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

缺乏先验知识条件下的模型窃取方法

硕士研究生 丁杨

2021年04月11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. 了解机器学习的云服务现状
  - 2. 了解模型窃取方法的发展历史
  - 3. 理解缺乏先验知识的模型窃取方法的技术原理
  - 4. 了解模型窃取在网络安全领域中的应用

背景简介

- 机器学习的发展
  - 在从图像分类到语音识别等各个领域都取得了最先进的性能
  - 使用大量敏感的训练数据来训练模型，并且通常需要花费大量计算资源
- 许多云提供商已经启动了机器学习即服务（MLaaS）

背景简介

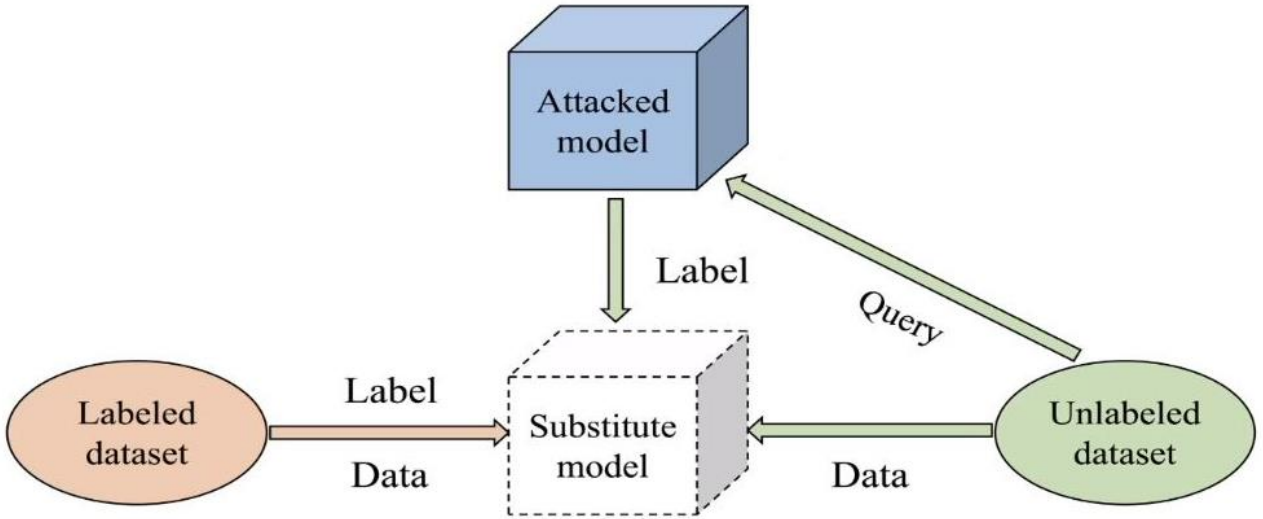
- 模型窃取提出的目的
  - 免费使用模型：模型训练者将模型托管在云上，提供访问模型的API，通过对其他用户每次调用API的方式来收费，恶意的用户将企图窃取这个模型免费使用
  - 破坏训练数据的隐私性：利用对模型的多次访问可以推断出训练数据信息
  - 绕过安全检测：在很多场景中，机器学习模型用于检测恶意行为，例如垃圾邮件过滤，恶意软件检测，网络异常检测。攻击者在提取到目标模型后，可以构造相应的对抗样本，以绕过安全检测

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、已知与训练数据有相似属性的数据集等条件下，实现了对DNN的窃取

背景简介

- 研究现状
  - 现有的大多数模型窃取攻击都需要有关目标DNN的训练数据或辅助数据的知识。实际上这些数据并不总是可访问的，由于数据保护意识的提升，很难获取如健康数据、生物特征等类型的数据

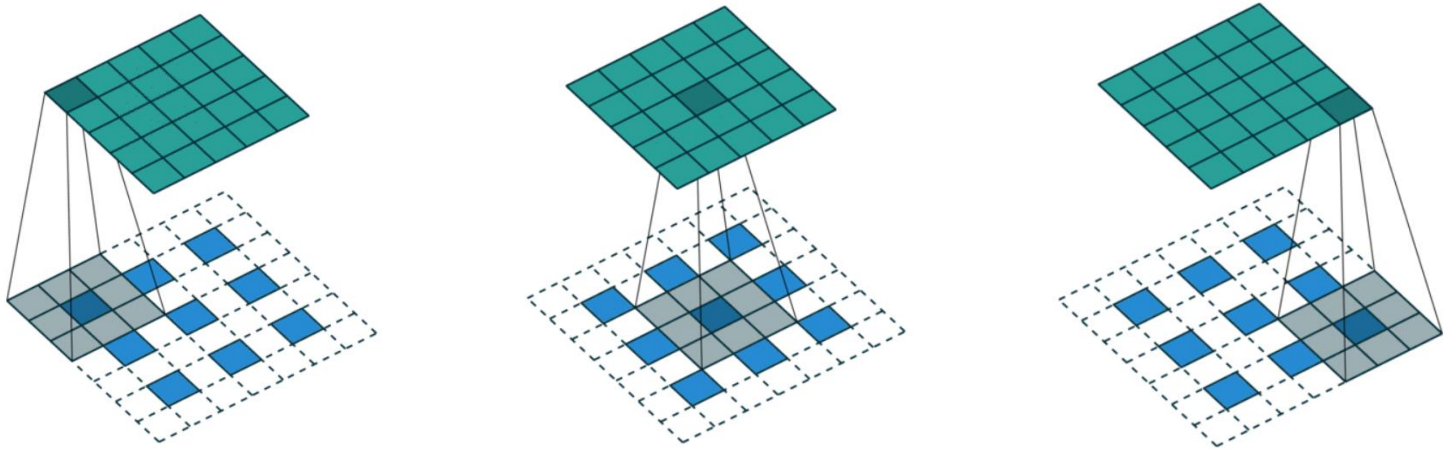


基本概念

基本概念

- 基本概念
  - 模型窃取
    - 一种攻击者通过循环发送数据并查看对应的响应结果，来推测机器学习模型的参数或功能，从而复制出一个功能相似甚至完全相同的机器学习模型的攻击方法
  - 对抗样本
  - 无目标攻击（Non-targeted Attack）
    - 只需让模型辨认的结果是错误即可，不关心会被识别成什么类别，只关心不让输入被识别成正确的类别
  - 目标攻击（Targeted Attack）
    - 不只要让模型获得错误的分类结果，而且要让其被分类成指定的类

- 基本概念
  - 图像上采样(Upsampling)
    - 一种让图像变成更高分辨率的技术，在原有图像像素的基础上在像素点之间采用合适的插值算法（双线性插值，反卷积，反池化）插入新的元素
  - 反卷积：先按照一定的比例通过补0来扩大输入图像的尺寸，接着旋转卷积核，再进行正向卷积



算法原理

算法原理

T	在无先验知识的条件下进行模型窃取
I	可查询的目标模型

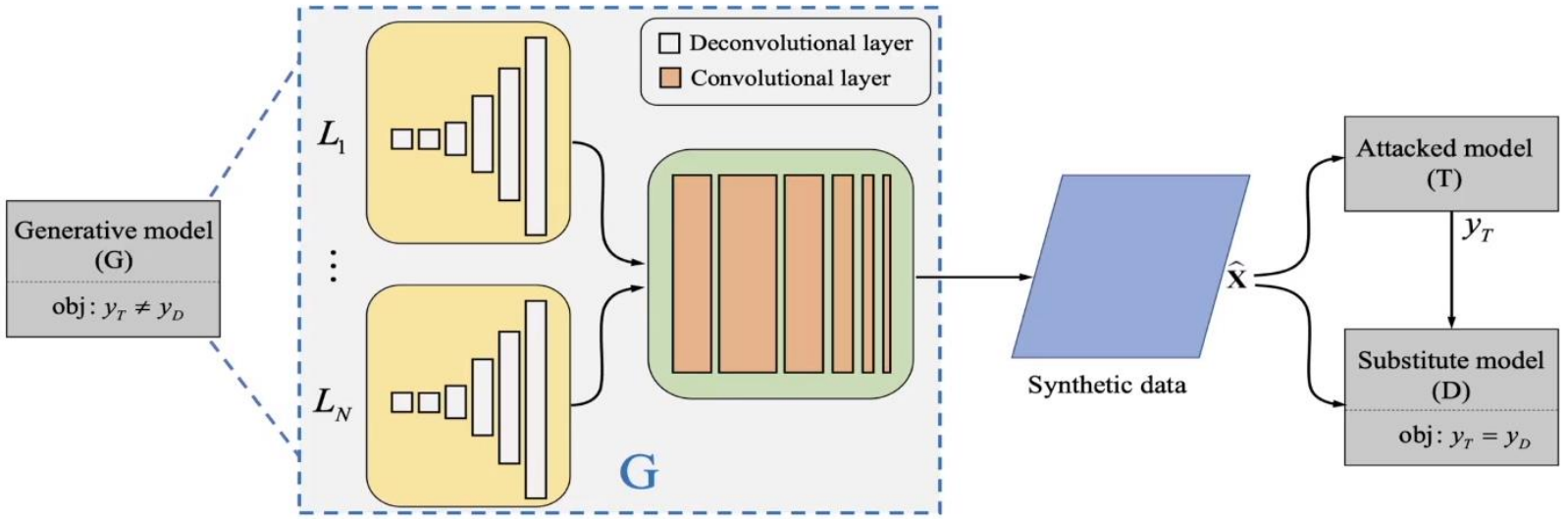
P	For {
	1. 生成器产生随机样本数据
	2. 通过目标模型为样本数据打标签生成训练集
	3. 通过训练集训练替代模型
	4. 优化生成器
	}

O	与目标模型功能相近的替代模型

P	无法获得原始的训练数据与内部参数
C	攻击者可以自由访问被攻击模型获得其输出
D	生成与原训练数据分布相近的训练数据
L	NDSS 2020

算法原理

- 算法流程图



the objective of G: generate samples $\widehat{\mathbf{X}} = G(\mathbf{X})$ and let $y_D(\widehat{\mathbf{X}}) \neq y_T(\widehat{\mathbf{X}})$

the objective of D: guarantee $y_D(\widehat{\mathbf{X}}) = y_T(\widehat{\mathbf{X}})$

算法原理

- 算法基本步骤
  - 使用生成模型G产生替代模型D的训练数据
  - 使用生成的数据探测被攻击模型T的输出T(x)
  - 由(x, T(x))训练替代模型D
  - 不断的迭代，直至达到收敛条件

Algorithm 1 Mini-batch stochastic gradient descent training of the proposed method DaST.

acc denotes the accuracy of D. *att* denotes the attack success rate for the attacks generated by D.

1 : **While** iteration $< \delta$ **or** acc, att do not increace

2 : Generate m examples$\{\widehat{\mathbf{X}}^{(1)}, \ldots, \widehat{\mathbf{X}}^{(m)}\}$by G.

3 : Update the substitute model :

4 : $\mathcal{L}_D = d(T(\widehat{\mathbf{X}}), D(\widehat{\mathbf{X}}))$.

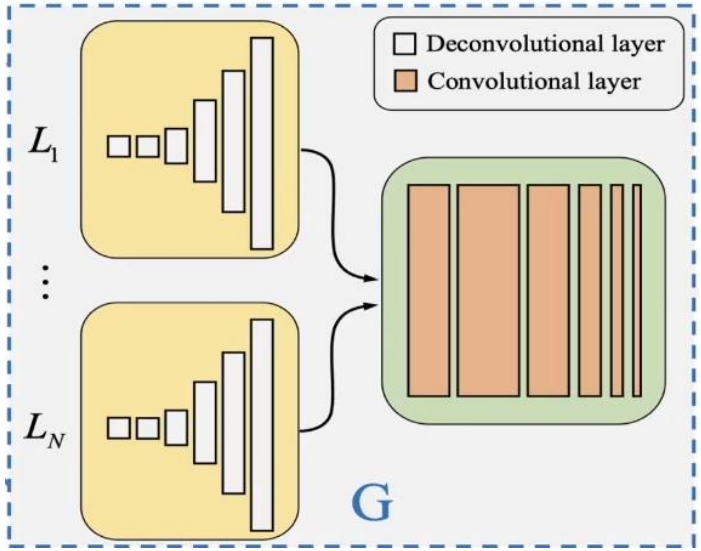
5 : Update the generative model :

6 : $\mathcal{L}_G = e^{-d(T,D)} + \alpha\mathcal{L}_C$.

7 : **end for**

算法原理

- 训练数据生成器G
  - G的目标是寻找能够使得T与D输出不一致的样本，挖掘T与D的不同之处。再用生成样本训练D，进而不断地逼近T，形成了一种对抗的形式
  - 在没有真实样本指导的情况下，G生成的数据可能会集中于某些分布区间之中
  - 设计了一个包含N个上采样反卷积生成网络
  - G从输入空间和可变标签值中随机采样噪声向量z
  - z被输入到第n个上采样反卷积网络和共享卷积网络，以产生数据X=G(z, n)



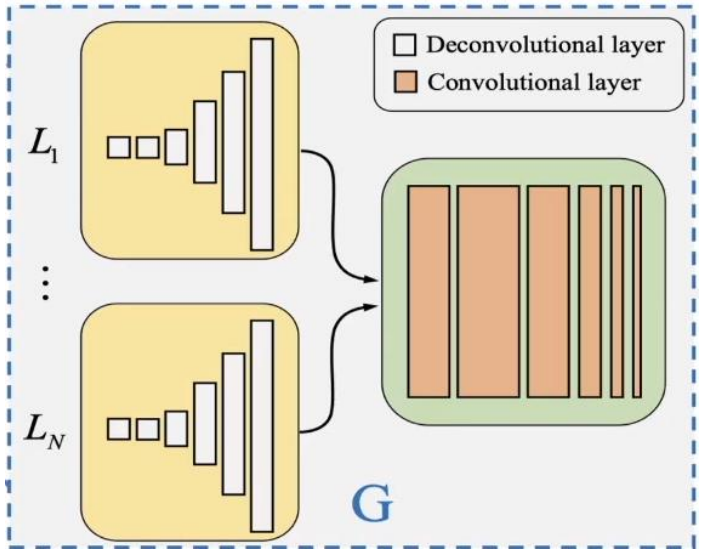
算法原理

- 训练数据生成器G
  - 建立模型G的附加标签控制损失公式Lc，Lc会约束T对生成图像的分类结果，使之能够与之前的控制信号n对应

$$\mathcal{L}_C = \mathrm{CE}(T(G(\mathbf{z}, n)), n)$$

  - 但这违背了对T无先验知识的条件，由于在训练过程中，D的输出将逐渐接近T的输出

$$\mathcal{L}_C = \mathrm{CE}(D(G(\mathbf{z}, n)), n)$$



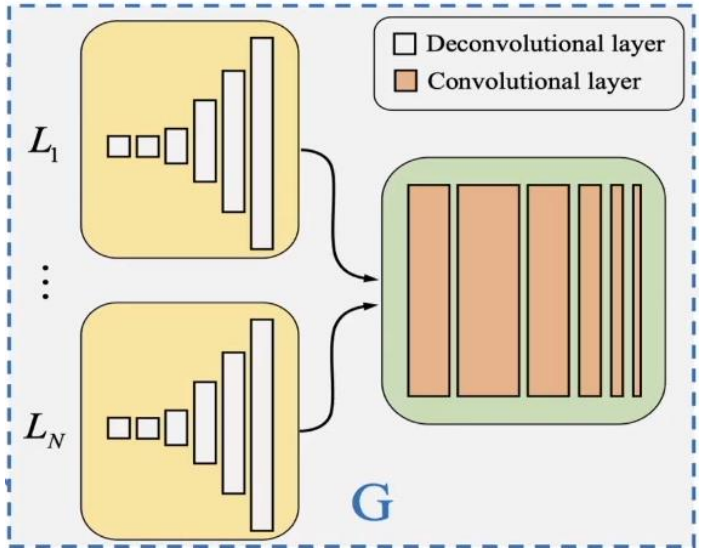
CE：交叉熵代价函数（Cross-entropy cost function）用来表征真实样本标签和预测概率之间差值函数

算法原理

- 训练数据生成器G
  - 为了避免T和D输出之间的差异越来越大，将G的损失函数更新为

$$\mathcal{L}_G = e^{-d(T,D)} + \alpha \mathcal{L}_C$$

  - 随着D的模仿能力增强，由T标注的合成样本的多样性会逐渐增强



算法结果

- 数据集
  - 实验在MNIST和CIFAR-10上进行评估，这两个数据集的测试集分别有10k个图像
- 实验条件
  - 实验场景中攻击者可以自由访问被攻击模型的输出
  - 被攻击模型的先验知识未知
- 实验设置
  - 被攻击型为训练好的VGG-16与中型网络
  - 替代模型设计了三种网络结构（3个卷积层的小型网络，4个卷积层的中型网络，5个卷积层的大型网络）
  - 评价方法：替代模型生成的对抗样本的攻击成功率

算法结果

- **MINST数据集上的实验**
  - **训练好的中型网络为T模型，使用三种网络结构作为D模型，将训练好的D模型分别用于Non-targeted与Targeted Attack**

Attack	Non-targeted		
	Pre-trained	DaST-P	DaST-L
FGSM	59.72 (5.40)	**69.76** (5.41)	35.74 (5.40)
BIM	85.70 (4.80)	**96.36** (4.81)	64.61 (4.82)
PGD	37.93 (3.98)	**53.99** (3.99)	23.22 (3.98)
C&W	23.34 (2.91)	**27.35** (2.74)	18.16 (2.75)
Attack	**Targeted**		
	Pre-trained	DaST-P	DaST-L
FGSM	12.10 (5.46)	**20.45** (4.49)	13.10 (5.46)
BIM	37.83 (4.90)	**57.22** (4.87)	29.18 (4.87)
PGD	28.95 (4.60)	**47.57** (4.63)	19.25 (4.63)
C&W	10.32 (2.57)	**23.80** (2.99)	12.31 (2.98)

Attack	Non-targeted		
	Small	Medium	Large
FGSM	62.61 (4.38)	56.21 (4.45)	**69.76** (5.41)
BIM	94.86 (4.85)	92.47 (4.84)	**96.36** (4.81)
PGD	45.31 (3.99)	43.62 (3.99)	**53.99** (3.99)
C&W	**30.61** (2.89)	24.34 (2.75)	23.80 (2.99)
Attack	**Targeted**		
	Small	Medium	Large
FGSM	19.92 (4.43))	20.45 (4.49)	**23.93** (5.45)
BIM	56.73 (4.89)	53.50 (4.84)	**57.22** (4.87)
PGD	39.42 (4.64)	40.76 (4.60)	**47.57** (4.63)
C&W	**24.86** (3.09)	16.25 (3.13)	23.80 (2.99)

  - **本方法生成的替代模型的攻击成功率明显高于其他预训练模型（在FGSM、BIM、PGB与C&W上分别为10.04%、10.66%、16.06%和4.01%）**

算法结果

- **CIFAR-10数据集上的实验**
  - **训练好的VGG-16网络为T模型，分别使用VGG-13、ResNet-18、 ResNet-50三种网络结构作为D模型**

Attack	Non-targeted		
	Pre-trained	DaST-P	DaST-L
FGSM	39.10 (1.54)	**39.63** (1.54)	22.65 (1.54)
BIM	59.18 (1.01)	**59.71** (1.18)	28.42 (1.19)
PGD	**35.40** (1.02)	29.10 (1.10)	17.80 (1.10)
C&W	9.76 (0.77)	**13.52** (0.74)	10.34 (0.74)
Attack	**Targeted**		
	Pre-trained	DaST-P	DaST-L
FGSM	**9.62** (1.54)	6.69 (1.54)	7.32 (1.54)
BIM	17.43 (1.00)	**20.22** (1.18)	15.26 (1.16)
PGD	10.46 (1.05)	**14.09** (1.12)	8.32 (1.10)
C&W	23.15 (2.05)	**26.53** (1.98)	19.78 (2.04)

Attack	**Non-targeted** (%)		
	VGG-13	ResNet-18	ResNet-50
FGSM	6.87 (1.54)	17.97 (1.54)	**39.63** (1.54)
BIM	**93.13** (1.18)	31.70 (1.54)	59.71 (1.18)
PGD	**56.14** (1.08)	10.04 (1.11)	29.10 (1.10)
C&W	**56.80** (1.64)	11.54 (1.64)	13.52 (0.74)
Attack	**Targeted** (%)		
	VGG-13	ResNet-18	ResNet-50
FGSM	**18.27** (1.54)	2.07 (1.54)	6.69 (1.54)
BIM	**62.23** (1.24)	8.00 (1.52)	20.22 (1.18)
PGD	**41.48** (1.17)	3.72 (1.26)	14.09 (1.12)
C&W	**33.65** (2.42)	7.31 (1.46)	26.53 (1.98)

  - **使用这三种网络作为D模型窃取VGG-16网络时，性能也优于其他与训练模型，其中VGG-13远超其他网络模型，这表明替代模型与目标模型的结构越相似，窃取效果越好**

算法结果

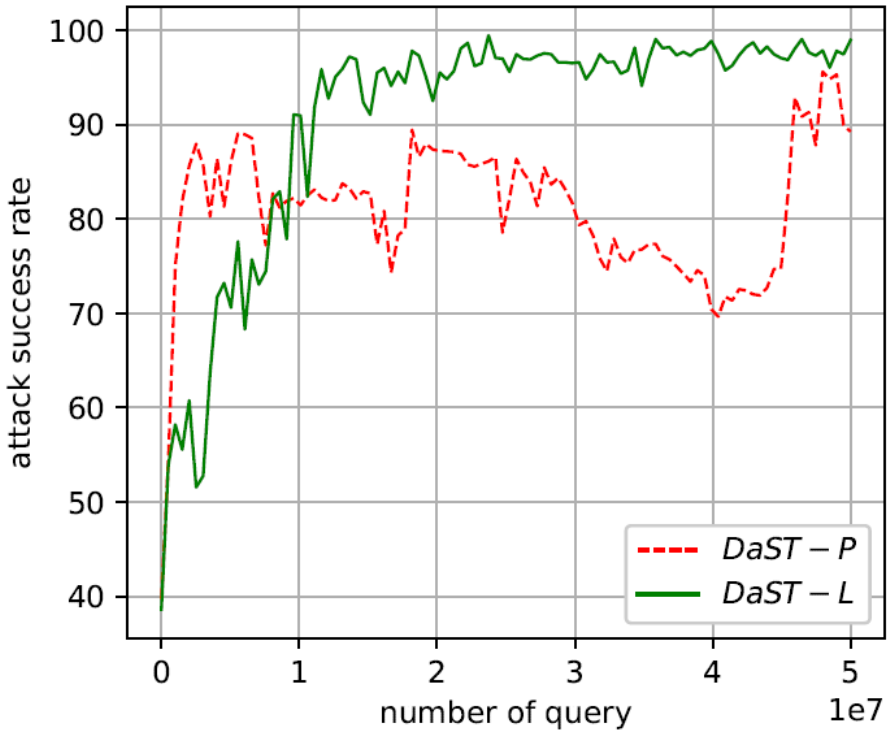
- 攻击Microsoft Azure上的在线模型（黑盒）
  - 训练好的大型网络（ 5个卷积层）为替代模型

Attack	Non-targeted (%)		
	Pre-trained	DaST-P	DaST-L
FGSM	77.96 (5.41)	96.83 (5.25)	**98.21** (5.36)
BIM	66.25 (4.81)	96.42 (4.79)	**98.35** (4.72)
PGD	59.23 (3.99)	90.63 (3.88)	**96.97** (3.96)
Attack	**Targeted** (%)		
	Pre-trained	DaST-P	DaST-L
FGSM	13.52 (5.46)	32.00 (5.21)	**43.99** (5.37)
BIM	19.31 (4.88)	50.21 (4.90)	**71.15** (4.56)
PGD	19.31 (4.60)	45.66 (4.46)	**65.91** (4.32)

  - DasT-L的性能优于DaST-P的原因是目标模型结构简单

算法结果

- 查询次数与准确率的关系
  - 当利用被攻击模型的输出概率训练时，准确率曲线快速上升，之后发生震荡
  - 获取被攻击模型的输出标签的情况下，反而能够让准确率曲线更加平滑



优劣分析

- 横向对比
  - 优势
    - 利用对抗生成网络的思想，使用生成器产生数据训练替代模型，再由替代模型与目标模型的差异修正生成器
  - 不足
    - 生成器的分支结构使得在窃取大规模模型时，需要耗费大量的训练资源
    - 需要生成大量数据进行目标模型的输入查询，为其生成标签；由于生成数据的不可控，还会有大量无效查询
- 纵向对比
  - 减少了对目标模型的先验知识的依赖性

应用总结

应用总结

- 算法的应用领域
  - 通过模型窃取生成对抗样本
  - 通过模型窃取获得原始训练数据

- 未来的发展
  - 模型窃取方法的可迁移性
  - 模型窃取的数据依赖性

参考文献

[1] Zhou M , Wu J , Liu Y , et al. DaST: Data-free Substitute Training for Adversarial Attacks[J]. IEEE, 2020.

[2] Yuan X , Ding L , Zhang L , et al. ES Attack: Model Stealing against Deep Neural Networks without Data Hurdles. 2020.

[3]F Tramèr, Fan Z , Juels A , et al. Stealing Machine Learning Models via Prediction APIs[J]. 2016.

[4] Papernot N , Mcdaniel P , Goodfellow I , et al. Practical Black-Box Attacks against Machine Learning[J]. ACM, 2016.

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。