**CloudLeak: Large-Scale Deep Learning Models**

**Stealing Through Adversarial Examples**

Honggang Yu et.al

**Abstract**

1. 提出一种黑盒攻击，针对云平台的大型DNN。

2. 提出的攻击特点在于：比现存的攻击需要的问询次数更少；也可以用于评估DNN网络的鲁棒性。

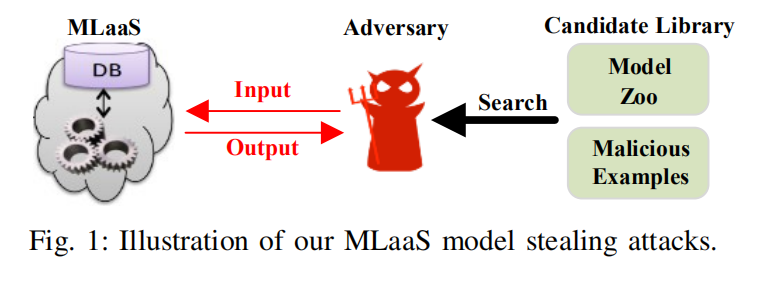
**Introduction**

MLaaS（基于云的机器学习服务）中，用户付费并通过访问API获得他们的input的预测。MLaaS中的DNN模型和训练数据都是不公开的。

但是最近的研究表明：简单的问询与得到模型的预测，就可以进行模型提取，model泄露；MIA不仅适用于过拟合的模型，对泛化良好的模型也很有效，训练数据泄露。

本文的攻击思想是：

通过使用基于边缘、对抗性、主动学习的算法，找到恶意样本，（这一步可以提高问询效率），然后用恶意样本问询黑盒API得到（input,output）对，构成合成数据集，并用它们来重新训练从Model Zoo中选出的**替代模型**。



**Contributions：**

1.提出了一种新的**对抗攻击方式**FeatureFool，FeatureFool攻击局部替代模型，这些替代模型采用内部表示来产生恶意样本的子集。恶意样本是被用来问询受害模型，得到受害模型和提取的模型之间的决策边界的距离，并大大减少了问询次数。

2.针对大规模DNN设计了一个**黑盒模型窃取攻击**，使用对抗主动攻击和迁移学习（从现有训练好的模型，如：AlexNet）来加快窃取过程。

3.在一些常见的商业平台**评估**了攻击框架。实验表明本文的窃取攻击在问询次数更少的情况下，得到的局部替代模型达到了与受害模型相近的效果。

**Related work**

**Transfer Learning：** 将某个领域（source domains）上学习到的知识应用到不同但相关的领域（target domains）中。（可简单理解为举一反三）

**Adversarial Attack in Deep Learning：**对纯净样本加入微小扰动生成对抗样本。

**Active Learning：**主动学习通过某种策略找到未进行类别标注的样本数据中**最有价值的数据**（informative data），交由专家进行人工标注后，将标注数据及其类别标签纳入到训练集中迭代优化分类模型（二分类为“有价值”“无价值”），改进模型的处理效果。本文利用一系列对抗样本生成算法，增大了分类边界有价值样本的密度，因此提高了问询的效率。

**Model Extraction Attacks：**通过问询提取到一个和目标模型表现几乎相同的模型。和以往研究不同，本文提出一种更有效的**黑盒攻击**来窃取大型深度学习模型，使用的是一种特殊的**迁移学习**和特意构造的**对抗样本**。

（这篇文章融合了迁移学习、对抗攻击、主动学习，具体的方式显示在之后的框架图中。）

**Background**

***名词解释（建议结合框架图理解）***

1.Victim model：受害模型

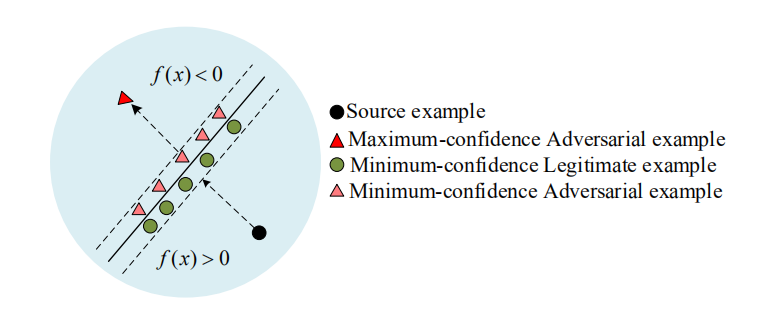
2.Substitute model：敌手提取到的模型，使敌手可以任意使用而不产生费用

3.恶意样本：敌手使用的样本，用于问询

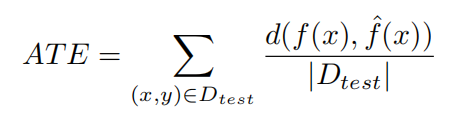
4.黑盒：敌手只能用恶意样本来问询受害模型并得到预测或置信度，而不知道内部的网络结构、精确数据集、参数、权重等。

5.FeatureFool：用于生成具有**不同置信度**的对抗样本。本质上是一种**对抗攻击方式**，利用基于特征的优化算法来产生对抗样本，达到误导模型的目的。

6.Synthetic dataset：包含两类样本--minimum-confifidence legitimate examples和minimum-confifidence adversarial examples（也就是说，文中这种主动学习方式就是目标是找到min-confidence 数据，这类数据会使样本的不确定性达到最大化，是很好的对抗样本）。这两类样本会提供更多关于分类器仿射超平面的信息。以二分类器为例，在图中可以看到，这两类样本（绿色和粉色）距离仿射超平更近，因此性能最好：



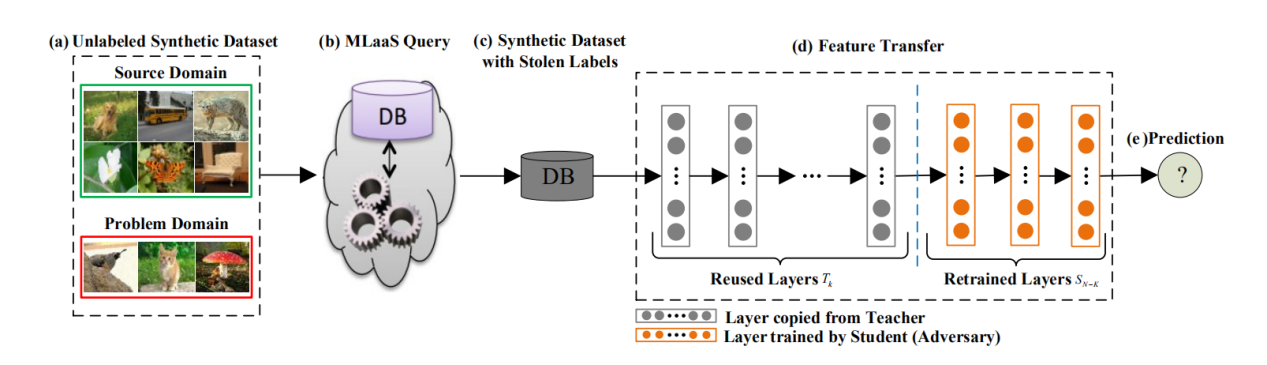
**7.评估指标ATE**：在测试集上使用Average Test Error (ATE)，ATE衡量了预测值与真实值之间的平均距离，对敌手来说，ATE越小说明substitute model与victim model越相近：



**Model Theft Attacks**

***Transfer Architecture Construction***

本文攻击的迁移框架如图：



流程是从左向右：

（a）生成unlabeled对抗样本，将其作为合成数据集

（b）使用合成数据集问询victim model

（c）用victim model的输出为对抗样本打上label

（d）用合成数据集训练local substitute model

（e）local substitute model做预测，要求local substitute model的表现要和victim model几乎一样好。

**具体理解：**

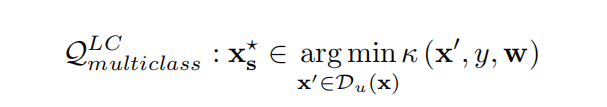
1. (a) 中采用FeatureFool找到basic informative data**+**margin-based adversarial AL找到unlabeled data中的informative data

【除了FeatureFool，本文也采用了另外四种方式进行对抗样本的生成，但FF效果最好】

1. (a)+(b)+(c)+(d)都属于主动学习的过程（找出informative data，打标签，优化substitute model）。
2. 迁移学习的过程主要在(d)，用合成数据集进行迁移学习**重训练**substitute model，得到的模型其边界和受害模型的相似的。对于source task，有很多针对不同任务、预训练好的模型可以直接利用，本文使用的是四种预训练好的模型AlexNet, VGG19, VGGFace 和 ResNet50作为基础结构，然后对其中的层进行了改变（具体的变化比较复杂，感兴趣的话可以在论文第四页右侧找到）。
3. (d) 训练substitute model（substitute model是怎么来的：从Model Zoo中预训练出的模型中构建出来），合成数据对substitute model进行**微调**，微调时只重新训练最后一部分全连接层，前面的层不作处理。

**这里再重点解释一下本文提出的主动学习方式：**

本文提出了一种新的主动学习方式--margin-based adversarial AL(基于边缘的对抗主动学习)。该方法认为，在unlabeled data里只有少数数据是有价值的，其余的数据都是多余的。Margin-based adversarial AL可表示为：

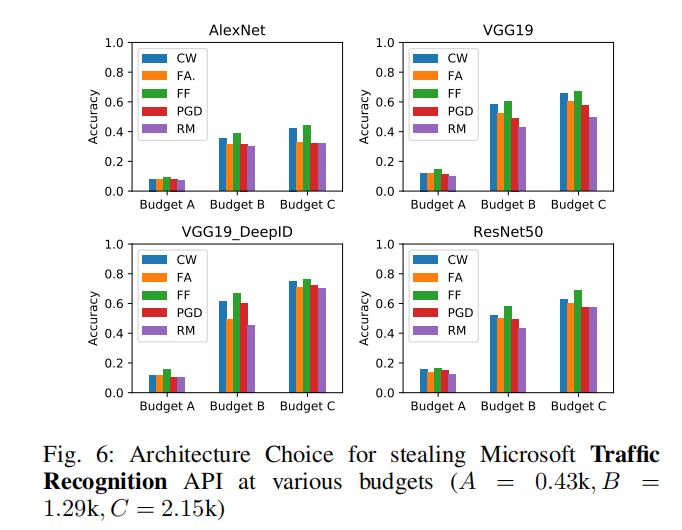


y：预测标签 w：受害模型的权重 k：x’的置信度

该方法的主要思想在于：从给出的unlabeled data里找出有最小置信度的作为informative data，这些数据作为合成数据集，训练substitute model。

大量的问询过于昂贵且会被检测出来，为解决此问题，本文设置两个目标：（1）用FeatureFool构建一个basic informative dataset Du(x)，其中的每个样本x有不同的分类置信度。（2）用margin-based adversarial AL产生informative dataset Dt(x)⊆Du(x)。实验表明，这样的对抗样本可以有效减少问询次数。

**Experimentation**

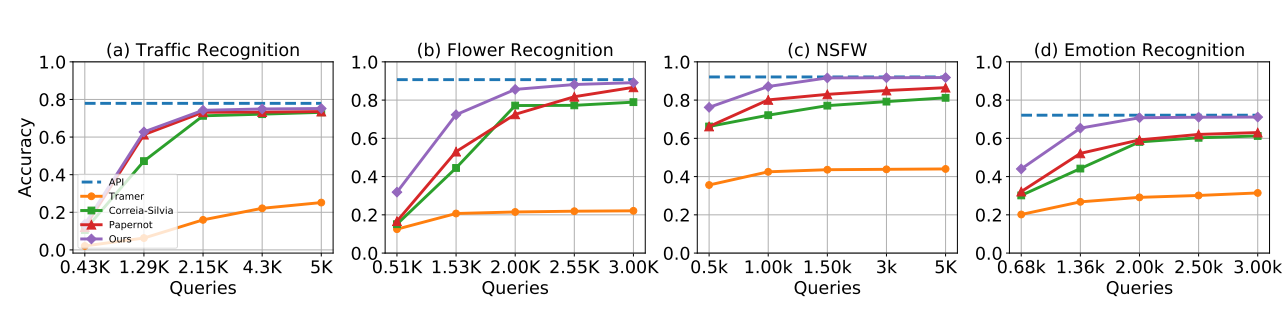


该图表示攻击的准确率。

（1）Substitute model用对抗样本（CW,FA,FF,PGD）训练比用随机样本（RM）效果好。

（2）Substitute model使用更复杂的网络和迁移框架，效果更好。（AlexNet网络比其他三个模型的表现要差）。

（3）FF（FeatureFool）准确率更高。



这张图对比了本文的攻击与其他攻击的效果。在这几类任务上，本文的攻击在窃取大规模模型时，保证了**高准确率**（纵轴）、**低问询**（横轴）和**低成本**（问询次数少自然成本低）。而其他方式至少在其中一个方面表现的不好。

**Limitations**

1. 改进对抗问询方式：FF产生的对抗样本中扰动较多，会造成合成数据集的污染，从而影响到精度，需设计更复杂的算法解决。
2. 扩展到多标签情况。
3. 扩展到其他领域：如音频、文本等。