Activethief: Model extraction using active learning and unannotated public data

Soham Pal, Yash Gupta, Aditya Shukla, Aditya Kanade,
Shirish Shevade, Vinod Ganapathy
Indian Institute of Science, Google Brain

Shilong Zhao †

August 21, 2023

† Institute of Computing Technology
University of Chinese Academy of Sciences

Outline

Introduction

Introduction

近年来，Deep Learning 作为一种特别成功和流行的 ML 模型，被越来越多地部署在生产中。但是，训练深度神经网络是一项昂贵的活动，需要高质量的数据集、强大的计算资源。许多公司将训练好的模型部署在云上，对公众提供付费访问接口，通常称为机器学习即服务 (Machine Learning as a Service, MLaaS) 平台。开发人员通过应用程序编程接口 (Api) 访问这些模型。

但这种模式已被证明容易受到 Model Extraction/Stealing Attack (Tramer et al 2016)。攻击者通过查询 MLaaS 提供的模型获得标记数据，然后训练替代模型逼近原模型。最终攻击者可以免费使用替代模型，将其作为竞争服务提供，或者使用它来帮助实施对原模型的攻击。

Introduction

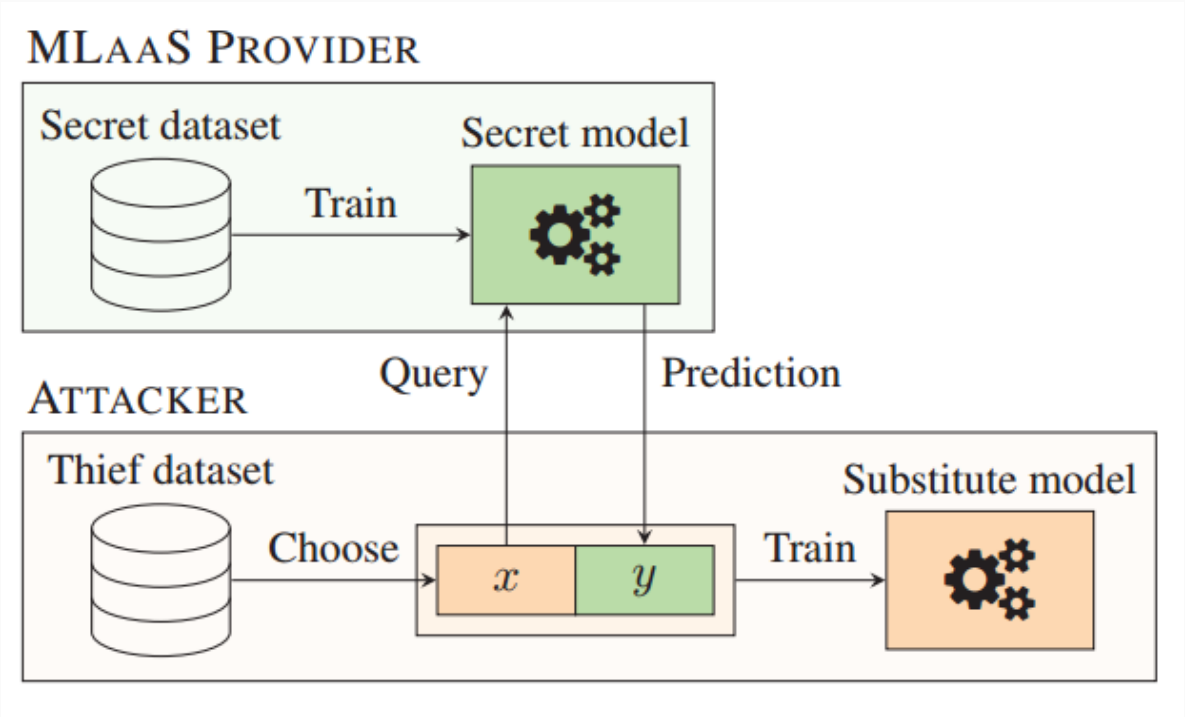

图 1: Overview of model extraction

Introduction

MLaaS 设定下的 Model Extraction/Stealing Attack (MEA) 通常面临以下两个挑战：

1. 如何获取攻击所用的数据集；
2. 如何在有限的查询预算下尽可能提高攻击的效率。

Contribution：

1. 提出了一个新的 MEA 框架，结合 Active Learning 利用未注释的公共数据提高了攻击效果；
2. 利用 Active Learning 提高了查询效率；
3. 隐匿性更强，攻击没有被最先进的检测方法 PRADA 检测到。

Framework

Framework

作者提出的 ActiveThief 框架如下：MLaaS 使用的模型为 Secret Model，数据为 Secret Dataset；攻击者使用的模型为 Substitute Model，数据为 Thief Dataset. 例如图像分类任务 {Secret Dataset = MINST, Thief Dataset = ImageNet}.

1. 攻击者从 Thief Dataset 中随机选取一个子集 S_0 作为初始样本；
2. 在第 i 次迭代 $(\mathrm{i} = 0, 1, 2, , \mathrm{N})$ 中，攻击者使用 S_i 中的样本查询 Secret Model f 查询，得到标注集合 $\mathrm{D}_\mathrm{i} = (\mathrm{x}, \mathrm{f}(\mathrm{x})) : \mathrm{x} \in \mathrm{S}_\mathrm{i}$；
3. 在所有标注集合 $\cup_{\mathrm{t}=0}^{\mathrm{i}} \mathrm{D}_\mathrm{t}$ 上训练 Substitute Model $\tilde{\mathrm{f}}$.
4. 攻击者使用剩余样本查询 Substitute Model $\tilde{\mathrm{f}}(\approx \mathrm{f})$，获得剩余样本的近似标签 $\tilde{\mathrm{D}}_\mathrm{i} = (\mathrm{x}, \tilde{\mathrm{f}}(\mathrm{x})) : \mathrm{x} \in \mathrm{S}_1 \cup \cdots \cup \mathrm{S}_\mathrm{i}$.
5. 使用 Active Learning 子集选择策略从 $\tilde{\mathrm{D}}_\mathrm{i}$ 中选择接下来要查询的 k 个样本集合 $\mathrm{S}_{\mathrm{i}+1}$.

Framework

以上过程在每次迭代中从头开始重新训练 Substitute Model。迭代次数 N，每次标注的样本个数 k、初始样本个数 $|S_0|$ 是超参数。

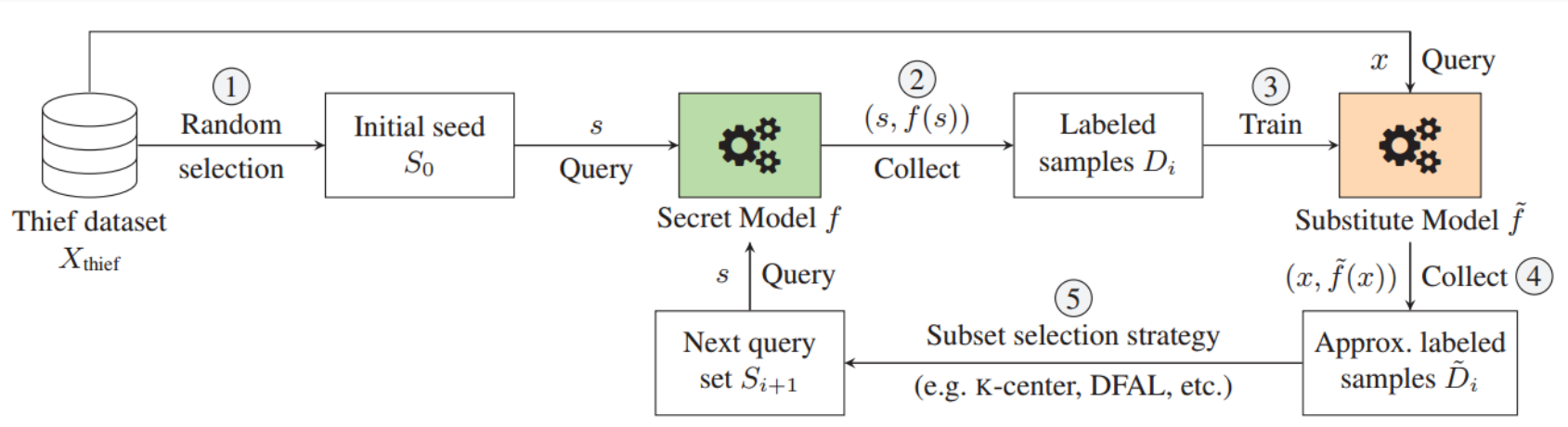

图 2: ActiveThief 框架

Metric

使用 Secret Model f 和 Substitute Model $\tilde{\mathrm{f}}$ 在 Secret Dataset 上表现的一致性衡量 MEA 的效果：

$$\mathrm{Agreement}(\mathrm{f}, \tilde{\mathrm{f}}) = \frac{1}{|\mathrm{X}^{\mathrm{test}}_{\mathrm{secret}}|} \sum_{\mathrm{x} \in \mathrm{X}^{\mathrm{test}}_{\mathrm{secret}}} \mathrm{I}(\mathrm{f}(\mathrm{x}) = \tilde{\mathrm{f}}(\mathrm{x}))$$

其中其中 $\mathrm{I}(\cdot)$ 为指标函数（相同为 1，不同为 0）。

Subset selection strategies

在每次迭代中，攻击者通过查询 Secret Model f，从 k 个 Thief Dataset 样本中选择一个新的 $\mathrm{S_i}$ 集进行标记。每个子集选择策略都将近似标签集 $\tilde{\mathrm{D}}_\mathrm{i} = (\mathrm{x_n}, \tilde{\mathrm{y}}_\mathrm{n})$ 作为输入，并返回集合 $\mathrm{S_{i+1}}$.

1. Random strategy：均匀随机选择 k 个样本；
2. Uncertainty strategy：这种方法基于不确定性抽样 (Lewis and Gale 1994)。根据 Substitute Model 输出的预测概率向量 $\tilde{\mathrm{y}}_\mathrm{n}$ 计算熵 $\mathrm{H_n} = -\sum_\mathrm{j} \tilde{\mathrm{y}}_{\mathrm{n,j}} \mathrm{log}\tilde{\mathrm{y}}_{\mathrm{n,j}}$，(其中 j 为标签 index)。该策略选择熵值最高对应的 k 个样本；
3. K-center strategy：使用 Sener 和 Savarese(2018) 的贪婪 k 中心算法。该策略首先将初始样本的概率被标记为聚类中心，在随后的每次迭代中，选择距离所有现有中心最远 k 个样本：

$$(\mathrm{x}_0^*, \tilde{\mathrm{y}}_0^*) = \mathrm{argmax}_{(\mathrm{x_n}, \tilde{\mathrm{y}}_\mathrm{n}) \in \tilde{\mathrm{D}}_\mathrm{i}} \mathrm{min}_{(\mathrm{x_m}, \mathrm{y_m}) \in \mathrm{D_{i-1}}} ||\tilde{\mathrm{y}}_\mathrm{n} - \tilde{\mathrm{f}}(\mathrm{x_m})||_2^2$$

然后标记选定的 x_0^*（此结果随后被视为一个中心，即 $\mathrm{D_{i-1}}$ 的成员）。重复这个过程，直到 k 个样本 $\mathrm{x}_0^*, \mathrm{x}_1^*, \ldots, \mathrm{X}_\mathrm{k}^*$ 被选中。

Subset selection strategies

4. DFAL strategy：使用 Ducoffe 和 Precioso(2018) 的 DeepFool-based Active Learning(DFAL) 算法。将 DeepFool (Moosavi-Dezfooli, Fawzi, and Frossard 2016) 应用于每个样本 $\mathrm{x_n}$，得到一个被 Substitute Model 错误分类的扰动后的样本 $\hat{\mathrm{x}}_\mathrm{n}$。计算扰动 $\alpha_\mathrm{n} = ||\mathrm{x_n} - \hat{\mathrm{x}}_\mathrm{n}||_2^2$，选取扰动 α_n 最小的 k 个样本 $\mathrm{x_n}$。
5. DFAL + K-center strategy：在该策略中，DFAL 策略首先用于选择 ρ 个样本作为初始子集（实验中设置 ρ= 总预算），然后使用 K-center 策略选择 k 个样本。

Experiment

Experimental setup

1. Datassets

Secret datasets. **对于图像分类，使用** MNIST、CIFAR-10 **和** GTSRB (Stallkamp et al . 2012)。**对于文本分类，使用** MR (Pang and Lee 2005)、IMDB (Maas et al 2011) **和** AG News2。

Thief dataset. **对于图像分类，使用** ImageNet。**对于文本，使用** WikiText-2 (Merity et al . 2017)。

2. Model architectures

Image classification. CNN.

Text classification. **首先使用** Word2vec **获得词嵌入** (Secret Model **使用预训练的嵌入**，Substitute Model **从头开始学习**)。**使用了** CNN **和** RNN **两种架构**。

3. **其他设置略**。

Experimental results

在知道 Secret Model 结构的情况下，实验证明了以下结论：

1. Active Learning 的有效性（图 3 和表 1）。使用 Active Learning 选择数据进行标注，模型的一致性比随机选择；
2. 联合策略（DFAL+K-center）的有效性（表 1）。多数情况下使用联合策略会有性能提升；
3. 查询预算的影响（表 1）。在增加查询预算时，一致性得到了改善；
4. 通用数据的有效性（表 1）。将使用通用数据的结果与使用均匀噪声（多维 U[0,1]）的 SNPD(Synthetic Non-Problem Domain) 数据进行比较。结果使用 SNPD 数据在所有数据集上的一致性都较低；
5. 迭代次数的影响（表 2a）。显示，随着迭代次数的增加，相同预算的一致性有所提高；

Experimental results

6. 访问输出概率得分的影响（表 2b）。访问 Secret Model 的输出概率可以提高一致性；
7. 对抗样本的可转移性（表 3）。使用 FGSM 方法利用 Substitut Model 的梯度信息生成对抗样本，计算这些样本在攻击 Secret Model 时的成功率。从结果可以看出，ACTIVETHIEF 比 Papernot 等人能够更好地实现对抗样本的可转移。

在不知道 Secret Model 结构的情况下，实验证明了以下结论：

1. 文本分类任务中模型结构的健壮性（表 4）。；
2. 图像分类任务中模型参数的影响（表 5）；

最后作者实验并尝试解释了 ActiveThief 如何逃避 PRADA 的检测（图 4）。

Conclusion

Conclusion

本文介绍了一种新的模型提取框架 ACTIVETHIEF，实验结果表明：

1. 仅使用一个未注释的公共数据集，就可以提取针在不同 Secret 数据集上训练的模型;
2. 对于查询预算有限的图像和文本域，以及跨架构的 MEA 都是可能的;
3. ACTIVETHIEF 没有被最先进的 MEA 检测方法检测到;
4. 使用该方法提取的模型与 Secret 模型具有很好的一致性;
5. 提高了对抗样本的可转移性。