DaST: Data-free Substitute Training for Adversarial Attacks

文章链接：<https://arxiv.org/pdf/2003.12703.pdf>（Accepted by CVPR2020）

代码：https://github.com/zhoumingyi/DaST

**Abstract**

本文是第一个提出不需要真实数据来训练adversarial attacks的替代模型（substitute model）的文章。针对黑盒攻击，本文之前的替代攻击需要预训练模型来生成对抗样本，真实情景任务下很难获得预训练模型。因而本文提出 data-free 替换训练模型（DaST）来获取替换模型，不需要real data。

**Introduction**

1. Score-based攻击（2017年提出，仅依赖于预测分数（例如类别机率或对数）。 从概念上讲，这些攻击使用数值估算梯度的预测）和decision-based攻击（2018年提出，score攻击的变体）这类攻击无需预训练替代模型，但代价是需要对攻击模型进行大量的询问。
2. Gradient-based攻击则无需问询，需要预训练模型。
3. 模型窃取攻击同样无需预训练模型，但需要大量的真实数据来训练替代模型。

因而需要一个data-free的替代攻击。

**本文贡献**

1. 首次提出针对对抗攻击的无需真实数据的替代模型DaST
2. 在local深度模型和online机器学习模型测试了DaST，证明了目前机器学习模型受到攻击的风险更大
3. 在两个攻击场景进行了测试：probability-only和label-only，均有效。

**Method**

**Label-only**：攻击者可以获得目标模型的输出标签，但很难获取输入空间内的data，该场景下的方法为DaST-L。

**Probability-only**：攻击者只能访问目标模型的输出概率，该场景下的方法为DaST-P。

无需图片数据的替代模型训练原理：

1. 生产模型G可以生成替代模型D所需的训练data，（z是从目标模型的输入空间的随机取样）；
2. 利用获取目标模型T的输出；
3. 利用image-pair，训练替代模型。

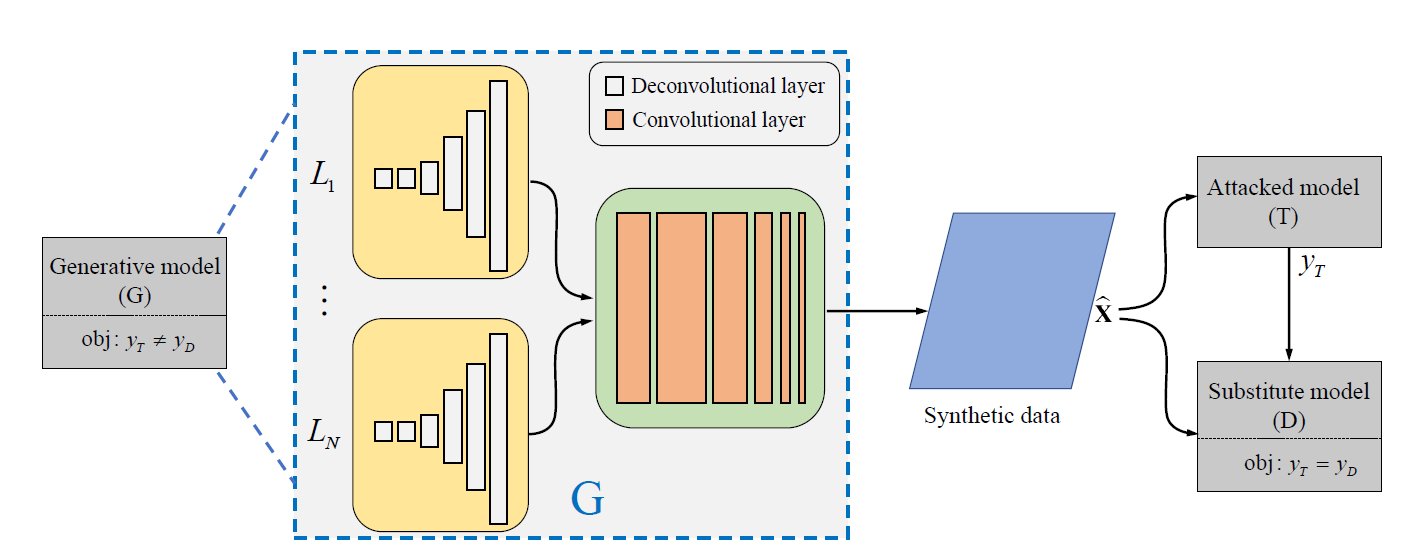
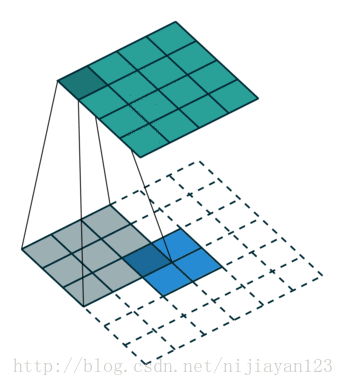


图1 DaST过程

训练过程如图1，这是一个两方模型，目标模型T是仲裁方。训练时，G随机生产出样本，保证T和D的样本标签是不同的，而替代模型D输出则要保证T和D的样本标签是相同的，使得D和T尽可能的接近。测试阶段，才能产生攻击所用的对抗样本。

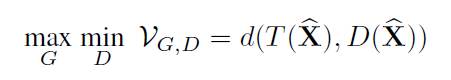
生产网络G中包含N个上采样反卷积组件，所有的上采样共享一个后处理卷积网络，那么G可以从T的输入空间随机取样噪音向量z，以及变量标签n，将z输入到n种类的上采样反卷积网络中，可还原出data =G(z,n)。

注：上采样反卷积：比于下采样，图片越来越小的操作，上采样操作使得整个图片越来越大的操作。



**替代模型复制目标模型信息，对抗训练详细过程为，：**

（1）以二元分类器分析，替代模型D的损失函数为：



测量T和D之间的输出距离，衡量差异性。

（2）在label-only场景中，距离函数可以写成：

C:\Users\LISHUANG\AppData\Local\Temp\1585893908(1).png

CE（.）为交叉熵损失，目标模型T的输出labels作为损失函数的标签，损失函数可以约束T和D之间的差异。（probability-only同理）

（3）为了保持训练的稳定性，将生产模型G的损失函数定义为：

C:\Users\LISHUANG\AppData\Local\Temp\1585894296(1).png

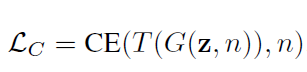
当且仅当时，才能获得全局最优的替代模型D，特别的时，。（即要使生成模型生产的随机样本，能让替代模型训练后，复制出目标模型的输出）

假定C:\Users\LISHUANG\AppData\Roaming\Tencent\Users\447253227\QQ\WinTemp\RichOle\V9T7DY5J)3JO@2IZVUC0M{4.png，R是T的输入空间，如果能D能实现，那么本文所设计的替代模型在无需T的梯度信息的情况下，实现和白盒攻击同样的成功率。

缺陷：在有限的时间内，很难保证D(.)=T(.)。如果无法约束生产模型G的输出空间，那么针对T的合成训练数据的空间会很小，分布不均，远达不到R，训练则没用。

**Label controllable Data Generation（解决G输出分布不均）：**

label controllable损失为：



因为目标模型T的梯度信息未知，且在训练过程中，相同输入情况下，替代模型D和目标模型T的输出会越来越接近，因而D替换T，作为标签控制，得到：

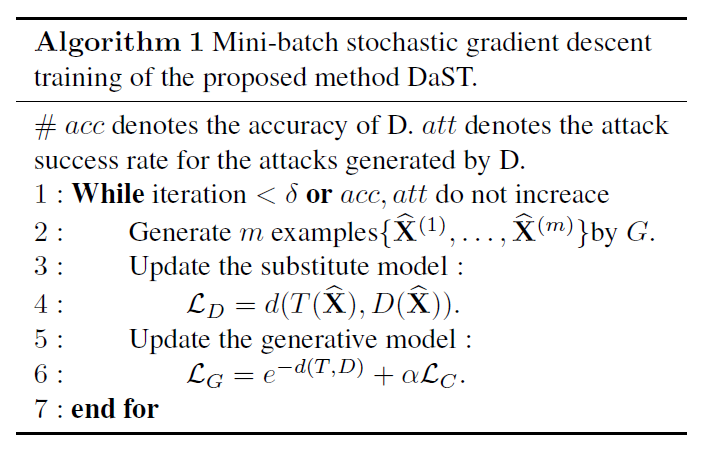
C:\Users\LISHUANG\AppData\Local\Temp\1585900484(1).png

因此，加上可控制标签损失后的模型G的损失函数为（LC是label controllable）：

C:\Users\LISHUANG\AppData\Local\Temp\1585900067(1).png

由此解决了模型G产生的样本不均匀的问题。

**DaST算法描述为：**



**Experiments**

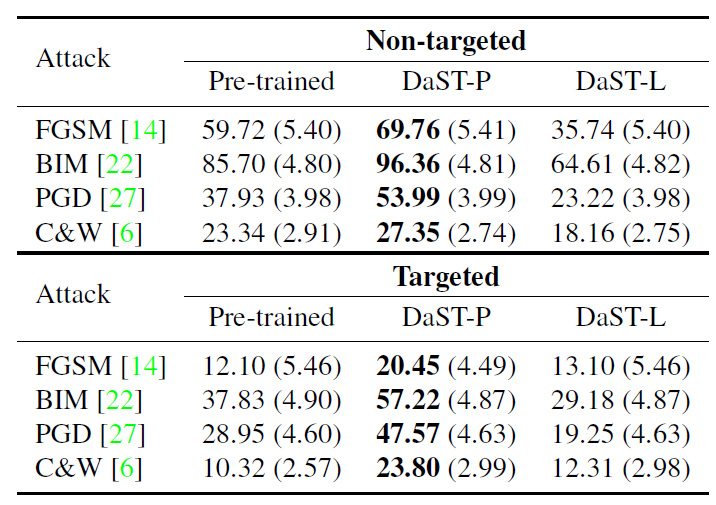
测试集：MINST和CIFAR-10，10k images

场景：label-only和probability-only，攻击者只能访问目标模型的输出

模型结构：

1. 替代网络不会下载任何预训练模型。在MINIST中，设计了3个网络，分别是small（3个卷积层），medium（4层），large（5层）
2. 对比：预训练medium网络和VGG-16

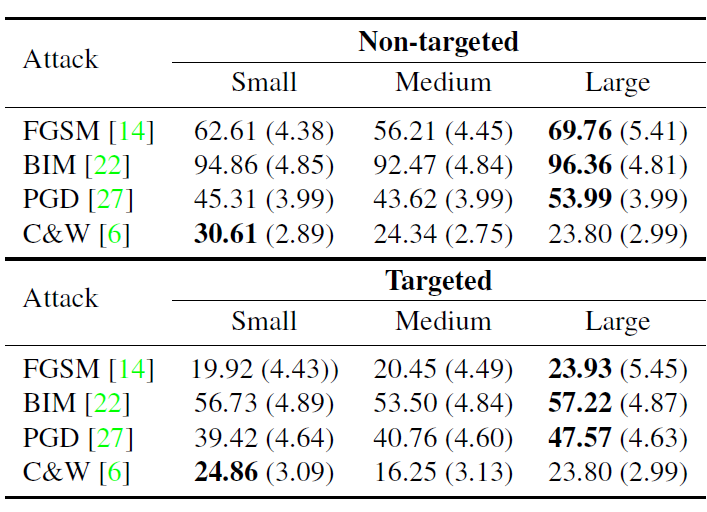
测试1. Performance of the proposed DaST on MNIST



在MINIST数据集进行测试表明，本文提出的DaST-P无论是在无目标攻击还是有目标攻击中，使用4种主流攻击方法的成功率都比预训练攻击的成功率高得多，甚至在基于DaST-L的攻击中，targeted attack的FGSM和C&W的效果也比预训练更好。

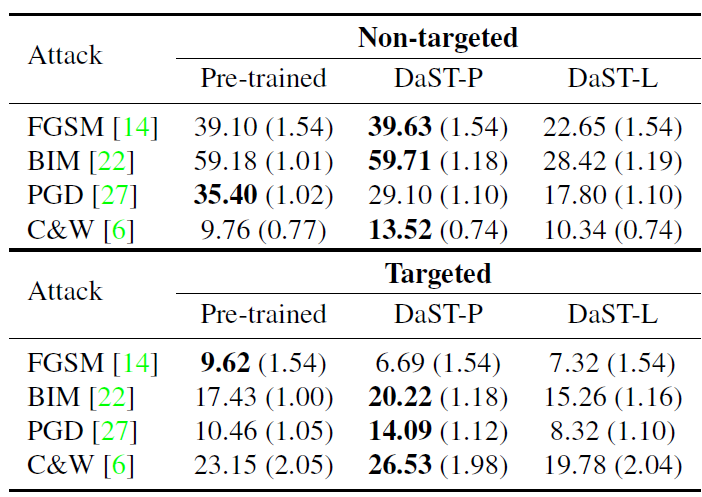
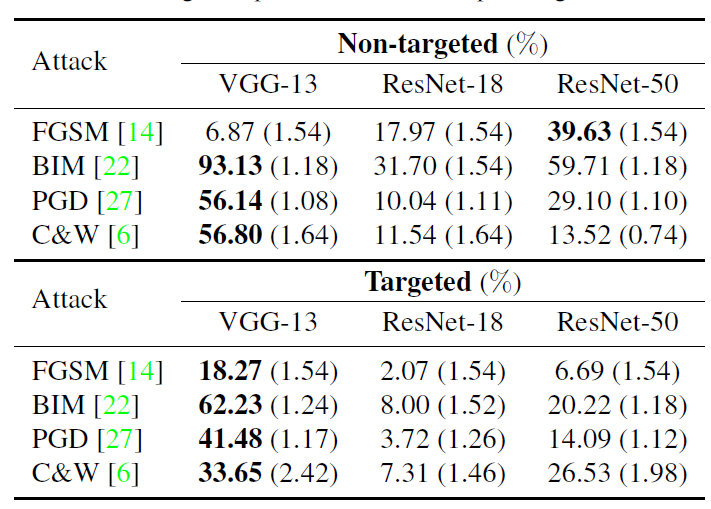
测试2. Performances of the proposed DaST with three different

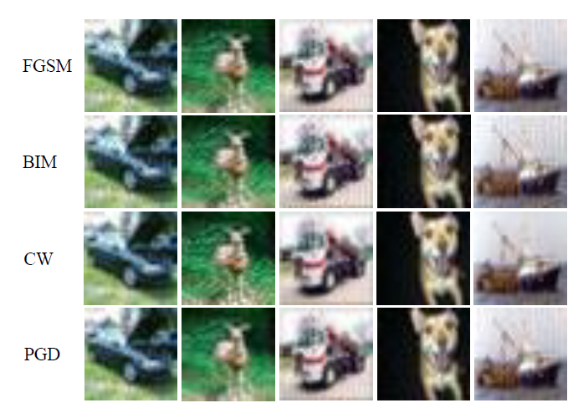
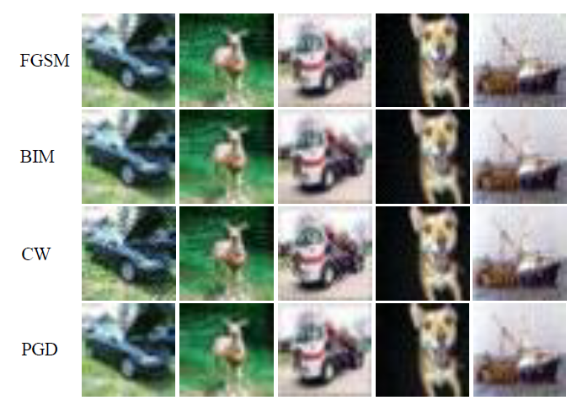
substitute architectures on MNIST



基于DaST-P，使用不同层数的卷积网络训练，在MINIST进行攻击测试的成功率，可以看到FGSM, BIM, 和PGD使用5层卷积层训练所得到的模型攻击成功率最高，而C&W则在3层最好。即**一般来说，结构越复杂的替代模型，对抗攻击性能更好**。

也在CIFAR-10进行了测试，效果也不错，在红框的表现略差



通过DaST-L训练后的攻击视觉效果 通过DaST-P训练后的攻击视觉效果

**Conclusion**

本文提出的DaST可以在无需真实数据下，无需问询，实现对抗攻击，

缺点在于该攻击方法无法单独进行对抗攻击，需要结合其他梯度攻击方法，且本文并未提出防御方法。