智源学者成果展示——智源青年科学家

作者　陈恺（中国科学院信息工程研究所）

2020年6月

项目负责人：陈恺

单位 / 团队：中国科学院信息工程研究所

成果名称：

Seeing isn't Believing: Towards More Robust Adversarial Attack Against Real World Object Detectors

眼见为虚：真实世界目标检测系统的高鲁棒对抗攻击

成果简介：

近年来，对抗攻击是 AI 安全研究领域内的研究热点。其中，对目标检测系统的对抗攻击严重威胁自动驾驶安全，如对抗攻击可能让汽车无法检测到 STOP 交通标识牌而导致事故。但近年来的对抗攻击研究始终局限于数字领域。在真实世界中，我们很难直接将数字世界中的对抗样本打印出并成功攻击。因此一直以来，数字领域的对抗攻击对真实世界难以构成威胁。直到 2018 年，美国加州大学伯克利分校研究人员提出了针对目标检测系统的物理对抗攻击。但其攻击距离十分受限 , 且未能验证多角度及不同环境下的攻击效果。因此，为解决以上问题，我们提出了高鲁棒的物理对抗攻击技术，使得攻击距离达到 25 米，攻击角度达 120° ，在不同光照及背景环境中，即使车辆以 30km/h 的速度动态行驶，也能保持较高攻击成功率。相关论文发表在 CCS 2019（CCF-A）。

文章链接：https://dl.acm.org/doi/10.1145/3319535.3354259

Demo 链接：http://kaichen.org/datas/CCS_2019_HA_IIE.mp4

项目负责人：陈恺

单位 / 团队：中国科学院信息工程研究所

成果名称：

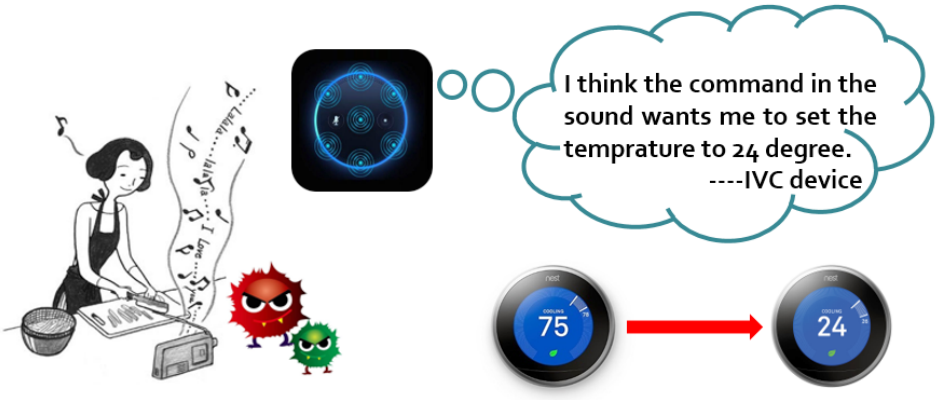
“恶魔音乐”实际对抗攻击商业化智能语音识别系统

Devil’s Music: A General Approach for Physical Adversarial Attacks against Commercial Speech Recognition Systems

成果简介：

随着大数据和深度学习算法的快速发展，语音交互被广泛应用于自动驾驶、医疗服务和教育等领域中。智能音箱和语音助手不仅连接了社交网络，也将众多智能家居设备联系起来。然而，由于机器识别和人类识别存在一定的差异，使得攻击者可以将语音命令嵌入到一段音频中形成对抗样本，在人们无法觉察的情况下恶意地控制语音识别系统，进而操控导航、智能家居设备，甚至获取用户的隐私信息。

由于商业化系统拥有庞大的数据集和复杂的深度神经网络结构，分析人和机器的差异原因并找到攻击点是构造对抗样本的难题。此外，如何在真实场景的攻击中克服播放、录音设备的电子噪声以及环境噪声的干扰，规模化地攻击语音识别系统是又一难点。本项研究通过分析语音识别深度学习算法的脆弱性，自动化地将命令嵌入到音乐中生成对抗样本——“恶魔音乐”，进一步模拟实际环境对样本的干扰，提高样本的鲁棒性，并通过网络和收音机信号广泛传播。“恶魔音乐”能够成功攻击 Amazon、Google、Microsoft 和 IBM 的语音转文本服务，并攻击语音助手（Google Assistant 和 Microsoft Cortana）以及智能音箱（Amazon Echo 和 Google Home），实现了在用户无法觉察的情况下通过“恶魔音乐”发送导航或拨打电话等命令。成果被发表于 USENIX Security 2018 和 USENIX Security 2020（CCF-A 类）。



图：攻击智能语音识别设备场景

文章链接：

http://kaichen.org/paper/conference/sec20summer_chen-yuxuan_prepub.pdf

http://kaichen.org/paper/conference/sec18-final449.pdf

Demo 链接：

https://sites.google.com/view/devil-whisper

https://sites.google.com/view/commandersong/

项目负责人：陈恺

单位 / 团队：中国科学院信息工程研究所

成果名称：

Understanding the Behavior of Skills in Large Scale，对大规模技能的行为理解

成果简介：

近年来，虚拟个人助理（VPA）在智能音响等物联网设备上得到了广泛应用。除了其内置功能，服务商鼓励第三方开发人员扩展 VPA 的新技能（亚马逊称之为 skill，谷歌称之为 action）以丰富产品功能。然而随着第三方技能的快速发展，一些具有潜在风险的技能也随之出现。据我们所知，目前还没有针对技能交互内容的系统性研究，主要因为存在以下两个难点：（1）技能完全处于黑盒状态，研究者无法得到相应的代码；（2）技能的输出内容（问题）是自然语言形式，对其理解并生成规范合理的输入内容（答案）难度很大。

我们尝试了首个关于技能行为的系统性研究，实现了针对技能的“聊天机器人”，包括自动化提取初始交互输入语句、理解问题和生成相应答案。本项研究构建了一个智能交互系统，并对技能行为进行了大规模的分析，样本总量达到 30801 个（28904 个来自于美国亚马逊市场，1897 个来自于谷歌市场）。如此大规模的分析使我们能更深入地理解技能及其开发人员的各种行为。研究结果发现，1141 个技能要求用户提供手机号码、姓名和地址等隐私信息而没有遵循开发者规范，例如未在隐私策略中声明这些信息或没有按照规定配置相应权限等。此外，我们还发现 68 个技能在用户发送停止命令后会继续窃听他们的谈话内容。相关论文发表在 USENIX Security 2020（CCF-A）。

文章链接：http://kaichen.org/paper/conference/sec20summer-final678.pdf

Beijing Academy of Artificial Intelligence

微信关注
北京智源人工智能研究院