**论文的核心思想：**

1. 目前存在两种对测试样本的攻击模型：逃逸攻击和基于特征空间的攻击。逃逸攻击是指直接修改样本实体（例如修改恶意PDF文件）的攻击，而基于特征空间的攻击是指修改样本提取而来的特征向量。这两种方法所做的修改，都要满足一定的约束条件。逃逸攻击需要让修改后的样本的功能属性不变（例如恶意PDF样本还得是恶意），基于特征空间攻击需要让修改满足Lp范数小于某个值。基于特征空间的攻击是目前研究时间最久，数学理论较为完善的理论。而逃逸攻击更多是组合优化搜索出对抗样本。于是作者想探究：拿基于特征空间模型得到的对抗样本重新训练得到的模型，是否真的对基于逃逸攻击得到的攻击样本具有强鲁棒性。

实验结果发现：基于特征空间得到的对抗样本重新训练得到的模型，其鲁棒性有限。

这里面的根源在于，基于特征空间在寻找对抗样本的过程中，是使用Lp范数来维持样本的功能属性。然而Lp范数是不能直接把这种维持样本功能属性的约束显式的表达出来。于是作者对这个Lp范数约束做了扩展：提出Conserved Feature的概念，这些特征一修改就会改变样本的功能属性，于是在寻找对抗样本的过程中显式让这些Conserved Feature不变。🡺如此可以寻找更多的逃逸对抗样本。实验结果发现加上这个约束后，基于特征空间的对抗模型的鲁棒性得到了很大的提升。

1. **Feature space model是迁移性最好的模型。**

**feature-space models of attacks** : 攻击者的特征空间模型。假设攻击者是在满足一定的Lp范数约束条件下直接修改特征向量的数值。 这种模型具有一些缺点：

1. 在实际问题中，并不能任意的修改特征向量的值。例如，在修改PDF的过程中，不能减小PDF的size
2. 在实际问题中，Lp范数的语义性不好。例如在对一个恶意PDF进行修改。Lp范数要求对这个PDF修改要越小越好。但实际上对PDF修改的多少，与其是否仍然是恶意的没有直接的关系。对恶意样本添加大量的nop，会导致Lp很大，但是PDF的恶意性并未受到影响。

**evadeML:** 逃逸机器学习，<https://evademl.org/docs/evademl.pdf> 是指通过对恶意的PDF的PDF主体进行修改，让修改后的PDF既保持了恶意行为，又能够让分类器分类错误。论文使用遗传算法寻找恶意样本。变异PDF的body部分：PDF的body是树状的，便于进行修改。通过删除恶意文件body中的某个object、从良性文件中选取一个object插入恶意文件中和从良性文件中选取一个object替换恶意文件中的某object这3个操作随机地修改恶意文件。 🡺这种操作是直接修改PDF主体的，而不是修改特征向量的数值。

**Conserved feature :** 那些不破坏恶意性就不能修改的属性。这种特征十分重要，但是这类特征不能通过传统的统计方法获得。

Realizable evasion attack : 直接修改样本实体，从而使得样本的功能属性不变、同时误导分类器

Feature space models evasion attack: 修改样本的特征向量，然后使用l2范数来保证样本的功能属性不变。