

**NDSS 2020 — 美国佛罗里达大学**

**一、Abstract**

本文结合主动学习（Active Learning，即在query expensive的情况下，从大量unlabeled data中根据选择策略选取一部分最重要的data，只对这些data进行标签标注，而不是整个unlabeled dataset）、迁移学习（Transfer Learning，即Student模型从Teacher模型继承模型架构和参数）以及敌对攻击（Adversarial attack，即在原始图像中加入人眼不可察觉的扰动，使目标模型对该图像错误分类），对目标模型进行黑盒（Black-Box）的模型提取攻击（Model Extraction Attack，即敌手利用query目标模型得到的信息，在本地训练一个与目标模型性能几乎一样的替代模型）。最后，文章提出防御措施抵抗该攻击。

**二、Introduction**

先前的模型提取攻击有**两个明显的缺陷**：

1. 只针对小规模模型，对于大规模模型攻击效果不好。
2. 需要query的数量与目标模型的参数成正比，在窃取大规模的模型时要想使替代模型具有可靠的性能，就需要大量的query，成本较高。

针对上述两个缺陷，本文的**idea**为：

首先，将主动学习获得的敌对样本进行黑盒查询，生成输入-输出对。随后，利用迁移学习的方法在这些输入-输出对上retrain替代模型。

由于所得到的敌对样本大致位于目标分类器的决策边界（分类边界）上，攻击者在生成用于对替代模型进行再训练的合成数据集时，可以大大减少标记工作。

因此，本文**贡献**总结为：

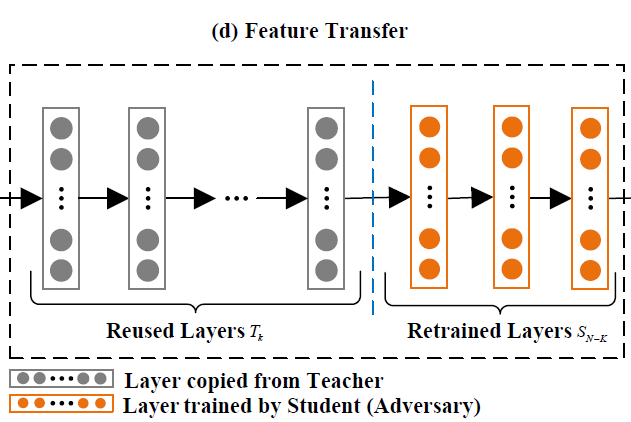
1. 用决策边界处的敌对样本对替代模型进行训练，极大减少了query的次数。
2. 通过对抗性的主动学习和从现有的训练有素的模型(如AlexNet)中迁移学习来加速模型提取速度。
3. 提出了防御方案。

**三、Background**

1）**主动学习（Active Learning）**

适用场景：欲对大量无标签数据进行分类，但没有任何可用的已标注数据训练分类器，且因成本有限，难以大量标注数据集以训练分类器。

核心思想：抽取信息量最大、不确定性高的样本进行标签标注，而不是从样本总体中随机取样。（注：“难分类的样本”即类与类之间的边界样本，即上文提到的决策边界上的样本。）

2）**迁移学习（Transfer Learning）**

许多用于各种任务的预训练模型（Teacher model）可供研究人员直接使用，如AlexNet、VGGNet、VGGFace和ResNet等。如图所示，敌手作为Student，从现存的Teacher model中继承前K层的模型架构和参数，然后再添加N-K层。在用敌对样本进行ratrain时，只训练后N-K层即可。

（注：具体可见0316\_USENIX\_SECURITY\_2018.docx）

3）对抗样本攻击（Adversarial Sample Attacks）。

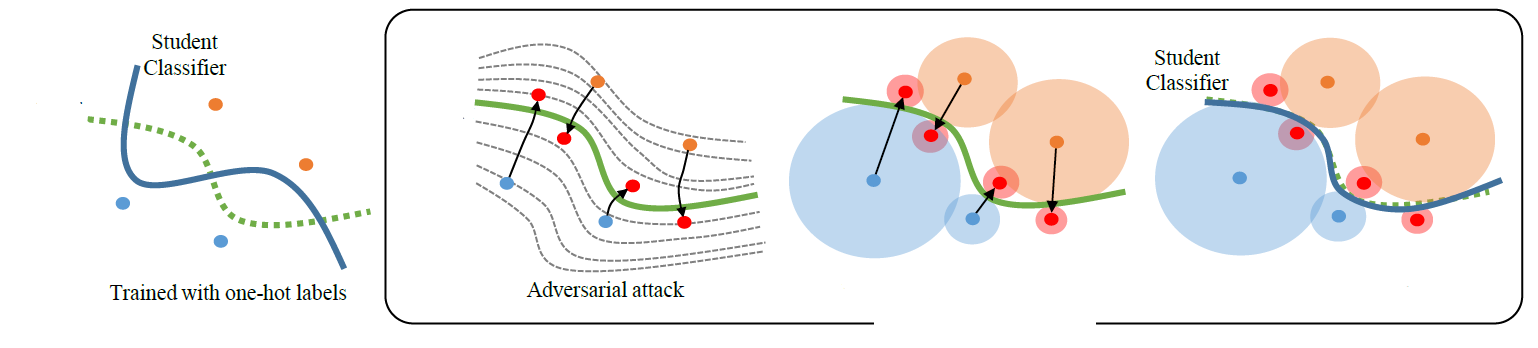
让模型对一个加上少量噪音的样本，进行错误的预测。模型对原始样本可以进行正确预测，并且引入噪音的样本不容易被察觉。

（注：具体可见0316\_USENIX\_SECURITY\_2018.docx）

**四、Method**

根据“二、Introduction”中讲述的idea可以看出，本文主要工作为：主动学习中如何构造敌对样本？看到这里，大家可能有一个疑问：为什么要将对抗样本作为主动学习中“信息量大”，也就是位于决策边界处的样本呢？下面给大家一一解答。

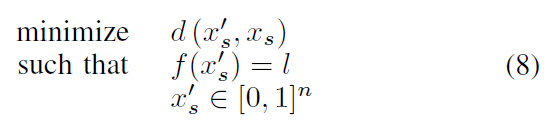
1）**为什么要将对抗样本作为决策边界处的样本？**



以二分类为例，如上图所示。设从左到右依次为图a、b、c、d。在图a中，如果我们用图中的四个样本点进行替代模型（student model）的训练，则训练后所得的决策边界很有可能是图中实线，与目标模型的决策边界（图中虚线）有非常大的差距，这样提取出来的模型是错误的。但是如果我们用图b、c、d中的红色样本进行替代模型的训练，便可很好的拟合目标模型的决策边界，从而达到模型提取的目的，而红色的点就是根据图a中的四个点找到的敌对样本。

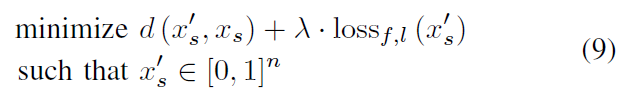
2）**如何构造敌对样本，使构造后的敌对样本刚好在决策边界处？**

本文提出了解决该问题的优化方法：

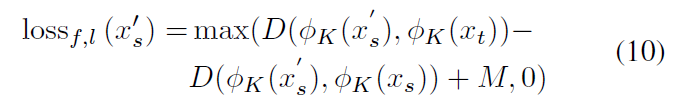


构造的敌对样本为，真实样本为，表示损失函数。该优化的思想是：最小化敌对样本与原始样本之间的差异，同时可将对抗样本误分类为。

因为解决有限制条件优化问题困难，因此本文利用penalty method（惩罚方法）进行求解，既是将条件转移到优化函数中：



其中，惩罚项为：



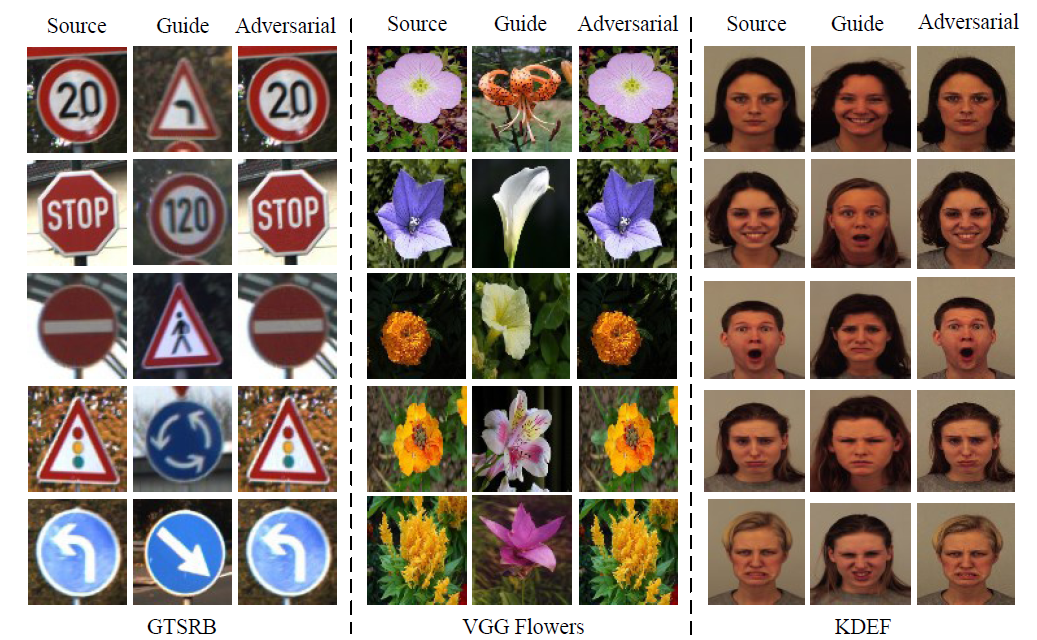
此处，惩罚项用“triplet loss (TL)”来表示。在TL中，表示标签为的样本，要想能被误分类为，应该满足：其特征与的特征尽可能相似，与原始样本之间的特征差异尽可能大。因此，小于0，M为与之间的最小间隔。

综上所述，如果一切正常，式（10）应该为0。

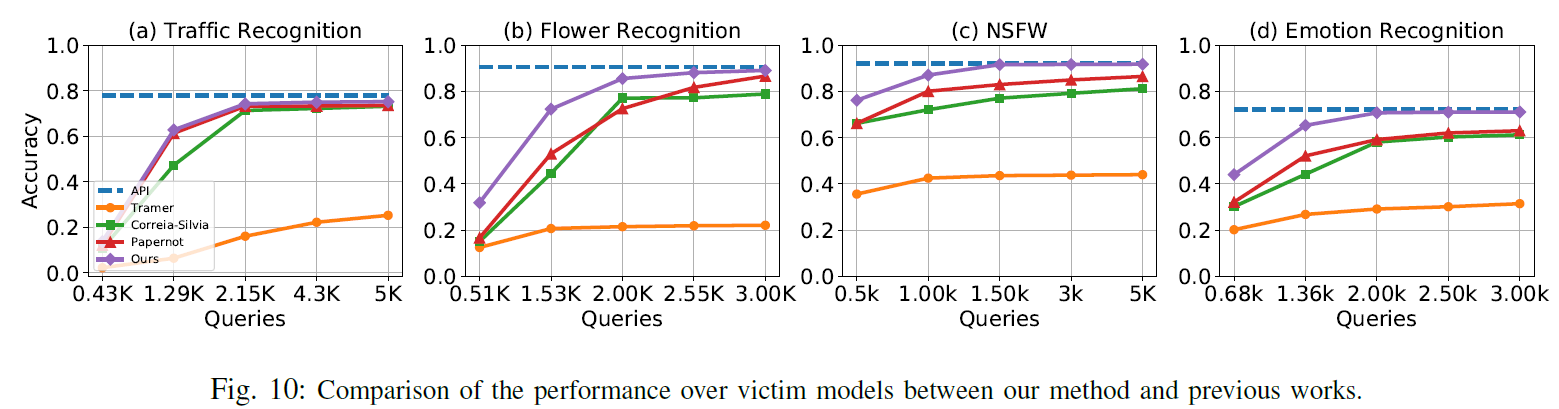
（注：关于TL，网上有详细解释，没看懂问度娘）。

3）找到敌对样本之后，将这些样本送入目标模型的API，得到预测结果。将这些输入-输出对作为训练数据，retrain本地替代模型即可。

**五、Experiment**



左侧第一列为原始图像，中间为误分类的目标类，右侧为生成的敌对样本。肉眼察觉不到原始图像与敌对样本的差别 。



将本文工作与先前的方案作对比。蓝色虚线为基准，紫色实线为本文方案，效果非常好。

**六、防御方法**

利用检测敌对样本的方法进行防御，分析敌对样本与原始样本的特征差异。