PREDICTION POISONING: TOWARDS DEFENSES AGAINST DNN MODEL STEALING ATTACKS

文章链接：<https://openreview.net/forum?id=SyevYxHtDB>

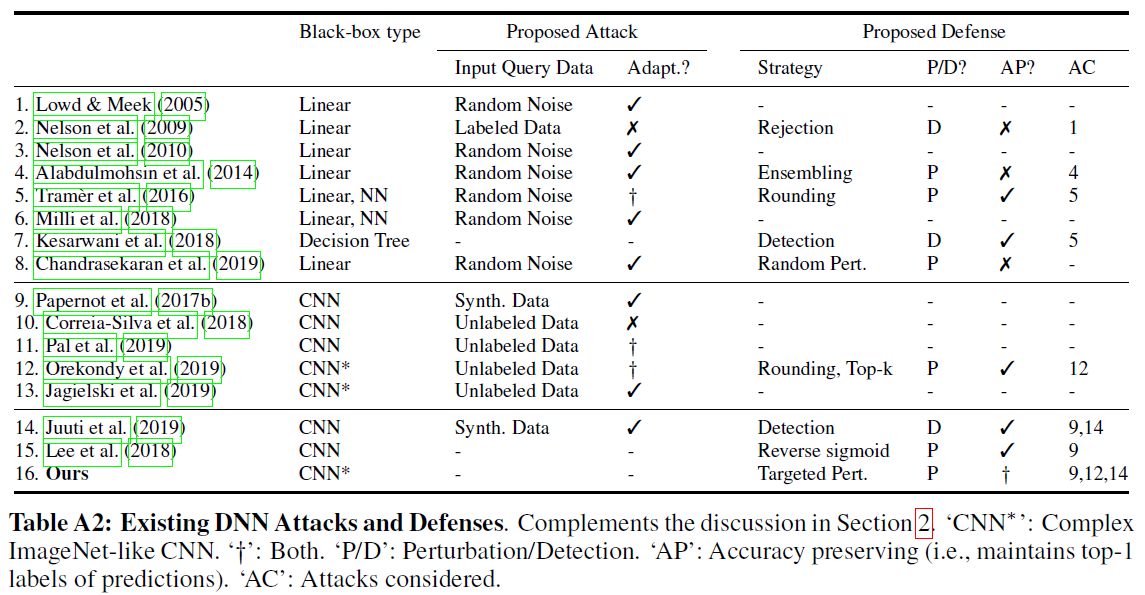
**ICLR文章公开审稿，链接里有审稿人的评论和论文作者的回复，相当精彩。**

**1 Abstract**

本文提出第一个针对model stealing attacks（敌手通过查询victim模型，来窃取模型的functionality）的主动防御。利用Poisoning attack的思想，对预测值进行投毒，从而降低敌手重建模型的精确度。

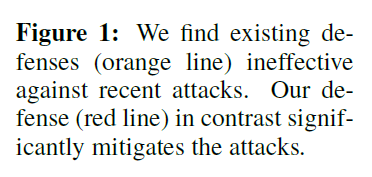
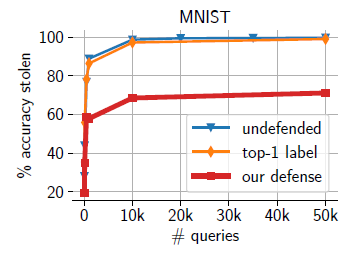
**2 Introduction**

现有的防御技术主要分为两大类，detect和perturbation。作者总结了现存model stealing attacks and defenses:



 从05年就有对ML的窃取攻击，近几年主要研究对复杂DNN模型的攻击。

现存的基于Peturbation的方法都保留了Top-1 label的可用性，作者发现其仍然会遭受窃取攻击：



最后，作者认为：**现存的防御方法都不能有效缓解攻击，而是通过增加攻击者对victim model的查询次数来延缓攻击**。

**4 Approach**

防御者的目标：

1. utility：被扰动的预测值仍然是可用的，即预测的label尽可能不变。
2. non-replicability：敌手不能利用query sample和victim model的预测值偷走模型。

Insight：

本文考虑，敌手收集query sample和预测值，再利用SGD训练本地模型。因此，为了防御窃取攻击，防御者首先扰动敌手query sample的预测值，从而使得由扰动后预测值计算的梯度，与真实预测值计算的梯度相比，梯度间夹角最大化，因此梯度偏差大使得训练损失高。

例如，敌手训练模型过程中，真实梯度和扰动梯度符号完全相反时，即夹角为180度，对模型的投毒攻击最成功。

Formal的方案如下：

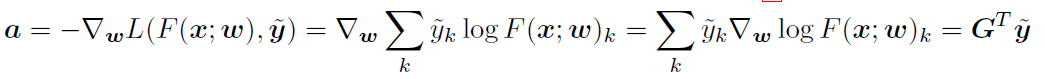


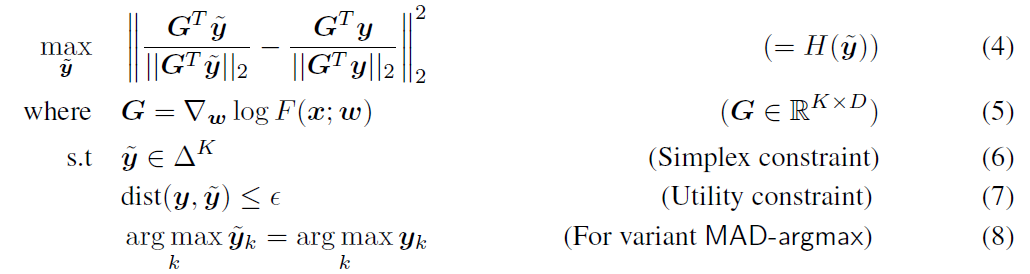
在敌手利用查询到的预测值训练模型时，u表示用真实预测值y计算的梯度，a表示用被扰动预测值计算的梯度。优化问题为最大话二者向量的夹角：



 因此，需要解决上述优化问题，即找到一个被扰动预测值，使得a和u的单位向量的夹角最大。

假设敌手利用cross entropy loss训练模型，则a可以重写为下式：

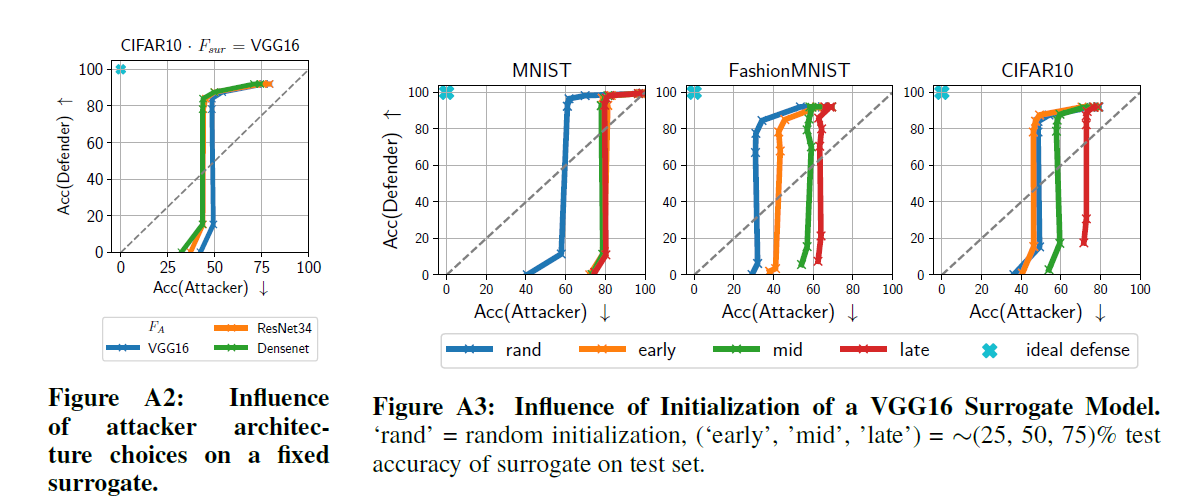
其中G表示Jacobian矩阵，即cross entropy loss对模型参数的导数。



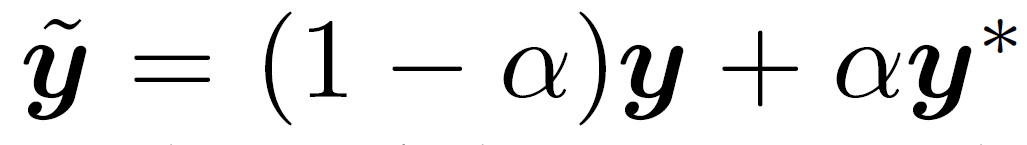
因此，优化问题转换为上述带限制的优化。优化函数（4）即为最大化a和u的单位向量的夹角；限制（6）表示被扰动的预测值需要保证所有元素的和为1；限制（7）表示添加的扰动要小于设定好的budget；限制（8）表示对预测值扰动前后top-1 label的不变性。

但是，解决上述优化问题仍然有两个挑战。

第一，因为对于防御者来说，敌手模型是未知的，因此优化函数中模型权重w未知。作者提出利用surrogate model，即本地模拟一个敌手model，估计模型权重W。通过实验发现模型的架构对防御能力无影响。但是，模型参数有很大影响，最好的参数是随机初始化的模型参数。实验结果如图所示：

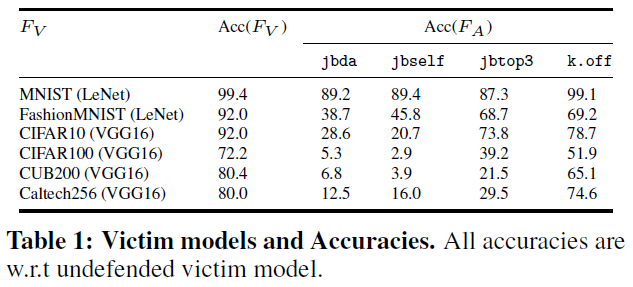


第二个挑战是，上述带限制的优化问题，所求结果通常为local minimal。为了解决这个挑战，作者提出Heuristic Solver，分两个步骤求解，第一步是寻找使得目标函数最大的one-hot预测值；第二步，利用如下公式求最终解：

。（这一步我也很疑惑，没理解求解的思想）

**5 Experimental results**

作者测试了不同模型下，victim model的精确度（F\_V），和被不同攻击方法（jbda、jbself、jbtop3、k. off）窃取的模型精确度（F\_A），如下：



 作者又测试所提防御方法的防御效果，如下图可以看出，大部分的窃取方法都能被防御，同时确保了高的可用性。（读懂此图需要看图下的描述）

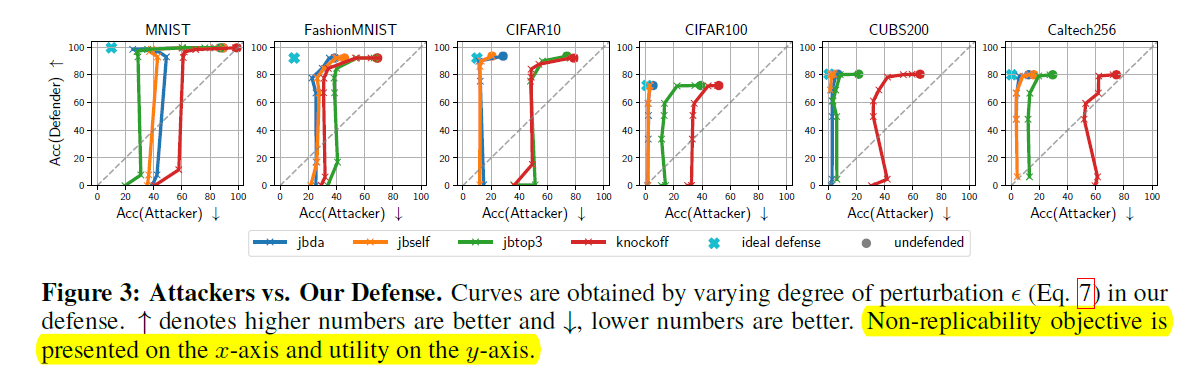
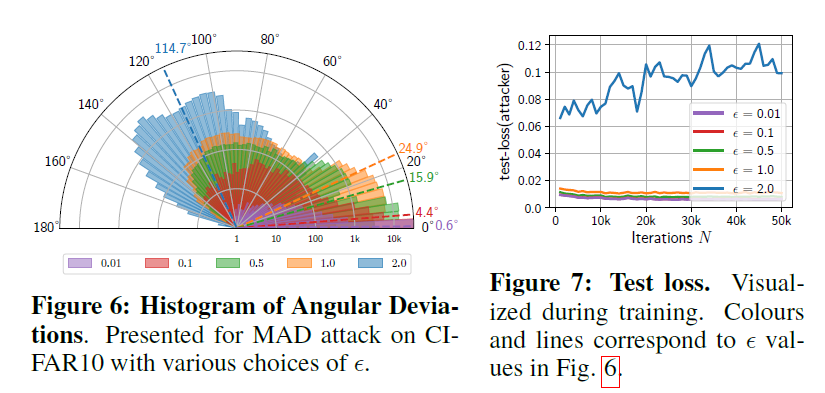


图6作者又分析了防御方法使得被扰动预测，与真实预测相比，所产生的梯度偏差。图7又分析了梯度偏差导致的攻击者模型的预测损失。如图可知，较大的budget导致大的梯度偏差，并导致大的预测损失。



**6 Conclusion**

这篇文章，将投毒攻击用在防御模型窃取攻击上，同样也有文章将对抗样本用在抵御成员推断攻击上。因此，AI方向攻击和防御是相互作用的，深入理解才能更好地应用。