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

基于知识蒸馏的模型窃取方法

硕士研究生 丁杨

2021年11月14日

内容提要

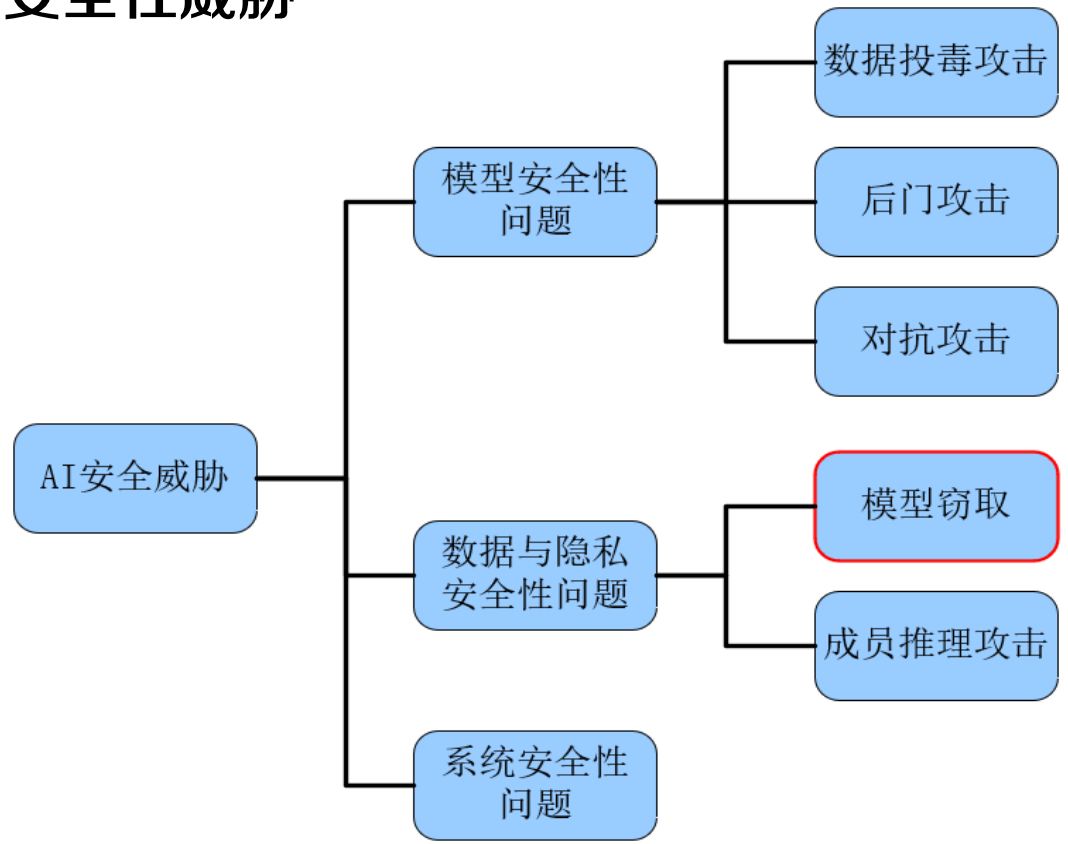
- 背景简介
- 基本概念
- 算法原理
- 优劣分析
- 应用总结
- 参考文献

背景简介

- 预期收获
  - 1. 了解模型窃取方法的发展历史
  - 2. 理解基于知识蒸馏的模型窃取方法的技术原理
  - 3. 了解模型窃取在网络安全领域中的应用

背景简介

- 人工智能的发展
  - 推动社会经济各个领域从数字化、信息化向智能化发展
  - 面临着严重的安全性威胁



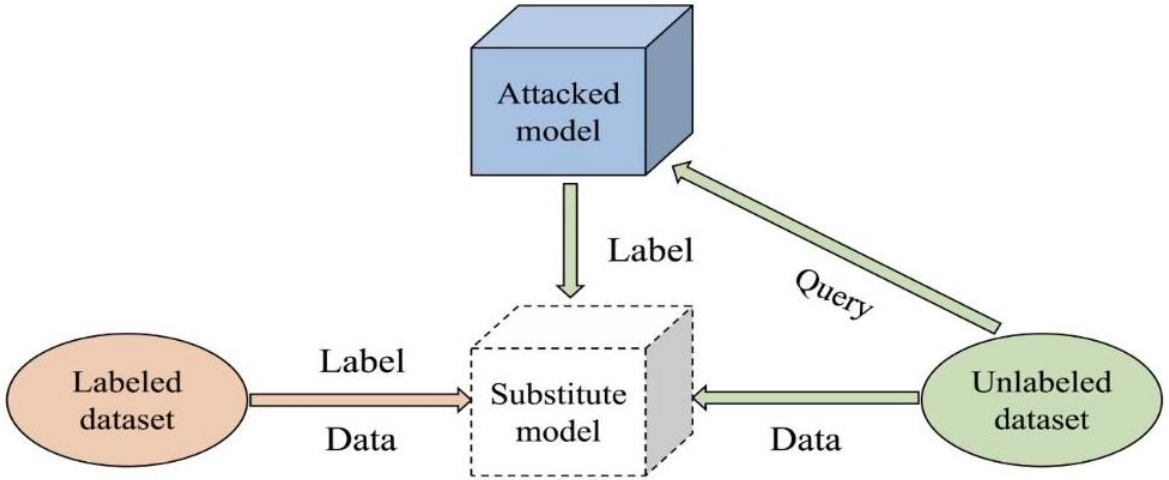
背景简介

- 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
  - 最新的研究则关注如何在未知训练数据的情况下进行模型窃取

基本概念

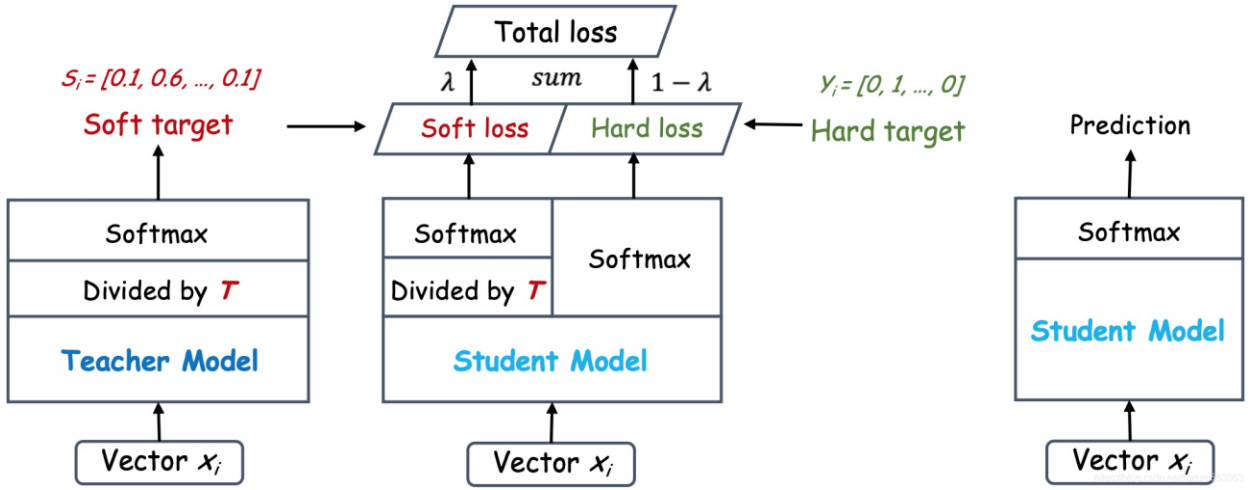
基本概念

- 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
    - 进行对抗攻击
  - 方法
    - 构建方程求解参数
    - 训练替代模型



基本概念

- 知识蒸馏
  - 概念：把复杂模型或者多个模型（Teacher）学到的知识迁移到另一个模型（Student）上
  - 本质：模型压缩，让小模型去学习大模型的知识，即让Student模型的输出拟合Teacher模型的输出
  - Teacher模型和Student模型没有网络结构的限制



基本概念

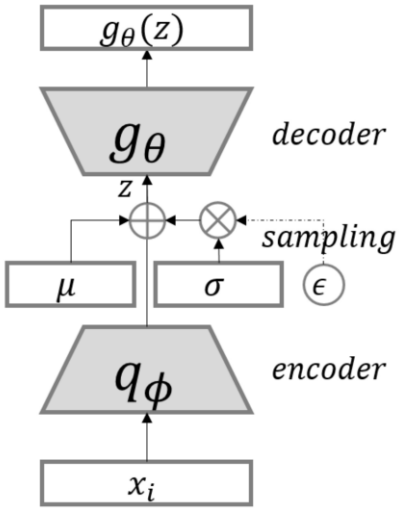
- 相对熵
  - 概念：又称为Kullback-Leibler散度（Kullback-Leibler divergence）或信息散度（Information divergence），是两个概率分布差异的非对称性度量
  - 设 $P(x)$ $Q(x)$ 是随机变量 X 上的两个概率分布，则在离散和连续随机变量的情形下，相对熵的计算公式分别为:

$$\mathrm{KL}(P\|Q)=\sum P(x)\log\frac{P(x)}{Q(x)}$$

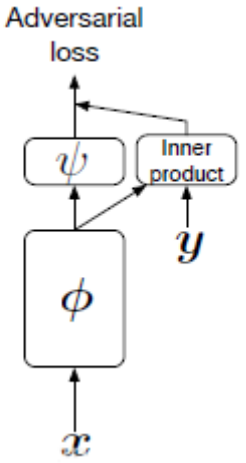
$$\mathrm{KL}(P\|Q)=\int P(x)\log\frac{P(x)}{Q(x)}dx$$

基本概念

- VAE（变分自编码器）
  - 通过编码过程生成输入样本分布的均值与方差，然后通过采样的方法来复原输入样本分布，并且使用复原的分布和真实分布的距离来进行参数的调节



- SNGAN（频谱归一化GAN)
  - 输入首先经过网络 ϕ 提取特征，然后把特征分成两路，一路与经过编码的类别标签 y 做点乘，另一路通过网络 ψ 映射成一维向量，最后两路相加，作为神经网络最终的输出



算法原理

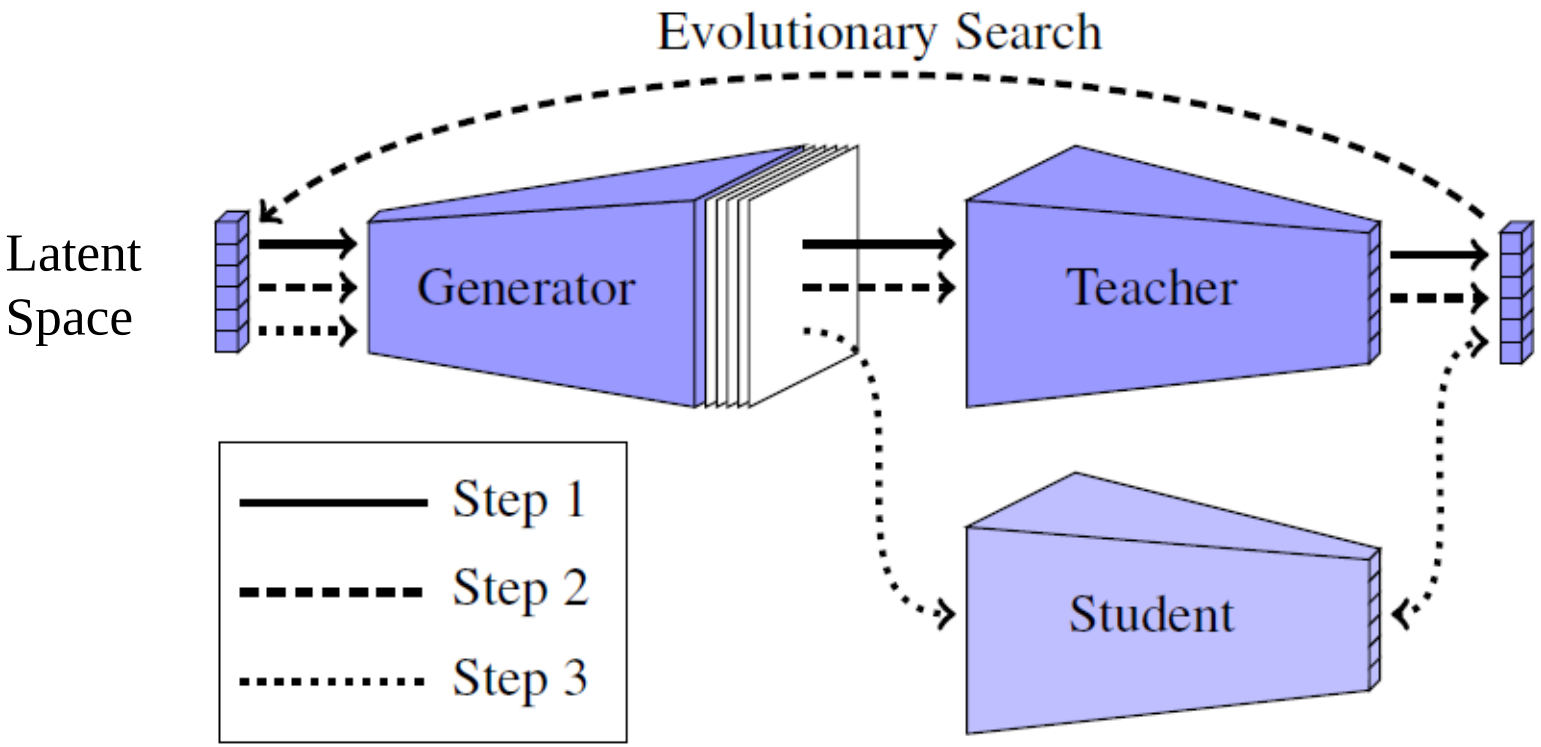
算法原理Black-Box Ripper

T	模型窃取
I	Teacher 模型
P	1. 生成随机样本，随机样本通过生成器得到数据样本 2. 调用Teacher 模型，得到数据样本的分类预测概率 3. 基于进化策略优化样本空间 4. 训练Student 模型
O	Student 模型

P	无法获得Teacher 模型原始的训练数据与内部参数
C	攻击者可以获得Teacher 模型输出的分类预测概率
D	生成与原训练数据分布相近的训练数据
L	NeurIPS 2020

算法原理

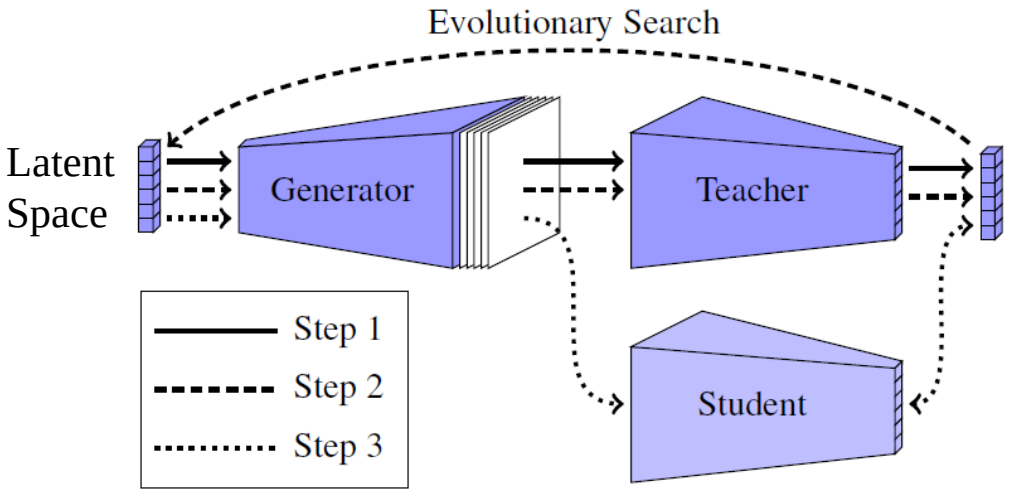
- 算法流程图



$$\min_{\theta_S} \| S(Z\downarrow\uparrow,\ X'\downarrow) - T(X\downarrow\uparrow,\ X'\downarrow) \|$$

算法原理

- 算法流程
  - 在潜在空间中生成一组随机样本
  - 随机样本通过生成器得到数据样本
  - 数据样本通过Teacher 模型得到对应的分类预测概率
  - 采用进化算法，利用类概率信息优化潜在空间
  - 使用优化的潜在空间生成数据，训练Student 模型



$$\min_{\theta_S} \| S(Z\downarrow\uparrow,\ X'\downarrow) - T(X\downarrow\uparrow,\ X'\downarrow) \|$$

算法原理

- 进化算法
  - 目的:在期望的类上生成高置信度的数据样本
  - 过程:

Algorithm 1 Evolutionary Optimization Algorithm

Input: y - desired class label, T - black-box teacher, G - generator trained on Z.
Hyperparameters: K - population size, k - elite size, u - latent space boundary, t - threshold for stopping criterion.
Output: p^* - generated data sample with high confidence on desired class y.

procedure OPTIMIZE(y, T, G, K, k)
  Initialize population from a uniform distribution: $P \leftarrow \{U(-u, u)\}_K$
  Select fittest latent vector: $p^* \leftarrow min_{p \in P} V(p, y, T, G)$
  while $V(p^*, y, T, G) \geq t$ **do**
    Select fittest k vectors: $P_e \subset P$
    Uniformly sample $K - k$ copies from P_e: $P_c \leftarrow \{\mathcal{U}(P_e)\}_{K-k}$
    Mutate copied vectors with Gaussian noise: $P_c \leftarrow P_c + \mathcal{N}(0, 1)$
    Replace old population with new one: $P \leftarrow P_e \cup P_c$
    Select fittest latent vector: $p^* \leftarrow min_{p \in P} V(p, y, T, G)$
  return $G(Z\downarrow\uparrow,\ p^*\downarrow)$

$$\min_v V(v, y, T, G) = \min_v \sum_{i=1}^{n} \left(T(X\downarrow\uparrow,\ G(Z\downarrow\uparrow, v\downarrow)\downarrow) - y_i\right)^2 = \min_v \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

实验结果

- 数据集
  - CIFAR-10、CIFAR-100 和Fashion-MNIST
- 实验条件
  - Teacher 模型的训练数据未知
  - 攻击者可以获得Teacher 模型的分类预测概率
- 实验设置
  - 预训练生成器(VAE，SNGAN)
  - 预训练Teacher 模型
- 评价方法
  - Student 模型的准确率

算法结果

- **CIFAR-10数据集上的实验**
  - **Teacher 模型使用AlexNet 架构**
  - **Student 模型使用Half-AlexNet 架构**

Proxy Dataset	CIFAR-100 90 classes	CIFAR-100 40 classes	CIFAR-100 10 classes	CIFAR-100 6 classes
Teacher Accuracy	82.5	82.5	82.5	82.5
Knockoff Nets [31]	74.5	65.7	46.6	36.4
ZSKD [28]	69.5	69.5	69.5	69.5
DeGAN [1]	**80.5**	76.3	72.6 ± 3.3	59.5
Black-Box Ripper (**Ours**)	79.0 ± 0.2	$\mathbf{76.5} \pm 0.1$	$\mathbf{77.9} \pm 0.3$	$\mathbf{69.9} \pm 0.2$

  - **随着代理数据集中类的数量越来越少，Student 模型的准确率下降幅度受影响较小**

算法结果

- **Fashion-MNIST 数据集上的实验**
  - **Fashion-MNIST 作为真实数据集，CIFAR-10 作为代理数据集**
  - **Teacher 模型分别选择VGG-16 与LeNet**
  - **Student 模型对应选择VGG-16 与Half-LeNet**

Architectures	VGG-16	LeNet → Half LeNet
Teacher Accuracy	94.2	89.9
Knockoff Nets [31]	82.9	77.8
ZSKD [28]	-	79.6
DeGAN [1]	-	**83.7**
VAE (no evolutionary optimization)	78.3	73.1
SNGAN (no evolutionary optimization)	87.6	80.0
Black-Box Ripper with VAE (**Ours**)	86.1	78.8
Black-Box Ripper with SNGAN (**Ours**)	**90.0**	82.2

算法原理

算法原理 DFME

T	模型窃取
I	Victim模型

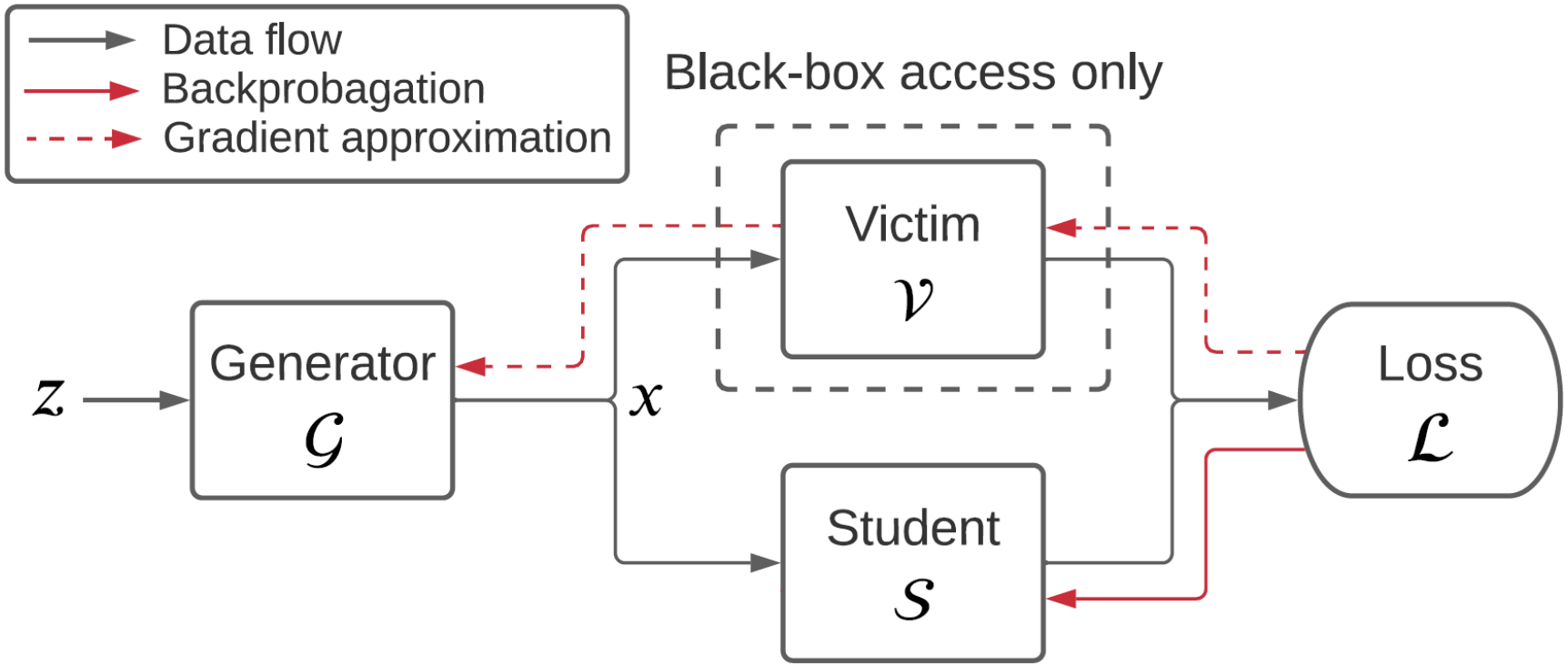
P	For {
	1. 生成器生成数据
	2. 将数据通过Victim模型与Student模型
	3. 计算损失函数
	4. 反向传播，优化生成器与Student模型
	}

O	Student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Victim模型获得其分类预测概率
D	Victim模型的梯度信息
L	CVPR 2021

算法原理

- 算法流程图



$$\min_{\mathcal{S}} \max_{\mathcal{G}} \mathbb{E}_{z \sim \mathcal{N}(0,1)} \left[\mathcal{L}(\mathcal{V}(\mathcal{G}(z)), \mathcal{S}(\mathcal{G}(z))) \right]$$

算法原理

- 损失函数的选择
  - 知识蒸馏中，大多数的损失函数使用相对熵

$$\mathcal{L}_{\mathrm{KL}}(x) = \sum_{i=1}^{K} \mathcal{V}_i(x) \log \left(\frac{\mathcal{V}_i(x)}{\mathcal{S}_i(x)} \right)$$

  - 随着Student 模型与Victim 越来越接近，相对熵的梯度快速衰减，使得生成器难以收敛，因此选择了L1范数作为损失函数

$$\mathcal{L}_{\ell_1}(x) = \sum_{i=1}^{K} |v_i - s_i|$$

算法原理

- 反向传播
  - Student 模型的参数梯度下降

$$
\begin{aligned}
&z \sim \mathcal{N}(0, 1) \\
&x = G(z; \theta_{\mathcal{G}}) \\
&\text{compute } \mathcal{V}(x), \mathcal{S}(x), \mathcal{L}(x), \nabla_{\theta_{\mathcal{S}}} \mathcal{L}(x) \\
&\theta_{\mathcal{S}} = \theta_{\mathcal{S}} - \eta \nabla_{\theta_{\mathcal{S}}} \mathcal{L}(x)
\end{aligned}
$$

- 梯度逼近
  - 通过函数在多个u_i方向上的一小步的平均变化来计算

$$
\nabla_{\mathrm{FWD}} f(x) = \frac{1}{m} \sum_{i=1}^{m} \frac{f(x + \epsilon \mathbf{u_i}) - f(x)}{\epsilon} \mathbf{u_i}
$$

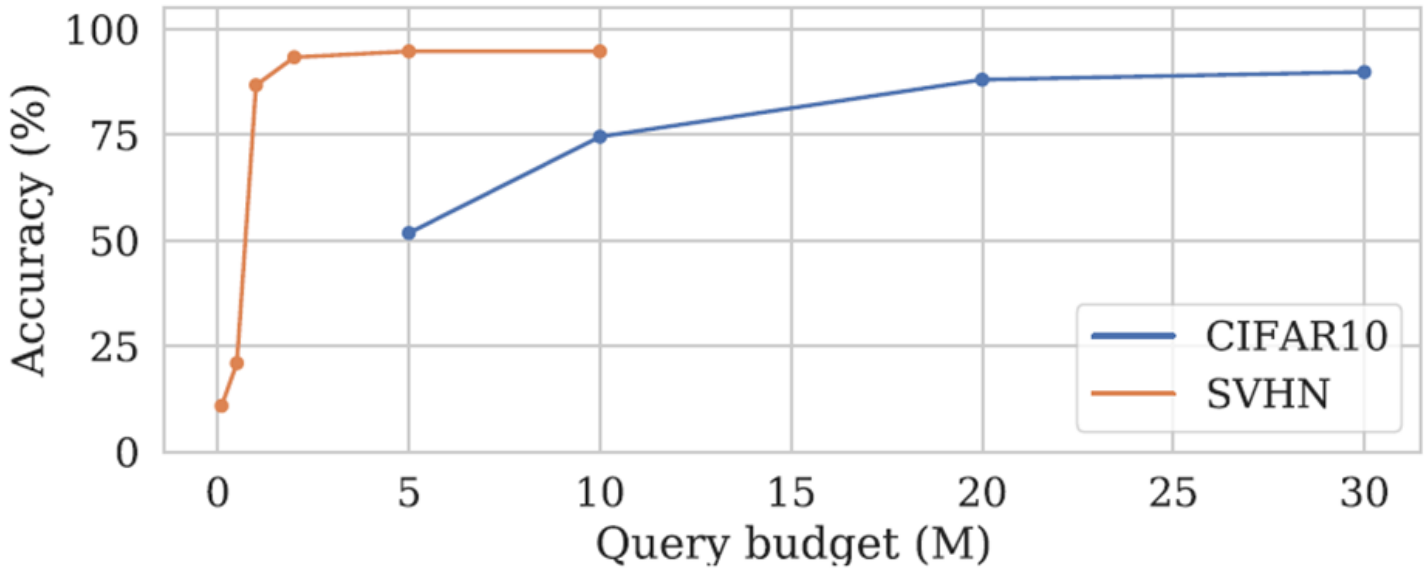
实验结果

- 数据集
  - SVHN、CIFAR-10
- 实验条件
  - Victim 模型的训练数据未知
  - 攻击者可以获得Victim 模型的分类预测概率
- 实验设置
  - 生成器使用三个卷积层，与线性上采样层、批量标准化层和除最后一层外所有层的ReLU 激活函数相互交织
  - 预训练Victim 模型
- 评价方法
  - Student 模型的准确率

实验结果

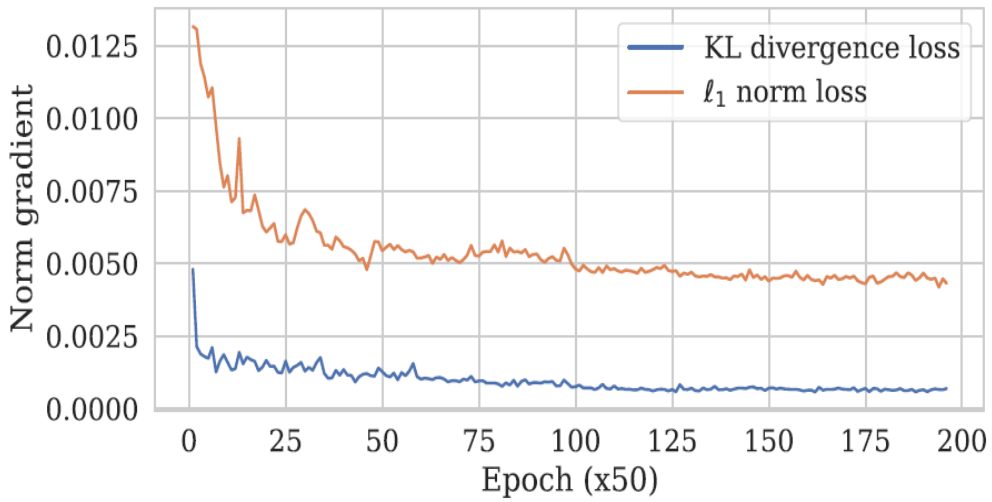
- **Student模型准确率实验**

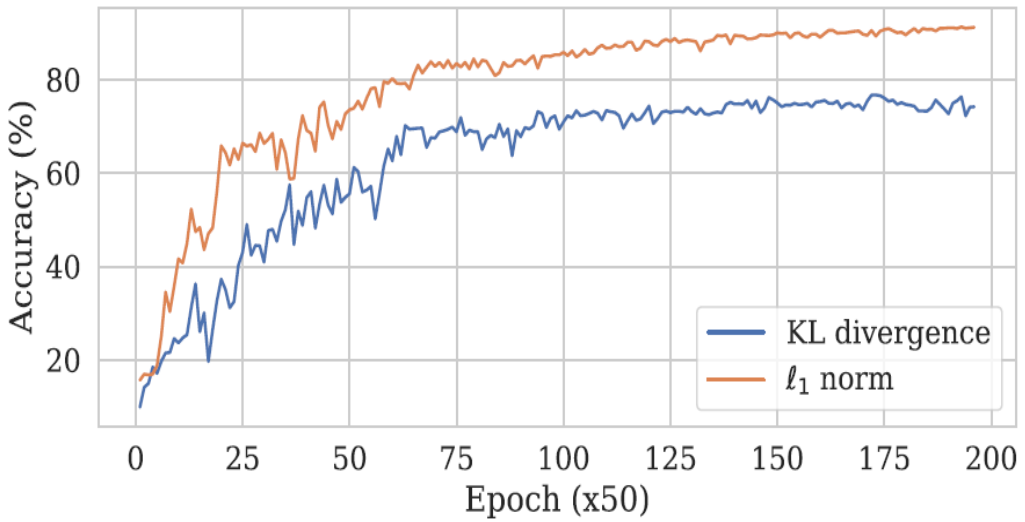
Dataset (budget)	Victim accuracy	DFME	DFME-KL	MAZE* [20]	Log-Probabilities
CIFAR10 (20M)	95.5%	88.1% (0.92×)	76.7% (0.80×)	45.6% (0.48×)	73.2% (0.77×)
SVHN (2M)	96.2%	95.2% (0.99×)	84.7% (0.88×)	91.1% (0.95×)	94.4% (0.98×)



实验结果

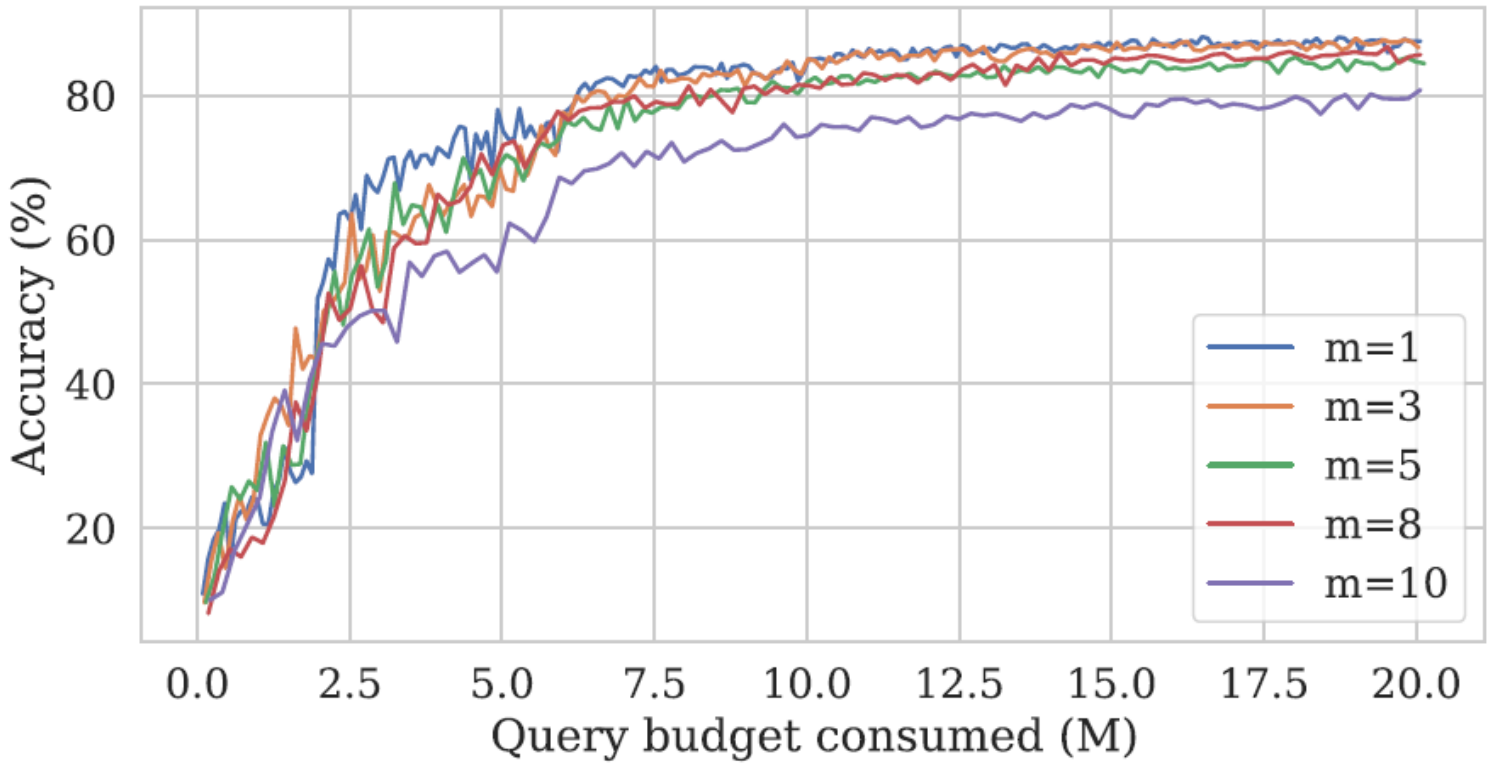
- 消融实验
  - 损失函数的不同选择





实验结果

- 消融实验
  - 参数m的选择



优劣分析

- 优势
  - 两种算法都利用知识蒸馏的思想，使用生成器产生数据训练Student 模型，无需 Teacher 模型的训练数据
  - DFME 利用对抗生成网络的思想，使用生成器产生数据计算模型差异，再由差异修正生成器与替代模型，生成器随实际情况调整
- 不足
  - 需要Teacher 模型的分类预测概率信息
  - 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，会有大量无效查询

应用总结

应用总结

- 算法的应用领域
  - 通过模型窃取获取目标模型的参数信息，使用白盒方法生成对抗样本
  - 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
  - 不依赖模型类概率输出的条件下进行模型窃取

参考文献

[1] Kariyappa S, Prakash A, Qureshi M K. Maze: Data-free model stealing attack using zeroth-order gradient estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 13814-13823.

[2] Truong J B, Maini P, Walls R J, et al. Data-free model extraction[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 4771-4780.

[3] S. Kariyappa, A. Prakash and M. K. Qureshi, "MAZE: Data-Free Model Stealing Attack Using Zeroth-Order Gradient Estimation," 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 13809-13818, doi: 10.1109/CVPR46437.2021.01360.

[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。