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

基于输入输出扰动的模型窃取防御方法

硕士研究生 张辰龙

2024年07月14日

问题回溯

- 总结反思
  - 增加一些基础性知识，保证讲解的完整性
  - 部分地方语速偏快，可以适当进行调整，把控整体节奏
  - 语音语调较为平缓，学习找到讲解的感觉，抓住听众的注意力

- 相关内容
  - 2023.09.17 张辰龙 《深度神经网络模型窃取防御方法》
  - 2023.03.12 邢凤桐 《深度神经网络模型水印保护方法》
  - 2023.03.05 张辰龙 《深度神经网络模型窃取检测》
  - 2021.01.03 王　琛 《深度神经网络对抗样本防御方法》

内容提要

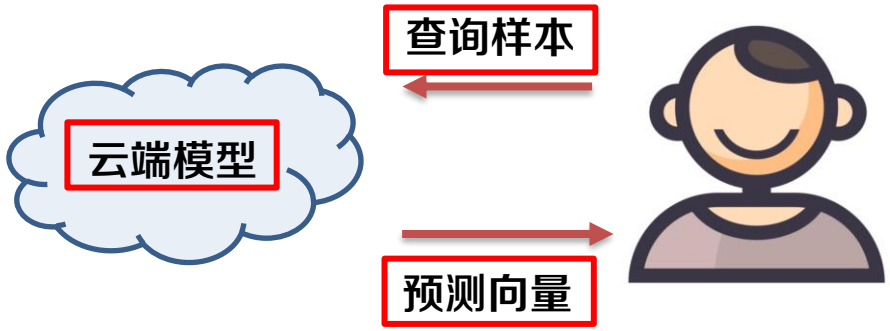
- 预期收获
- 题目内涵解析
- 研究背景与意义
- 研究历史与现状
- 知识基础
- 算法原理
  - APGP
  - APMSA
- 特点总结与工作展望
- 参考文献

预期收获

- 预期收获
  - 了解深度神经网络模型防御整体框架
  - 理解深度神经网络模型窃取防御的算法原理及其理论问题
  - 通过学习最优化问题的构建、求解、优化思路，为其他研究方向提供灵感
  - 了解深度神经网络模型窃取防御的重要意义

目标内涵

- 研究目标
  - 通过对模型的输入输出进行修改，减小查询样本引发的信息泄露
  - 降低攻击者窃取所得替代模型的预测准确率
- 题目内涵解析
  - 模型窃取防御的三类方式
    - 扰动输入、扰动输出、扰动模型决策边界
  - 输入扰动
    - 将输入样本经过修改后再经过模型预测
    - 通过更大的噪音覆盖掉攻击者精心设计的微小扰动
  - 输出扰动
    - 将输出预测向量进行修改，提供错误的或具有误导性的预测向量



背景意义

- 研究背景
  - 部署在云端的模型（黑盒），向用户提供查询接口
  - 攻击者：构造样本→利用查询接口获得预测向量→利用样本、向量训练本地模型
- 研究意义
  - 保护权益：模型训练需要很大的代价，模型拥有者通过向每次查询收费来收回成本，窃取模型后可以绕过付费查询
  - 保护隐私：窃取后的模型与原模型具有相似的决策边界，可以为对抗样本攻击、成员推理攻击、模型反演攻击等提供跳板
  - 保护知识产权：保证深度神经网络模型的知识产权，促进数据共享

研究历史与现状 模型窃取攻防

2016 Tramer 等人提出针对简易模型结构的模型窃取，可以通过方程求解实现窃取

2017 Papernot等人利用迭代方式生成逼近目标模型分类边界的样本，实现窃取

2018 Lee等人提出对输出预测向量添加高斯噪声可以有效防御模型窃取

2019 Orekondy等人指出可以利用公共数据集，基于随机和贪婪策略完成模型窃取

2019 Taesung等人构造sigmoid映射，对输出预测向量进行普遍的向量变换

2020 Orekondy 等人利用替代模型训练机制，通过输出向量误导模型梯度更新方向

2020 sanjay等人首次提出分类扰动的概念，结合OOD进一步提升了防御效果

2021 DFME: 结合生成对抗网络生成随机数据，同步训练替代模型和生成器

2021 MAZE:通过将DFME的损失函数更新为一阶范数，同样实现了无数据模型窃取

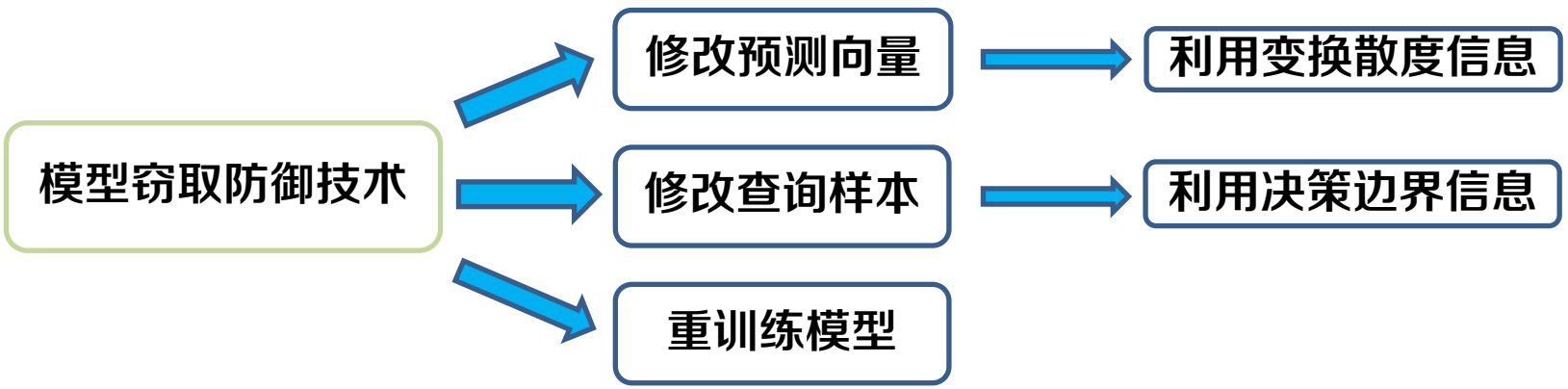
2019 Wang等人提出信息净化框架，同步净化输入样本和输出样本，减少信息泄露

2022 Dzie: 根据信息泄露评估指标要求用户提供wof工作量证明，以提高攻击成本

2022 GRAD2: 通过采用常数向量约束，误导替代模型的梯度更新方向偏移

2023 APGP:训练一个扰动模型，使得将每一个输出的预测向量修改的KL散度最大

2024 Liang等人通过辅助数据训练受害模型，以模糊受害模型的决策边界



基本概念

- 模型窃取
  - 攻击者构造无标签的“攻击者数据集”，利用预测模型的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
- 模型窃取防御（扰动输入）
  - 对攻击者的查询样本进行样本变换，将变换后的样本输入模型
  - 对所有查询样本均执行相同的变换
- 模型窃取防御（扰动输出）
  - 直接修改每个查询样本的预测向量，向攻击者提供错误的信息
- 约束条件
  - 攻击者无条件使用模型的所有输出结构，不考虑硬标签模型窃取
  - 防御要兼顾对于正常样本的预测功能

对样本的变换最终仍是影响的模型的输出！

APGP: Accuracy-Preserving Generative Perturbation for Defending Against Model Cloning Attacks

TIPO APGP

T	目标	通过扰动预测向量减少信息泄露
I	输入	1 组模型预测向量
P	处理	1.训练扰动模型，使得每一个预测向量修改变化的KL散度最大，但各类别次序不变 2.构造最优化问题，使用拉格朗日乘子法求解 3.设定控制因子，可自主控制最优化方程保证预测向量的top-n不变
O	输出	1组扰动后的预测向量

P	问题	现有防御方法计算开销大、存在严重的效用权衡、存在信息泄露风险
C	条件	攻击者可以利用一切受害模型返回的信息
D	难点	如何通过修改预测向量尽可能向攻击者提供错误信息
L	水平	IEEE ICASSP（2023 顶会）

算法原理 APGP

- 核心思想
  - 对模型输出的预测向量进行重映射，使映射前后“距离”最大，但保证置信度顺序不变
- 算法步骤
  - 定义差异：采用KL散度等距离计算指标定义映射前后的向量差异
  - 确定约束：以置信值排序作为约束条件，构造不等式约束
  - 最优化问题求解：根据优化目标和约束条件采用拉格朗日乘子法求解，并设计增广方式，增加惩罚项以确保排序正确

明确核心思想，以正向思维优化求解

预测向量的“距离”：用于衡量两个预测向量之间的差异性，常见的衡量方式有：KL散度、范数距离、余弦距离等。常用字母d表示，d越小，距离越小，表示二者越相近。

拉格朗日乘子法：是一种寻找变量受一个或多个条件所限制的多元函数的极值的方法。将一个有n 个变量与k 个约束条件的最优化问题转换为一个有n + k个变量的方程组的极值问题，其变量不受任何约束。

算法原理 APGP

- 变量定义
  - 训练数据：$D = \{(x^n, y^n)\}_{n=1}^N$，其中$N$为训练集容量
  - 预测向量：$p^n = f_T(x^n; \omega)$，其中$f_T(x; \omega)$为预测模型
- 优化目标
  - 获得对预测向量的扰动模型G_θ，实际是一个映射函数
  - $\max_\theta \frac{1}{N} \sum_{n=1}^N d(\sigma(G_\theta(p^n)), \sigma(p^n))$
  - $s.t. argsort(G_\theta(p^n)) = argsort(p^n)$
  - 但是！$argsort$对数组进行排序，离散操作，不可微
  - $s.t. G_\theta(p^n)_{r_i^n} > G_\theta(p^n)_{r_{i-1}^n}, \forall i \in \{2, \dots, M\}$
  - 其中$r^n = argsort(p^n)$

构造原始拉格朗日形式，化离散约束为不等式约束

$argsort$函数：返回预测向量从小到大索引的排序。

例如

$argsort([0.7, 0.1, 0.2]) = [2, 3, 1]$

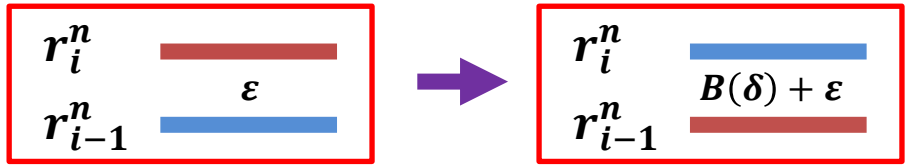
思考：为什么softmax函数σ，要位于$G_\theta()$的外侧？

答：可以减少一个约束条件，不需要对G_θ的输出做归一化约束。

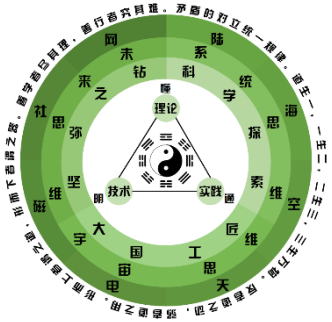
算法原理 APGP

- 优化目标
  - $\max_{\theta} \frac{1}{N} \sum_{n=1}^{N} d(\sigma(G_\theta(p^n)), \sigma(p^n))$
  - $s.t.\, G_\theta(p^n)_{r_i^n} > G_\theta(p^n)_{r_{i-1}^n}, \forall i \in \{2, \dots, M\}$

ELBO思想：提高下确界，强制提高模型的类间差异



- 构造求解
  - $\max_{\theta} \frac{1}{NM} \sum_{n=1}^{N} [\sum_{i=1}^{M} d_i \left(\sigma(G_\theta(p^n)), \sigma(p^n)\right) - \lambda \sum_{i=2}^{M} max\left(0, G_\theta(p^n)_{r_{i-1}^n} - G_\theta(p^n)_{r_i^n}\right)]$
  - 原始的拉格朗日乘子法在面对冗余约束等问题时会出现难以收敛的问题
  - 引入可学习参数δ
  - $\max_{\theta,\delta} \frac{1}{NM} \sum_{n=1}^{N} \sum_{i=1}^{M} d_i(\sigma(q^n), \sigma(p^n))$
  - $s.t.\, q_{r_1^n}^n = G_\theta(p^n)_{r_1^n} \;\; and \;\; q_{r_i^n}^n = max(G_\theta(p^n)_{r_i^n}, G_\theta(p^n)_{r_{i-1}^n} + B(\delta))$
  - 其中$B(x) = \exp(x)$，取其非负性



算法原理 APGP

- 优化目标
  - $\max_{\theta,\delta} \frac{1}{NM} \sum_{n=1}^{N} \sum_{i=1}^{M} d_i\left(\sigma(q^n), \sigma(p^n)\right)$
  - $s.t. q^n_{r_1^n} = G_\theta(p^n)_{r_1^n} \quad and \quad q^n_{r_i^n} = max(G_\theta(p^n)_{r_i^n}, G_\theta(p^n)_{r_{i-1}^n} + B(\delta))$

从数学角度构造最优化问题能够展现出较好的可扩展性和灵活性

- 新的需求
  - 在符合实际使用要求的前提下，仅保证top-K的预测向量排序不变
  - $s.t. q^n_{r_i^n} = G_\theta(p^n)_{r_i^n}, \forall i \in \{1, \dots, M-K\}$
  - $s.t. q^n_{r_i^n} = max\left(G_\theta(p^n)_{r_i^n}, \max_{j<i}\left(G_\theta(p^n)_{r_j^n}\right) + B(\delta)\right), \forall i \in \{M-K+1, \dots, M\}$

少量排序可以引入更强的约束

强调理论引领
发挥技术优势
实现安全保障

实验设计与实验结果 APGP

- 数据源：4个经典的图像数据集
  - CIFAR-10、CIFAR-100、CUB-200、Caltech-256
- 评价指标
  - #Stealer Acc：攻击者窃取所得模型的测试准确率
  - #Defender Acc：防御模型对正常样本的测试准确率

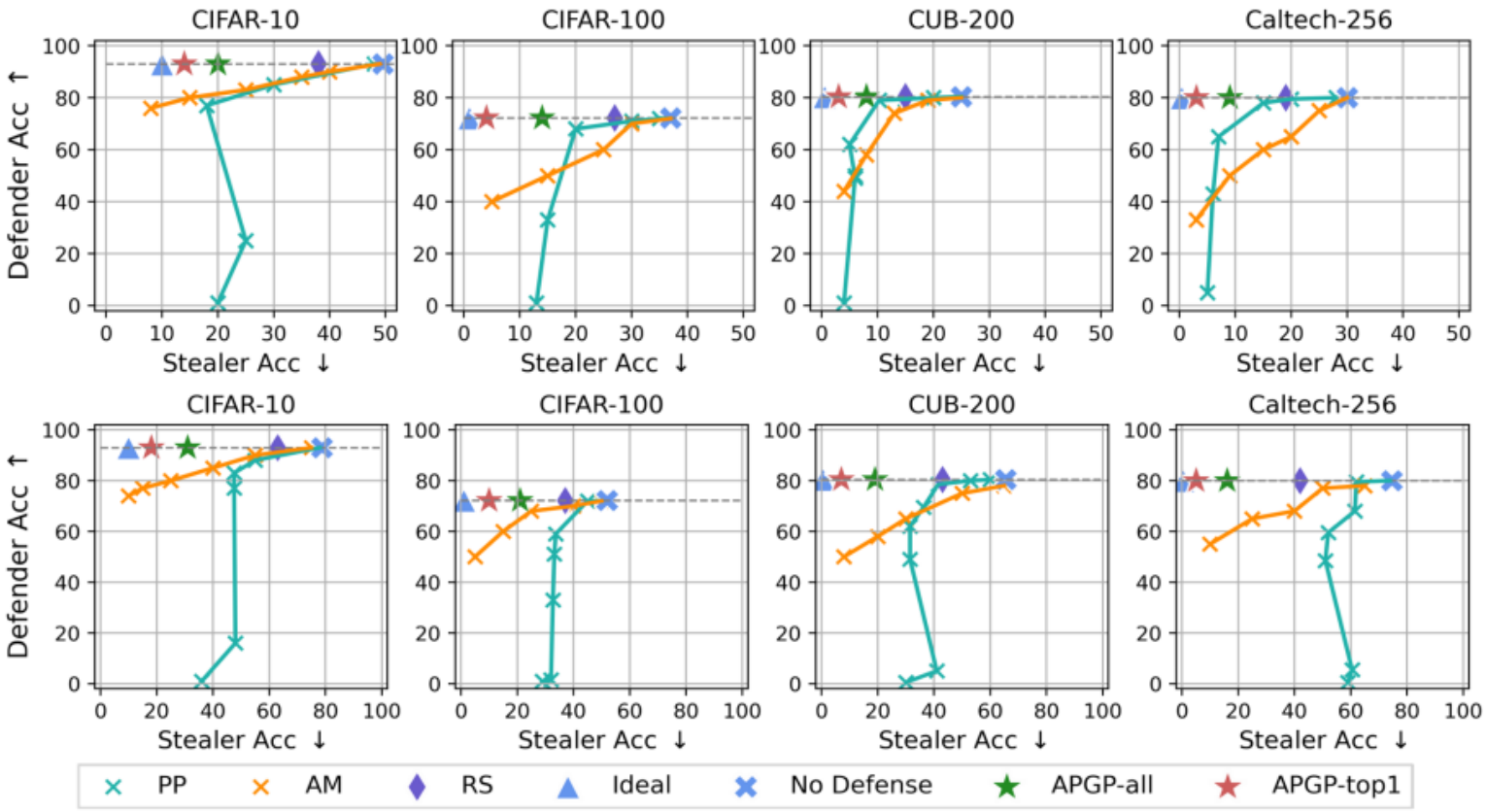
防御领域通常会引入No Defense方法，即在没有任何防御措施的情况下，Stealer Acc和Defender Acc分别为多少，加上防御方法后主要查看的是这两个指标的下降程度

Dataset	CIFAR-10			CIFAR-100		
Target Model	ResNet18	ResNet50	ResNeXt29	ResNet18	ResNet50	ResNeXt29
No Defense	95.13	95.48	95.76	77.44	78.12	81.85
NT [7]	94.56 (−0.57)	94.70 (−0.78)	94.62 (−1.14)	77.42 (−0.02)	77.14 (−0.98)	80.26 (−1.59)
APGP	95.13 (−0.00)	95.48 (−0.00)	95.76 (−0.00)	77.44 (−0.00)	78.12 (−0.00)	81.85 (−0.00)

受害模型预测的TOP-1准确率

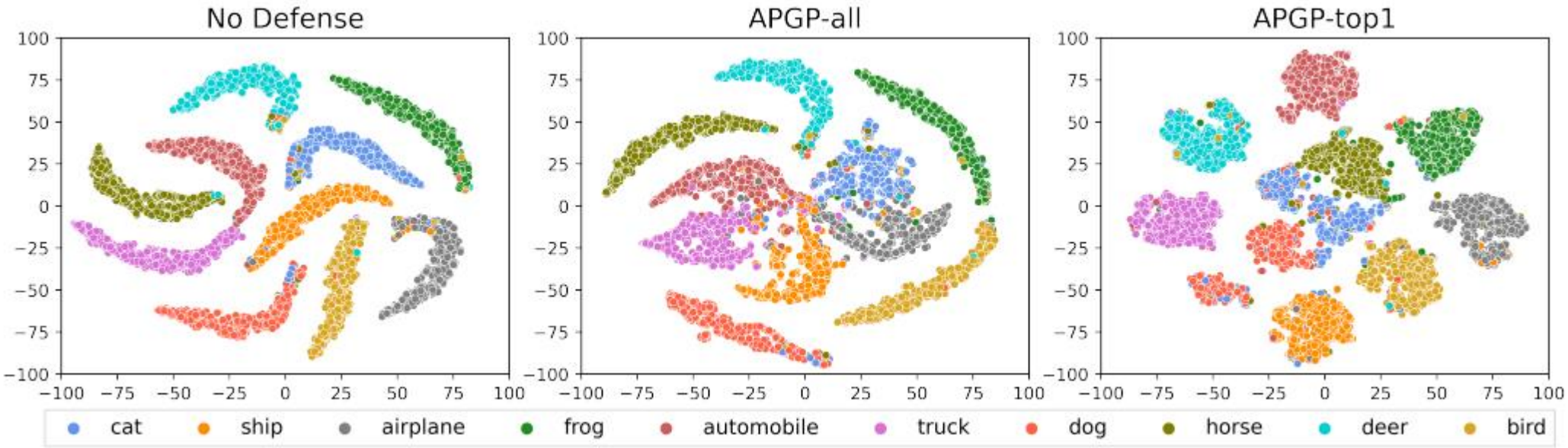
实验结果 APGP

- 效果评估
  - 对4种数据集和2种攻击方法，APGP均能实现较低的Stealer Acc和较高的Defender Acc
  - APGP-top1的效果好于APGP-all（符合预期）
- 其它分析
  - 直接的映射：速度快
  - 对于Defender Acc的定义
    - [0.9, 0.06, 0.04]
    - [0.5, 0.26, 0.24]
  - 可用的定义方式
    - 余弦相似度
    - 范数距离



实验结果 APGP

- 效果评估
  - APGP算法改变了从预测向量体现出的数据分布，迫使攻击者训练的替代模型拟合不到正确的决策边界
  - t-Distributed Stochastic Neighbor Embedding (t-SNE)绘图方法，在二维平面上直观展示出数据分布情况，python中有直接的函数库调用



APMSA: Adversarial Perturbation Against Model Stealing Attacks

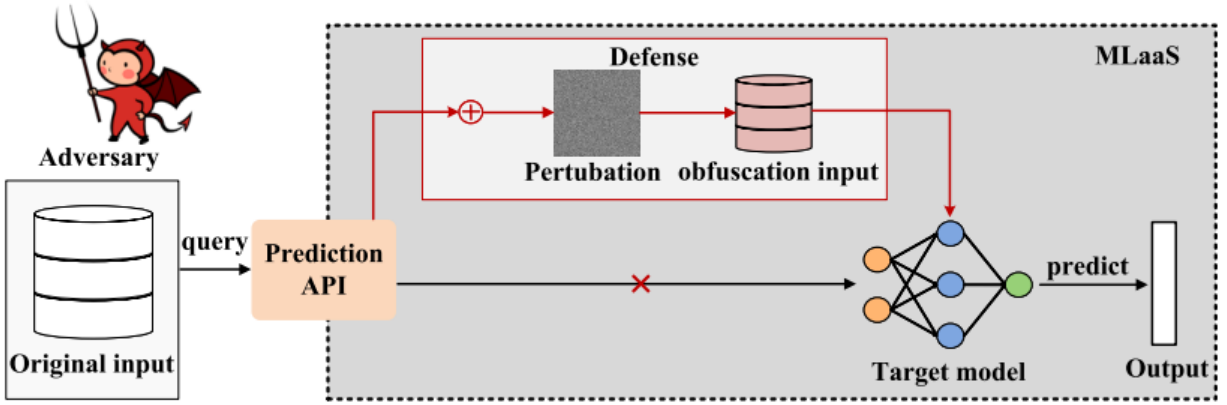
TIPO APMSA

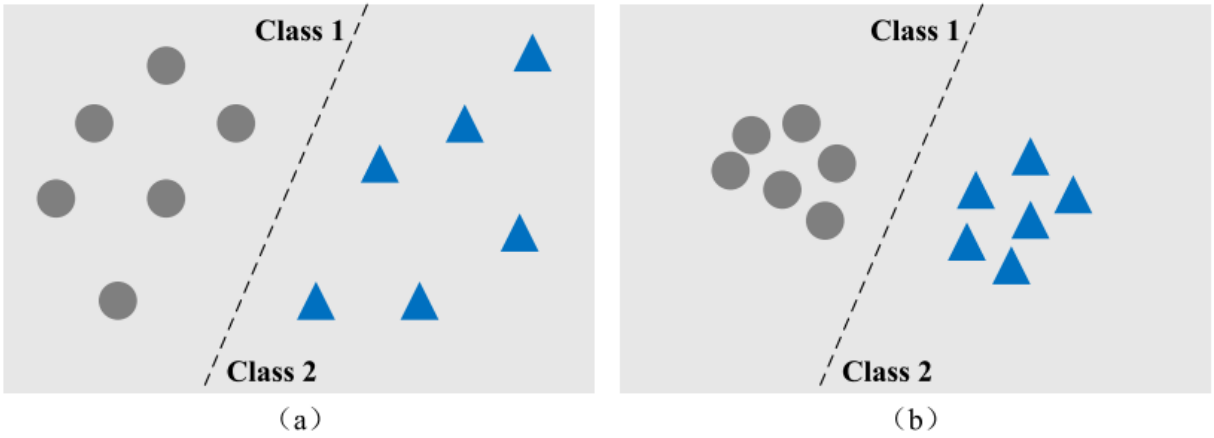
T	目标	通过扰动模型输入样本，间接影响输出，减少信息泄露
I	输入	1 组输入图片样本
P	处理	1.对每个输入样本添加对抗噪声 2.循环添加，确保最后样本处于刚好改变硬标签的前一阶段 3.返回修改后样本对应的预测向量
O	输出	1组图片样本对应的预测向量

P	问题	现有方法难以保障模型的内部信息
C	条件	攻击者可以利用一切受害模型返回的信息
D	难点	如何修改查询样本以覆盖掉攻击者精心设计的扰动
L	水平	IEEE TIFS (2023 CCF A)

算法原理 APMSA

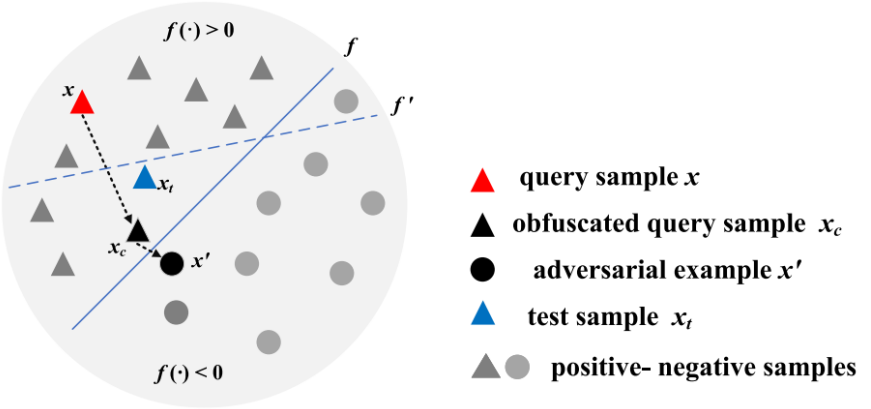
- 核心思想
  - 通过改变输入样本，间接影响预测向量，减少信息泄露
- 算法步骤
  - 构造对抗噪声：以白盒对抗样本构造方式，对每一个输入样本构造其对抗性示例
  - 样本确定：取刚好改变样本硬标签的前一个样本为真正的输入样本
  - 返回结果：将改变后的样本的预测向量返回给查询用户

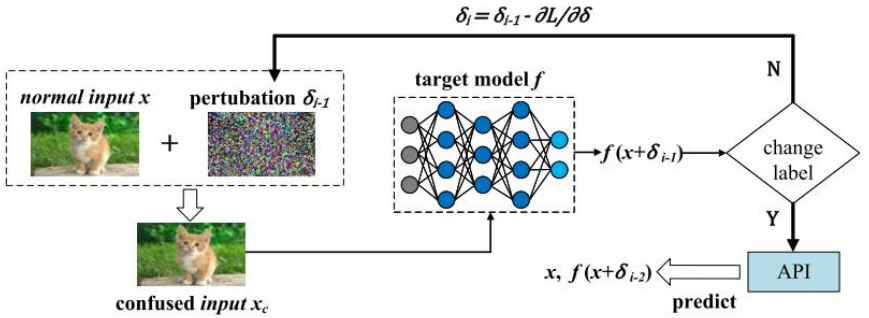




算法原理 APMSA

- 变量定义
  - 输入样本：x　　预测模型：f_v
  - 预测向量：$y = f_v(x)$ 其中y的标签为$argmax(y)$
- 算法步骤（以FGSM为例）
  - 确定标签：定义预测向量的标签M为该样本的标签
  - 计算损失：采用交叉熵损失函数
    - $l = CrossEntropyLoss(y, M)$
  - 修改样本：计算l对x的偏导，循环修改x，使l增大，迫使硬标签发生改变
  - 样本选取：选取硬标签改变前的样本作为最终输入模型的样本





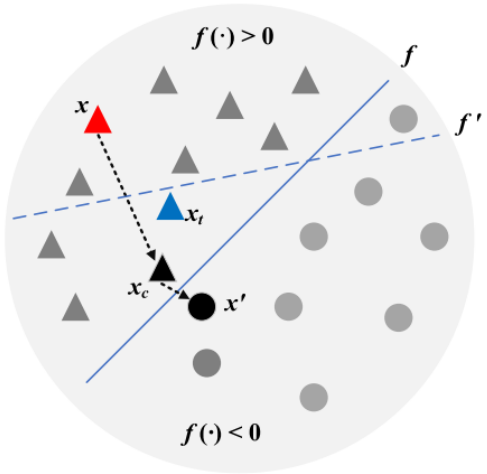
实验设计与实验结果 APMSA

- 数据源：2个经典的图像数据集
  - CIFAR-10、GTSRP
- 攻击方法：基于种子样本生成的攻击模式
- 评价指标
  - #Accuracy：攻击者窃取所得模型的测试准确率

为了适应算法性质，在实验选取上，该方法仅仅采用了种子样本生成这一种攻击方法，缺乏普遍性

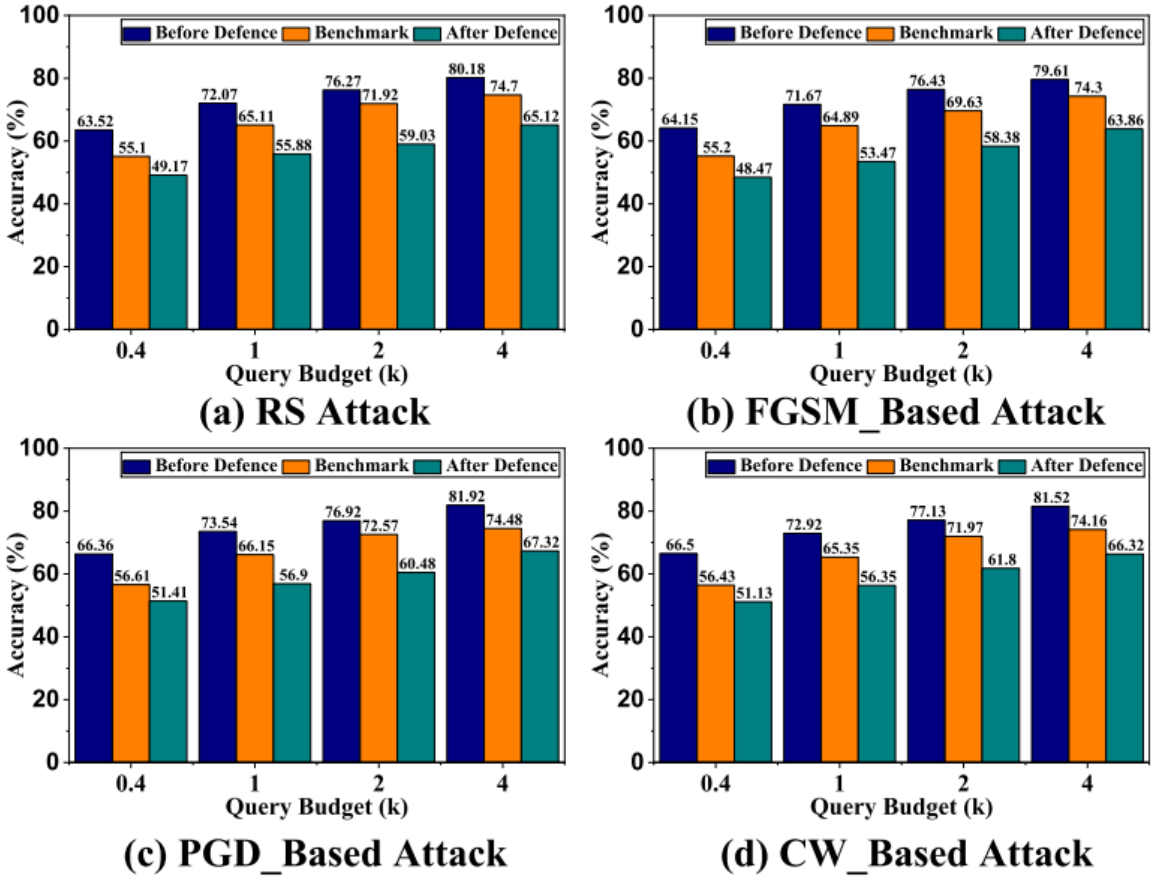
从多层向量机的角度考虑，单个点可以解释决策边界的偏移，而点的数量增多后，决策边界仍有可能回归到正常情况

Queries	5k	6k	7k	8k
RS	82.54 (89.77×)	83.36 (90.67×)	83.32(90.62×)	84.05(91.42×)
FGSM_Based	80.57(87.63×)	79.70(86.69×)	81.07(88.18×)	81.23(88.35×)
PGD_Based	82.31(89.53×)	83.23(90.53×)	83.66(90.99×)	84.66(92.08×)
CW_Based	79.82(86.82×)	80.37(87.42×)	80.56(87.62×)	81.40(88.54×)



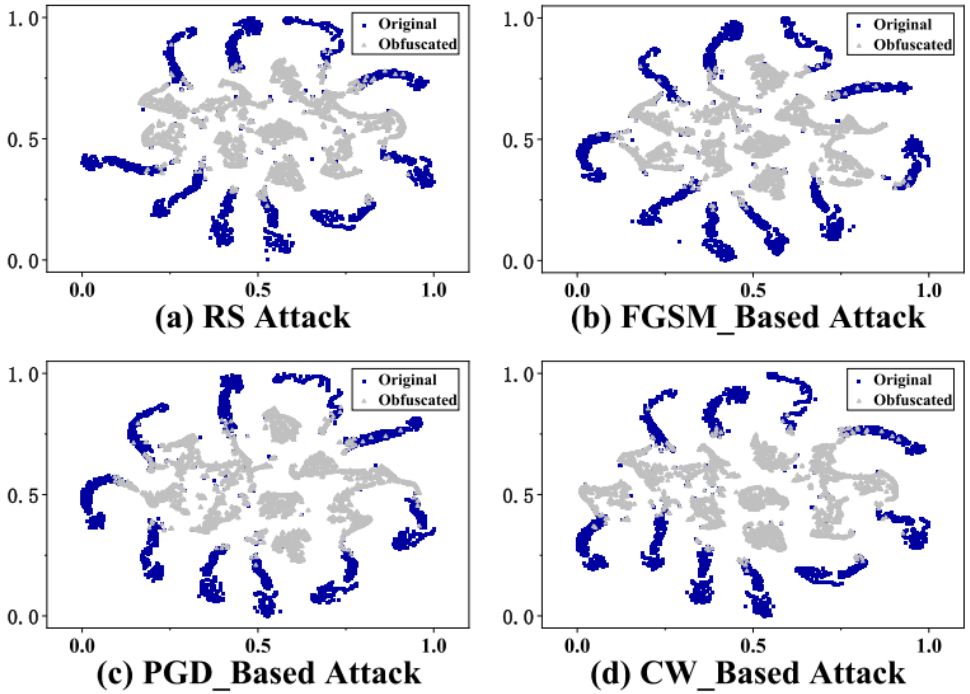
实验结果 APMSA

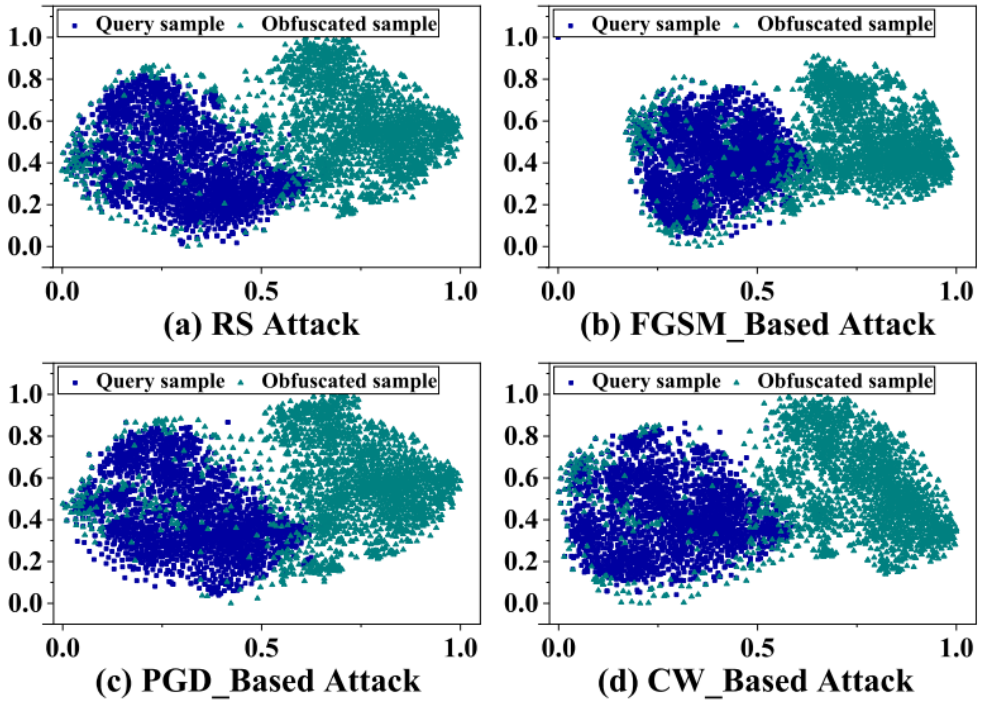
- 效果评估
  - 设计Benchmark：攻击者仅通过硬标签进行模型窃取（解释了算法的条件）
  - APMSA算法能够有效降低模型窃取所得替代模型的准确率
- 其它分析
  - 有新意：是改变输入进行模型窃取防御的新的尝试，改变了传统的噪声添加模式
  - 速度慢：对于每个样本平均需要进行数十次的迭代
  - 缺少实验分析：在文章中反复提及是对模型内部信息的保护，但实验论证不够充分



实验结果 APMSA

- 效果评估
  - 修改后的输入分布于修改前的分布具有明显的差异性，修改后的输入分布更加集中
  - 从单个样本的角度看，修改先后的样本具有明显的距离差异





特点总结与工作展望

- 特点总结

算法	APGP	APMSA
优势	1.以映射的方式修改预测向量，速度快 2.从数学角度解释了模型训练的原理，可扩展性和灵活性强	1.首次给出了以样本变换进行防御的形象解释，不再使用传统的噪声覆盖 2.能够很好的保护模型的内部信息
劣势	1.对预测向量的修改程度大，定义的评价指标有“自娱自乐”的嫌疑	1.对于样本变换迭代轮次多，耗时非常长 2.将样本靠近决策边界后可能起到反向效果

- 未来发展
  - 直接修改预测向量：和APGP类似，未来面向防御的方法更多的会采用直接修改预测向量的方式，因为无论是修改查询样本还是修改模型，最终的结果都是体现在预测向量的改动上
  - 分类防御：在普遍性防御的基础上进行适当的分类，以更好的平衡模型的防御能力和实际使用效果

参考文献

[1] Zhang J, Peng S, Gao Y, et al. APMSA: Adversarial perturbation against model stealing attacks[J]. IEEE Transactions on Information Forensics and Security, 2023, 18: 1667-1679.

[2] Cheng A, Cheng J. APGP: Accuracy-Preserving Generative Perturbation for Defending Against Model Cloning Attacks[C], ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023: 1-5.

[3] Jiang W, Li H, Xu G, et al. A comprehensive defense framework against model extraction attacks[J]. IEEE Transactions on Dependable and Secure Computing, 2023, 21(2): 685-700.

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！