**PRADA: Protecting against DNN Model Stealing Attacks**

**Code：**https://www.dropbox.com/s/7owpxde1ze0xkob/code.zip?dl=0

**1 Abstract**

本文提出新的模型stealing攻击（相比于两篇经典的16-USENIX和17-AsiaCCS），考虑了生成的查询样本和正则化方法对攻击效果的影响。攻击能力从预测精确度和可迁移对抗样本来考虑。并提出了generic的检测方法，PRADA，通过分析连续查询的分布进行检测。

**2 Framwork**

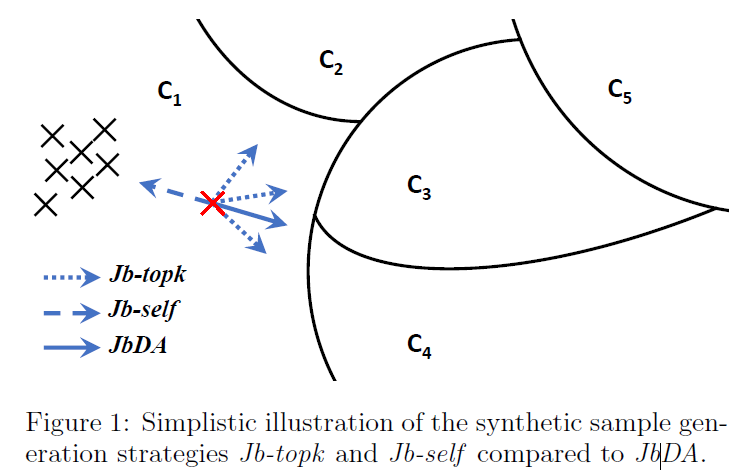
作者给出一个系统化的模型窃取框架：

1. 选择替代模型F ’架构和超参数。（替代模型架构对窃取能力的影响作者没有研究）
2. 初始化一个un-labelled种子样本集U（即，与目标模型训练集同分布的数据）。
3. 查询目标模型F获得带标签的训练集（U，F(U)）
4. 使用specific训练策略训练F ’
5. 根据种子样本U生成合成样本，来增加替代模型训练集，覆盖更多的输入空间。
6. 重复过程3-5，直到终止条件。

**3 Method**

作者提出的方法主要关注：步骤4中F ’的训练策略，和步骤5中合成样本生成方法。除此之外，还关注了种子样本集（本文考虑的是与目标模型同分布的数据集）的数量对攻击效果的影响。

首先描述步骤5中合成样本生成方法（本文提出Jb-topk和Jb-self）：JbDA是papernot17年提出的方法，是利用FGSM生成无目标的对抗样本来作为合成样本。三种方法的示意图如下所示：



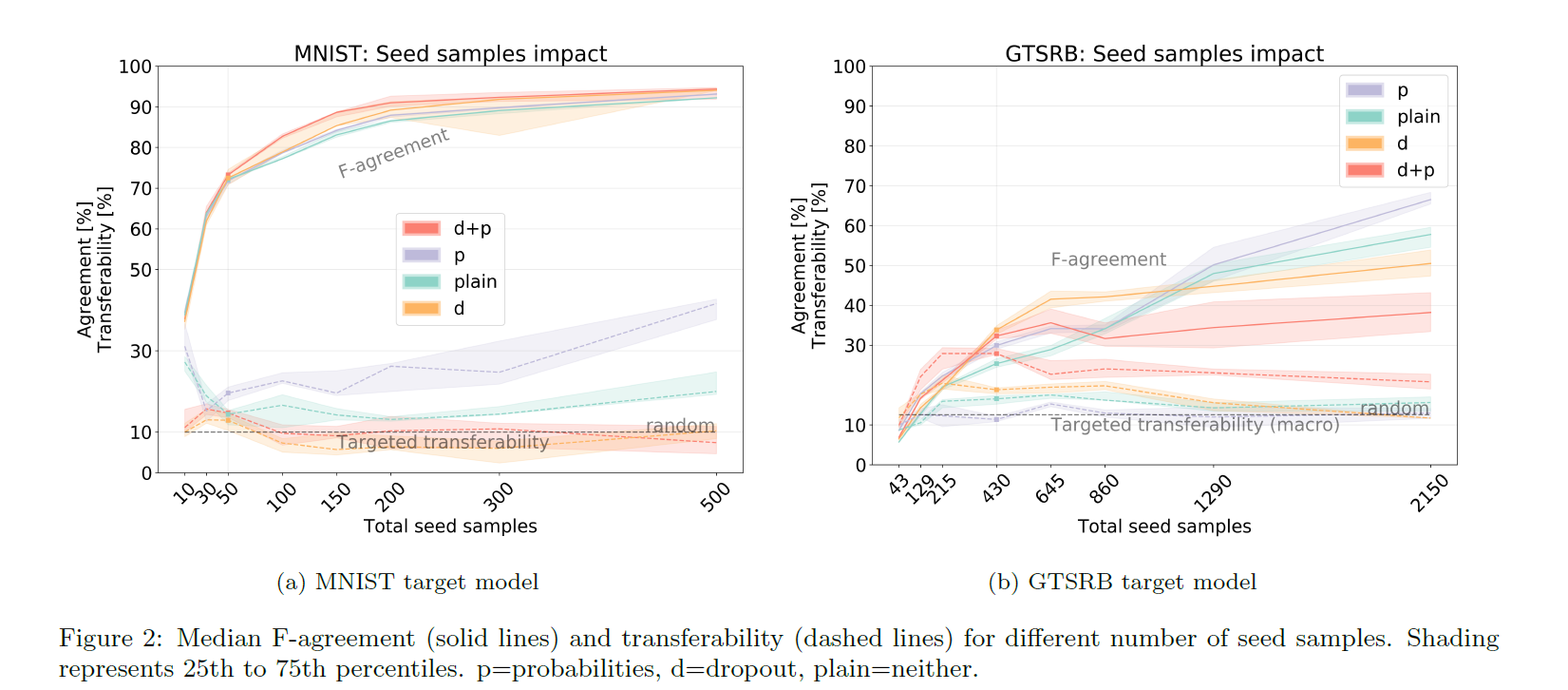
Jb-topk是指生成k个targeted对抗样本。方法是：用x查询替代模型F ’，对预测进行排序，除去最高的分类概率，挑选k个较高的类别来生成对抗样本。

Jb-self是指优化样本，使得更靠近最高的分类概率。方法是：用x查询替代模型F ’，进行样本优化朝着概率最高的分类。两种对抗样本生成都是利用PGD攻击方法。

步骤4中F ’的训练策略，本文首先考虑target模型返回结果对攻击影响，即label或者probabilities。然后又提出dropout的正则化方法对攻击影响。

实验结果：

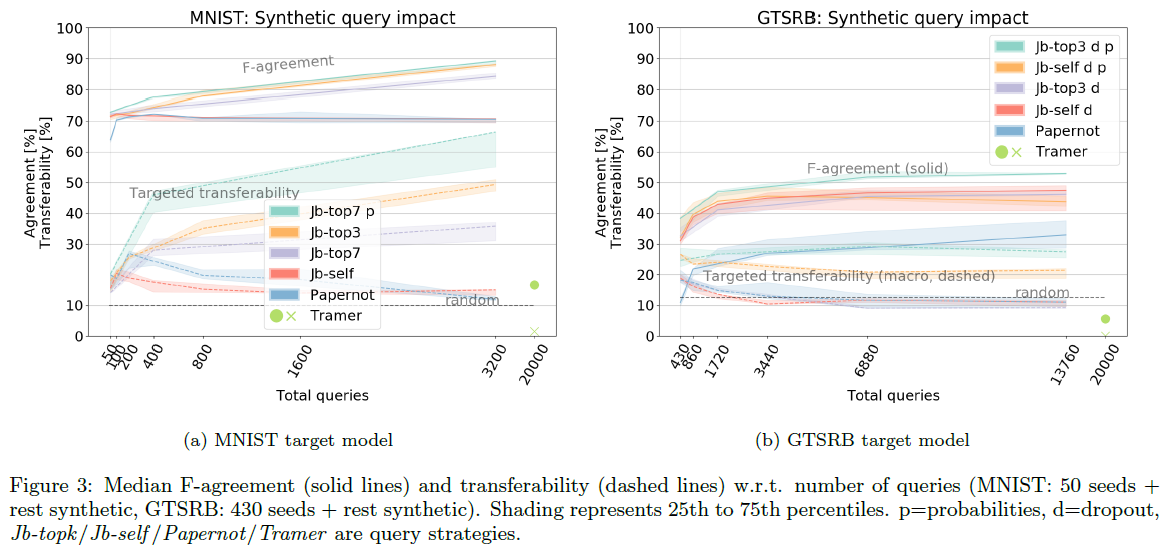
本文首先考虑种子数据集对攻击影响。不生成合成数据，也不查询目标模型。只是利用已有的种子数据集。值得注意的是，种子数据集与目标模型训练集同分布。例如MNIST，敌手种子数据集即MNIST数据集的一部分，并且与目标模型训练集不重合。



在四种情况下（d指dropout，p指probability，plain是没有d和p），分析模型训练结果。上图中实线为预测精确度（本文采用F-agreement，即F ‘ 和F预测结果的一致性），虚线为targeted对抗样本的迁移性。Dropout和probability都会带来较大的提升。

合成数据：

为了测试本文合成数据的效果，选取少量的种子样本，例如，只有50个MNIST数据。

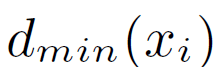


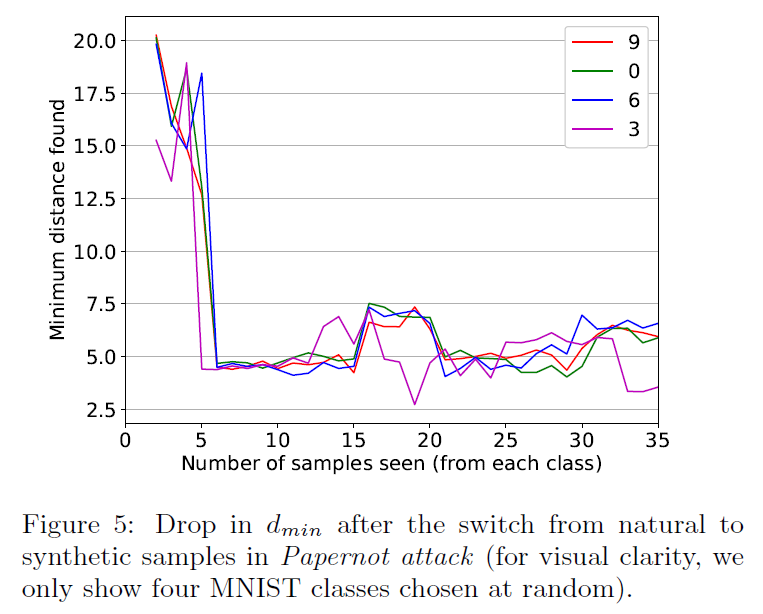
在MNIST数据集中，Jb-topk无论在精确度还是在目标对抗样本迁移性上都有好的表现。Jb-self则不好，也不知道作者提出来这个方法是干啥，单纯说明它不行吗？哈哈哈。

**4 Defense**

**PARDA：一个generic的防御方法，无论模型是什么类型，数据是什么格式。**

**核心思想是：敌手查询的样本，与常规的样本相比，具有不稳定的分布。**

具体方法是记录在结果为同一类别的样本中，新的查询样本，与之前的样本的最小距离。个人观点：方案的可行性在于：比如二分类0和1，一个常规的样本x0被目标模型分类为0，根据替代模型F ‘计算的对抗样本，仍然被目标模型分类为0，这两个样本最小距离很小。然后另外一个常规样本x1被目标模型分类为1，根据替代模型F ‘计算的对抗样本被目标模型分类为0，这时，这个对抗样本与x0的最小距离就很大。因此最小距离会发生异常的变化。



上图为MNIST中每个类计算的最小距离，因为每个类只有5个常规样本。图中5个样本之后都是合成样本，因此最小距离异常下降。因此可以检测根据对抗样本生成的查询样本。

**5 Conclusion**

个人观点：本文提出了新的合成样本生成方法，但是在种子样本很少的情况下，仍然不能产生很好的结果。例如MNIST上，50个种子样本，3000个生成样本只能达到接近90%的精确度。精确度有待提高。后续方法，如20-AAAI也提出了新的不相关数据集提高精确度。本文最大亮点是检测方法，PRADA，有效的防御了利用对抗样本生成增强数据的方法。但是对于常规样本仍然不能防御。

另外一点，模型窃取需要解决的问题就是利用少量带标签样本训练模型。解决这个问题，即可用到现有的active learning, semisupervised learning 等等方法。利用对抗样本的方法的局限在于，只能决定本地模型的边界，然后这个边界很大可能不是目标模型边界。如果可以寻找到目标模型边界，将会提高攻击精确度。另外，如果考虑一个经过对抗训练的目标模型，敌手想要窃取他的robustness和accuracy，是否可行，对数据集要求如何，以及对模型的训练方法要求如何，都是有待考究的，我也正在进行相关方法验证。