**Model-Agnostic Adversarial Detection by Random Perturbations**

**一、Abstract**

本文提出一个model-agnostic（模型不可知的）方法，分析模型在随机扰动下的响应，并在MNIST、CIFAR-10和ImageNet数据集上进行了评估。结果表明，本文检测方法对黑盒攻击和具有四种对抗性的强大CW攻击都是有效的。

**二、Introduction**

本文提出了一种简单而有效的对抗样本检测方法，它可以很容易地应用到所有现成的深度学习模型中。

本文想法的直觉来自于一个观察：对抗子空间的**决策边界**与在对抗方向上的合法数据的很相近。因此加上随机扰动扩大对抗空间后，有一定概率可以将对抗样本back为合法数据。

**Contributions：**

1. ·将置信度的相对变化作为对抗检测的判别特征。
2. ·本文的方法属于model-agnostic，也即无需知道模型的细节知识，适用于所有模型。
3. ·将随机扰动的tail bound与norm-bounded adversarial distortions联系起来，扩展了文中的理论。

**三、Related Work**

现有的很多攻击都是使用显式的模型信息构建的，一种自然的补救方法是隐藏这些知识或误导攻击者（如：模糊梯度）。

对抗样本具有可转移性的内在属性，给防御带来极大挑战。面对可转移的对抗样本，在不知道protected model的细节时，model-agnostic方式有很大优势。这种检测器通常建立在从输入样本统计中或输入变换的预测差异提取的特征上。

同时，面对自适应白盒攻击时，很多分类器和检测器无法识别出敌手构造的强大的对抗样本。

面对以上两种大多检测器不太“擅长”的点，文中对检测器评估包含了自适应攻击的treat model，并证明了检测器对攻击可转移性的鲁棒性。

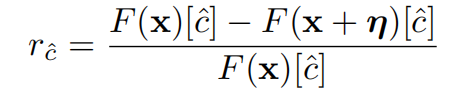
**四、Proposed Approach**

将对抗样本表示为xadv=x+*δ*（δ是一个微小常数）。

文中的创新点在于，**对输入样本加上一些适当大小的“干扰”信号，当这个输入是对抗样本时，将结果推回到原始分类的概率是无法忽略的（比较大），这会体现在label与相关分数的明显变化上。**因此，可以测量模型响应，并总结一些统计数据就可以区分出xadv和x。

**1. Main Steps**

输入x，模型F预测的类为c尖，从高斯分布*N*(0*,*diag(*σ*))中得到独立同分布的扰动η，测量c尖的相对分数差为：



考虑到信号的随机性，对过程重复进行m次处理，以从分布中得到更稳健的特征。

Step 1：训练一个二分类器作为检测器。其训练数据分别来自原始训练数据（标记为正常样本）与对正常样本使用特定攻击算法得到的样本数据（标记为对抗样本）。

Step 2：训练好的检测器也是一个概率生成器，输入样本，输出置信度，设置一个阈值，高的即为对抗样本。

Step 3：（也是文中添加的步骤，也叫noise augmentation，可以提高检测的性能）鼓励目标模型对**随机扰动（即良性噪声）**的正常样本与非扰动的样本归到同一类，实现方式是简单地将similar噪声扰动注入到模型的训练中。经过再训练的模型对有着良性噪声的正常输入有更强鲁棒性，也增大了对正常输入与对抗样本之间响应的差异。

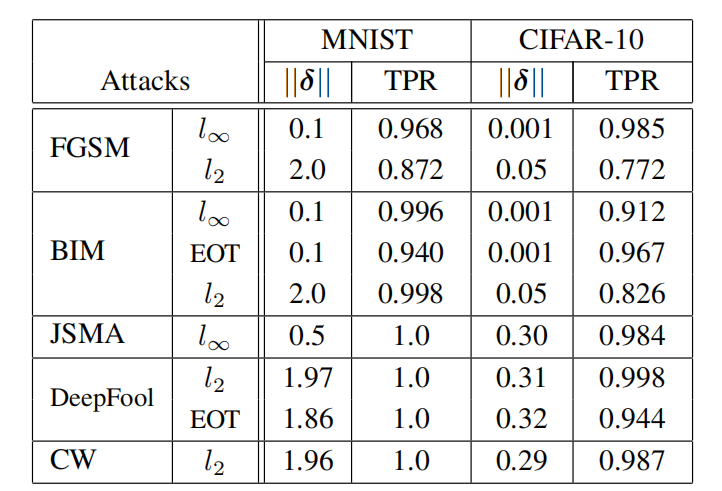
**2. Cross-Boundary Probability**

这部分先证明了在最坏情况下，文中方法具有可实施性，因此可以扩展到一般情况。这部分的分析只针对输入是对抗样本的情况，因此依赖于模型对随机噪声的鲁棒性，不过经验表明，大多模型都有较高的鲁棒性，进行额外的早生增强训练（Step 3）后，得到的模型鲁棒性更高。（这部分证明比较复杂，不放在这里，感兴趣可以在原文的第三页找到）。

**五、Evaluations**

**1.对抗检测的性能**

设置对抗样本为正类，自然图像为负类。TPR表示真实阳性率。下图为使用了文中方式（+step 3）后在各种攻击下的检测精度，可以看到精度整体都比较高：



**2. Transferred Adversary**

（a）本文（b）另一种检测方式SafetyNet，用列的攻击进行训练，用行的另一种攻击当做测试，得到了泛化能力，明显本文方式繁华能力更强：

