	VGG16 + *CINIC10*	42	75%
	5-layer + *CIFAR10 seed*	**33**	**75%**
Query Blinding	Crop	21	71%
	Brightness	38	61%
	Scale	24	73%
	Rotate	41	75%
	Contrast	14	61%
	Uniform	64	62%
	Gaussian	65	60%
	Translate	**23**	**73%**
	Auto-encoder	31	60%

	Random	OOD[1]	SD[7]	PRADA[22]		SEAT (Ours)	
Seed Images	Vanilla	Vanilla	Vanilla	Vanilla	Adaptive	Vanilla	Adaptive
CIFAR10	5	5	1	203	2	41	**23**
TinyImageNet	5	5	1	218	5	29	**19**
CINIC10	5	1	1	66	1	70	**16**

有效应对Adaptive Attacks是因为SE及其随机性

实验分析——SEAT

- 实验分析
  - 情景假设
    - 正常用户一不小心重复查询了部分数据，导致被认定为模型窃取
    - 正常用户长时间不间断查询，导致被认定为模型窃取
    - 攻击者使用公共数据集查询，绕过检测
  - 理论问题
    - 对恶意样本的特征定义宽松，囊括了少量正常样本
    - 忽视了计数方式中“数量”的递增性质
    - 忽视了取样自公共数据集的非生成式攻击样本特征



重复样本的查询会触发检测条件，但并不是模型窃取

一对双胞胎的照片

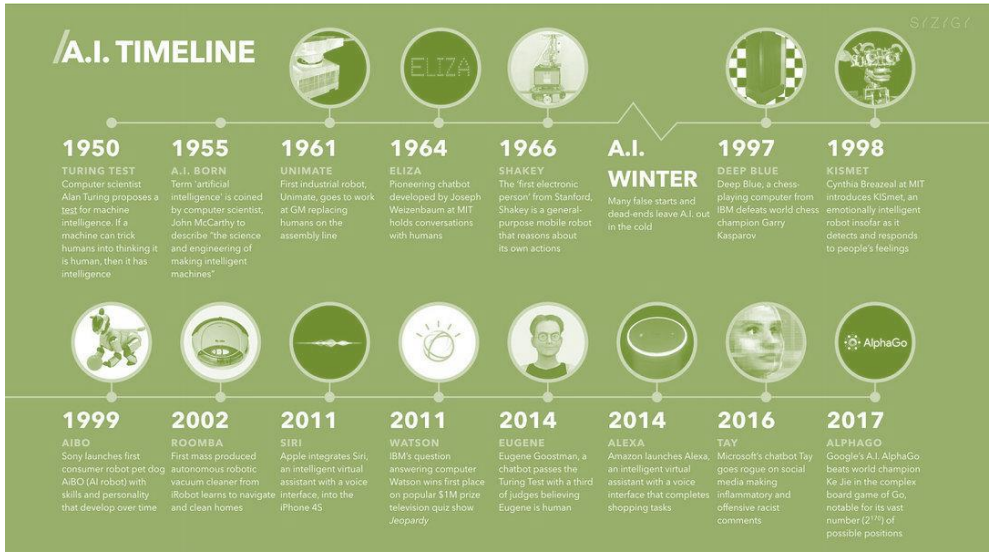
正常查询中必定会遇到相似样本

SEAT算法对正常用户的查询存在影响！

算法思想

- 强统计思想（PRADA）
  - 强统计思想要求连续的查询样本符合某一分布
  - 对攻击者行为考虑片面，忽略了伪查询的存在
- 弱统计思想（SEAT）
  - 从一组样本中寻找部分样本间的关联信息
- 非统计思想（OOD）
  - 仅考虑样本个体特征，丢弃样本统计特征
- 计数思想（SEAT, OOD）
  - 对违反规则的样本计数，以数量为检测阈值

有无相生，难易相成
长短相形，高下相倾
音声相和，前後相随

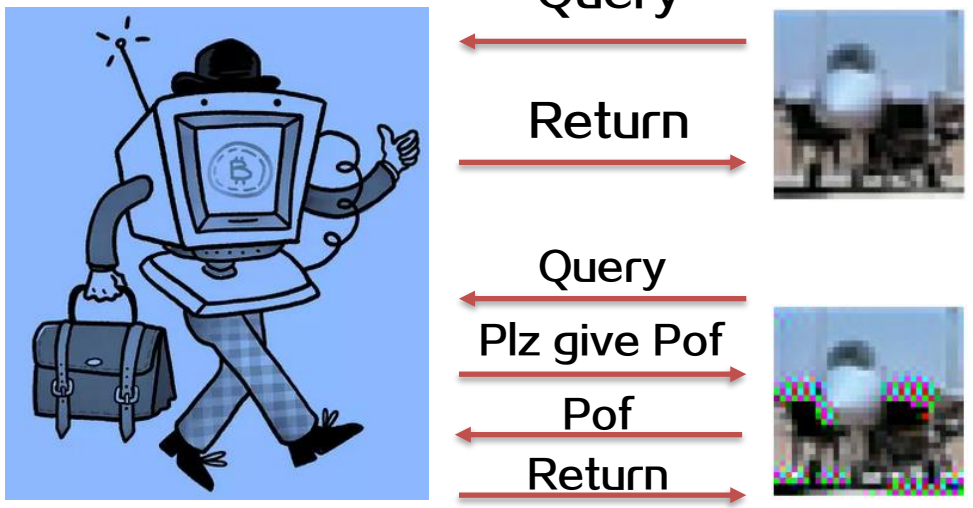


各类算法的思想也体现矛盾的对立统一！

总结

总结

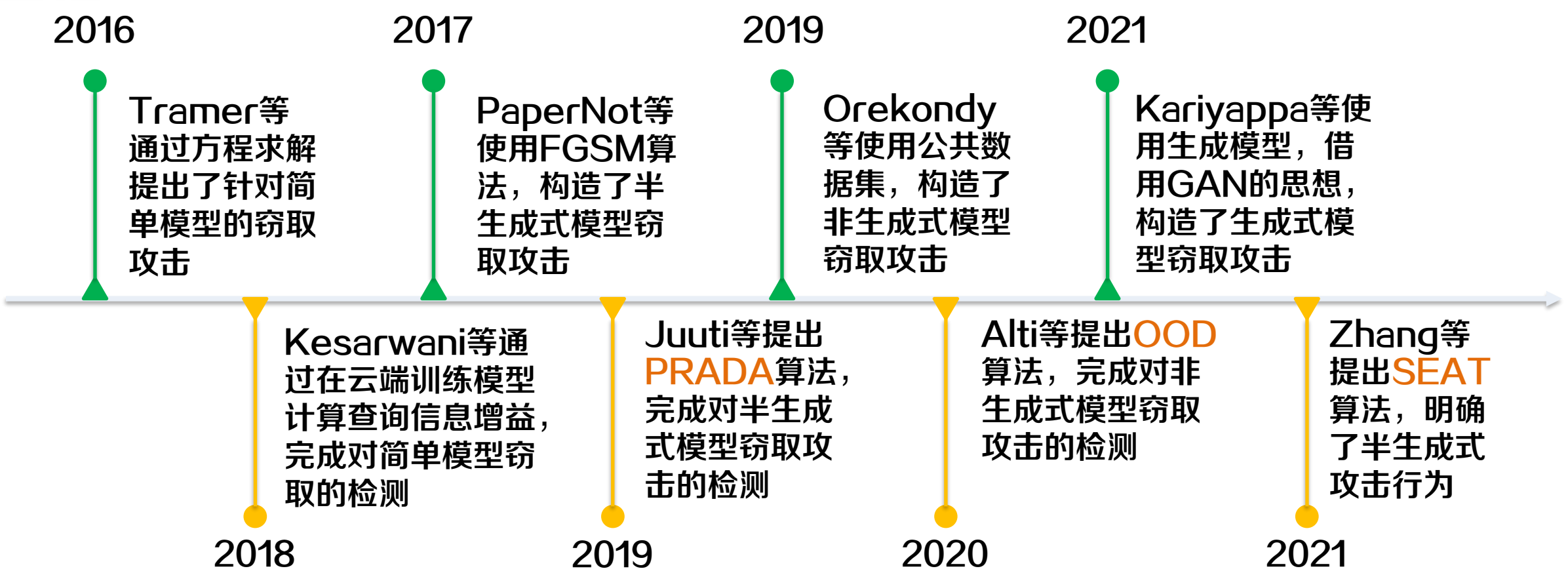
- 做过的思考
  - 从样本个体分析
    - 被验证单个样本无法证明模型窃取的存在
    - Dziedzic于2022年以Proof of Work思想巧妙绕过，使攻击者的成本显著提升
  - 类似蜜罐的形式
    - 对模型输出预测标签加上特征，检测新的查询样本中是否存在特征
    - Hard Label攻击打破了这一想法
    - 已经被应用于模型水印和模型窃取防御领域



Proof of Work思想简单概述图

模型安全领域许多想法是相通的！
问渠那得清如许,为有源头活水来。

总结

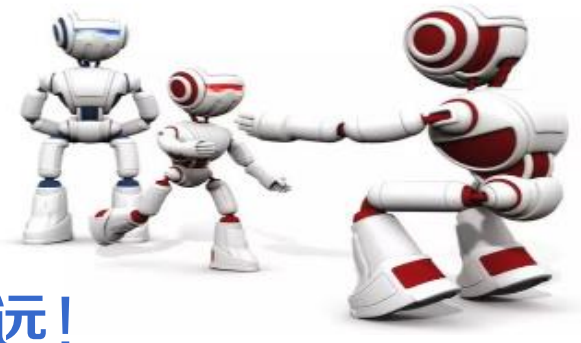


技术正是在攻与防的对抗中不断发展！

总结

- 模型窃取检测的重要性
  - 能够及时的发现并阻止攻击者，与防御技术联合保护模型安全
  - 减小模型信息或隐私信息泄露的风险
  - 维护模型知识产权，保障人工智能领域健康发展
  - 阻止基于替代模型的白盒攻击，维护模型使用的安全性
  - 不会对正常用户的预测结果产生干扰（相对于模型窃取防御）

- 模型窃取检测的劣势
  - 仅在样本查询阶段起到作用，替代模型一旦生成，便发挥不了作用
  - 无法检测多用户的联合攻击
  - 容易对正常用户产生误判，增大正常用户的查询成本
  - 相对于层出不穷的攻击方法，检测方法的更新速度较慢



保护模型安全任重而道远！

总结

- 问题回答
  - 为什么关于模型窃取检测的文献仅仅到2021年，之后的文献没有了吗?
    - 模型窃取检测是一项看起来容易，实际做起来难度大的任务
    - 检测器获取的信息少，仅有查询样本这一信息
    - 多数研究方向转为模型窃取防御、模型水印和攻击阻塞等
    - 有一篇文献（2022年B类会议），但算法模糊，算法思想重复

  - 基础概念部分提到了三种攻击方式：半生成、生成和非生成式攻击，为什么检测方法仅对半生成和非生成攻击进行分析?
    - 时间上，有效的生成式攻击是在2021年提出的
    - 效果上，生成式攻击模拟真实场景的攻击行为（攻击者无法拥有任何信息）
    - 攻击的样本特征包含于提到的检测算法的检测范围之内

参考文献

[1] JUUTI M, SZYLLER S, MARCHAL S, et al. PRADA: protecting against DNN model stealing attacks[C]. Proceedings of the 2019 IEEE European Symposium on Security and Privacy (EuroS&P), Stockholm, Sweden, F. 2019: 51-63.

[2] ATLI B G, SZYLLER S, JUUTI M, et al. Extraction of complex dnn models: Real threat or boogeyman?[C]. Proceedings of the International Workshop on Engineering Dependable and Secure Machine Learning Systems, New York City, NY, USA, F. 2020: 135-147.

[3] ZHANG Z, CHEN Y, WAGNER D. Seat: Similarity encoder by adversarial training for detecting model extraction attack queries[C]. Proceedings of the Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security, Virtual Event Republic of Korea, F. 2021: 28-41.

[4] PAPERNOT N, MCDANIEL P, GOODFELLOW I, et al. Practical black-box attacks against machine learning[C]. Proceedings of the Proceedings of the 2017 ACM on Asia conference on computer and communications security, Abu Dhabi United Arab Emirates, F. 2017: 71-81.

知人者智，自知者明。胜人者有力，自胜者强。知足者富。强行者有志。不失其所者久。死而不亡者，寿。

谢谢！