Evading Classifiers by Morphing in the Dark

Curtis Hung Dang

National University of Singapore

[hungdang@comp.nus.edu.sg](mailto:hungdang@comp.nus.edu.sg)

**中文标题：**

滥用恶意软件检测器中的PDF解析器

## Questions:

## 一句话概述：

## 中文摘要：

通过对抗数据操作，基于学习的系统已被证明容易逃避。这些攻击的研究是基于以下假设：对手对目标模型内部结构，训练数据集或至少分配给输入样本的分类评分有一定的了解。在本文中，我们研究了一个更加严格和逼真的攻击场景，其中目标分类器最低限度地暴露给敌手，只显示其最终分类决策（例如拒绝或接受输入样本）。而且，攻击者只能使用黑箱变形器来操纵恶意样本。也就是说，敌人必须通过在“黑暗中”对恶意样本进行变形来逃避目标分类器。我们提供了一个评分机制，可以根据可用的有限信息为每个样本指定一个真实值评分，该评分反映了每个样本的逃避进度。利用这种评分机制，我们提出了一种逃避方法 - EvadeHC？并针对两种PDF恶意软件检测器（即PDFRate和Hidost）进行评估。实验评估表明，所提议的规避行为是有效的，对评估数据集的规避率为100％。有趣的是，EvadeHC优于已知的基于分类器输出的分类评分运行的分类器规避技术。尽管我们的评估是针对PDF恶意软件分类器进行的，但所提出的方法是不可知的，并且更广泛地应用于其他基于学习的系统。

## 亮点解读

## 论文总结

我们提出了一个限制对手逃避分类器的模型，只有黑匣子访问。这种限制性设置允许我们调查利用有限领域知识和分类器特定信息（例如模型内部结构和训练数据集）的有限信息的通用躲避攻击。一方面，这种模式为后卫提供了对手能力的基准。另一方面，攻击者可以通过将领域知识纳入泛型攻击来获得更有效的躲避攻击。

我们已经描述了EvadeHC，这是一种逃避二进制结果检测器的通用爬山方法。我们的技术假定对检测器和数据处理机制都知之甚少。针对两种PDF恶意软件分类器进行的实验研究已证明了所提出方法的有效性。虽然实验是在恶意软件分类器上执行的，但所提出的技术及其安全性影响可能也会更广泛地应用于其他基于学习的系统。