GeekPwn 极棒

人工智能的安全问题

人攻智能 洞见未来

GEEKPWN 2018 · 国际安全极客大赛

GeekPwn – 白帽黑客的舞台

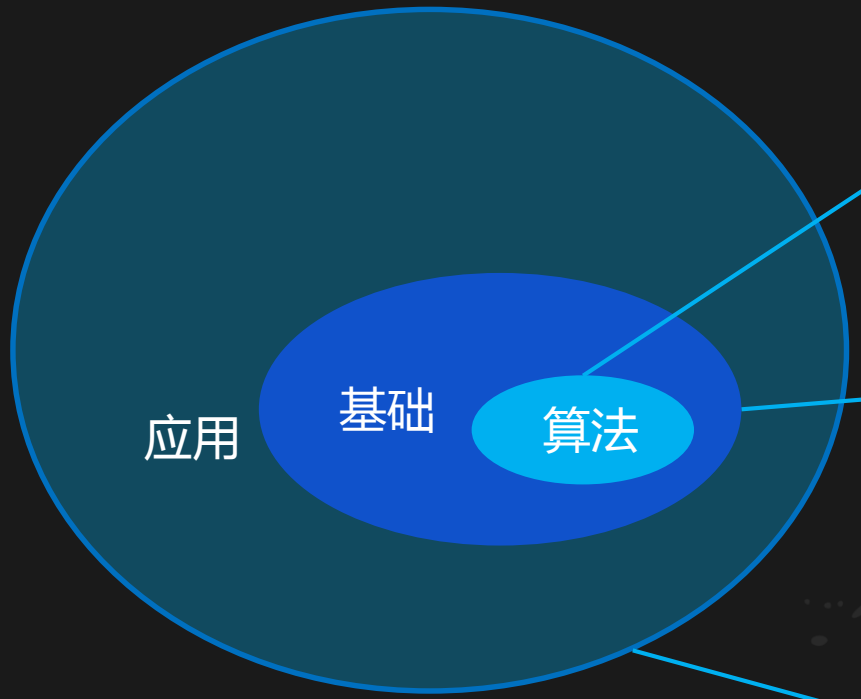
KEEN – 亚洲第一个Pwn2Own上攻破iPhone的世界级黑客大赛冠军团队

GeekPwn – KEEN公司主办的全球首个智能领域黑客大赛，自2014年起已举办8届。攻破智能设备和应用数百个，帮助厂商修复数百严重安全漏洞，为一百多位选手发放数百万奖金。攻破设备有汽车，POS机，摄像头，路由器，手机，平板电脑，智能家居，网络基础协议，智能插座，无人机……

人类对AI的恐惧

- 对人造智能的恐惧是人的天性
- 悲观派与乐观派
- 新技术出现必然会经过的阶段
- 人类需要做什么准备
- AI Safety 和 AI Security

AI安全的层次性



对抗样本
污染学习数据

缓冲区溢出
整型数溢出
运算精度降低
通讯未加密

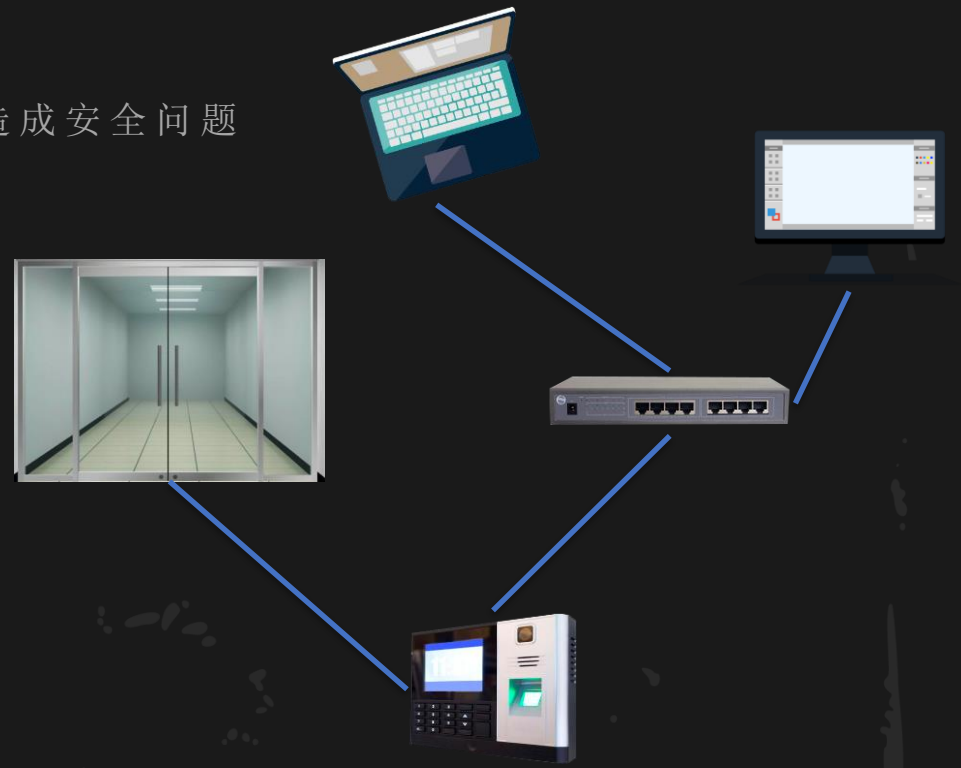
操作系统漏洞
应用层逻辑漏洞

错误的决策，判断

系统被控制

AI应用系统的安全性

- AI应用系统是一个整体，任何环节出现漏洞都可能造成安全问题
- 传统的攻击方法 - 利用系统漏洞控制门禁
- 最终效果，任意人脸可顺利通过门禁



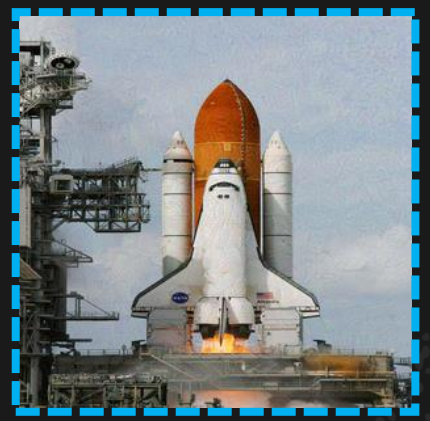
AI基础系统的安全性

- 库 OpenCV, Malheur, Scikit-Learn ……
- 框架 TensorFlow, TorchNet, Caffe ……
- 云服务 Google, Amazon, Baidu ……

AI算法的安全性

- 2016.10.24 极棒硅谷站
- Ian Goodfellow
- 生成式对抗网络（GAN）的发明人
- 主题演讲：Physical Adversarial Examples

蘑菇



蘑菇

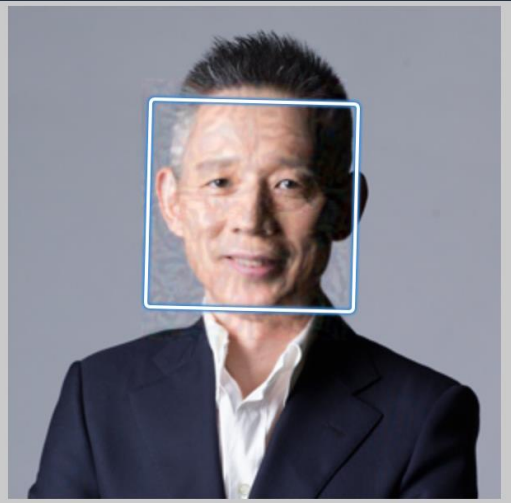
航天飞机

北极狐

https://ap-northeast-1.console.aws.amazon.com/rekognition/home?region=ap-northeast-1#/celebrity-detection

服务 资源组 Connor 东京 支

ekognition



选择示例图像

使用您自己的图像

图片必须是.jpeg 或 .png 格式，不得大于5MB。没有存储您的图像。

上传 或者拖放

使用图像 URL 前往

结果

Arnold Schwarzenegger

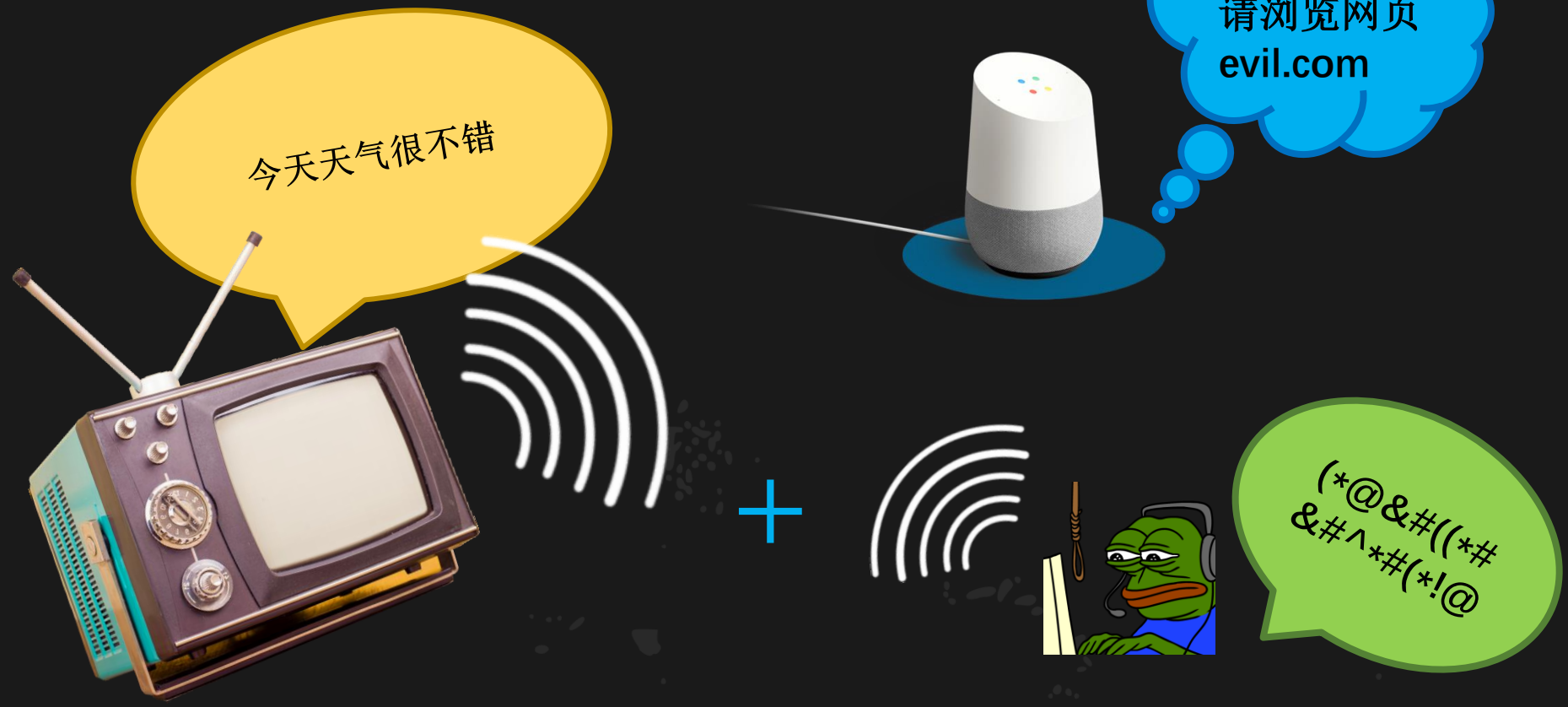
了解详情

匹配置信度 75

请求

响应


```
{
   "CelebrityFaces": [
      {
         "Urls": [
            "www.imdb.com/name/nm0000216"
         ],
         "Name": "Arnold Schwarzenegger",
         "Id": "3AI8Ss5",
         "Face": {
            "BoundingBox": {
               "Width": 0.42809364199638367,
               "Height": 0.42809364199638367,
               "Left": 0.28093644976615906,
               "Top": 0.19063545763492584
            },
            "Confidence": 99.99835968017578,
            "Landmarks": [
               {
                  "Type": "eyeLeft",
                  "X": 0.42850828170776367,
                  "Y": 0.3572295904159546
               },
               {
                  "Type": "eyeRight",
                  "X": 0.5558915734291077,
                  "Y": 0.3606841564178467
               },
               {
```


今天天气很不错
ok google
请浏览网页
evil.com
(*@&#((*#
&#^*#(*!@

对抗样本的危害

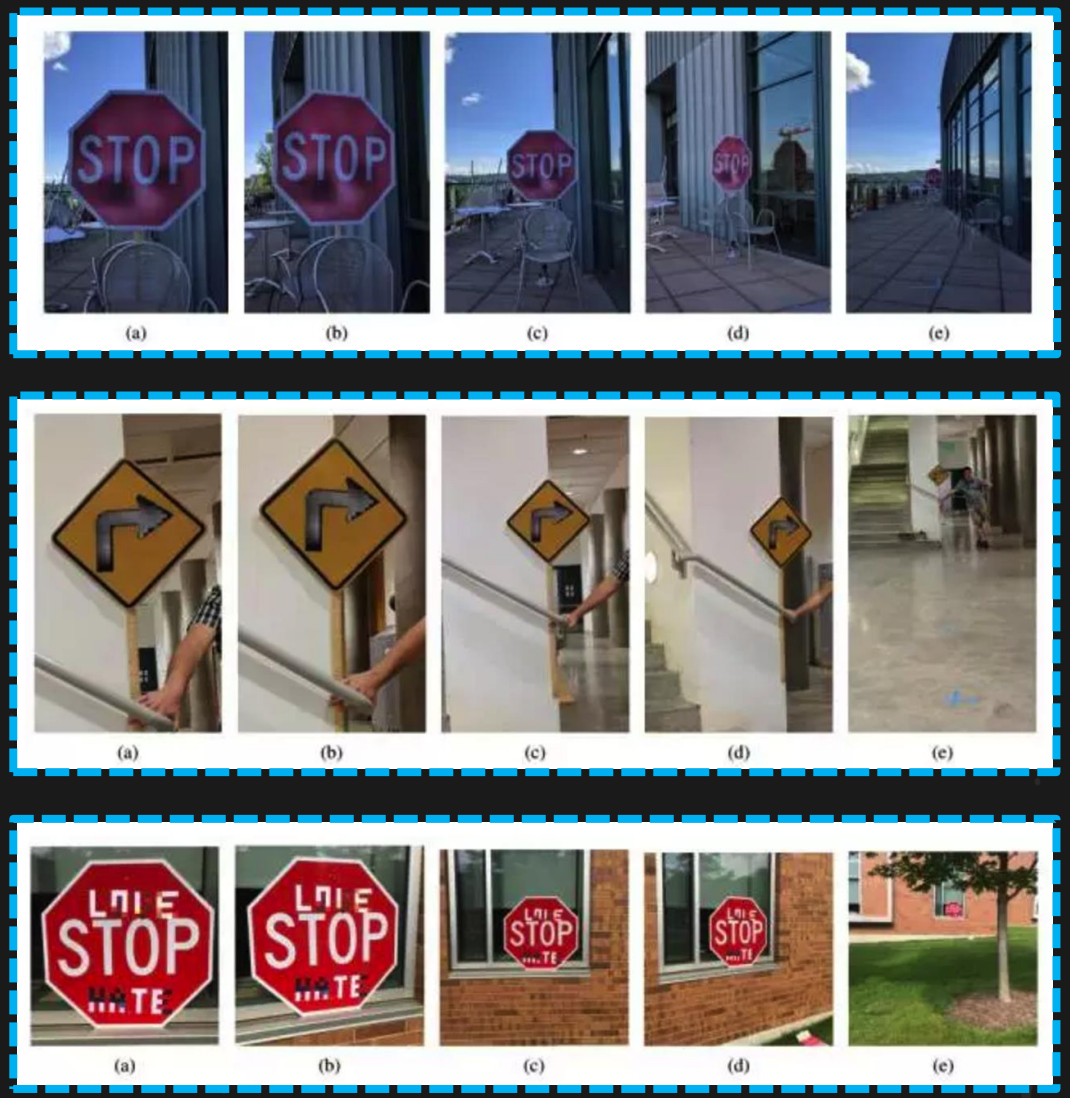
干扰人脸识别系统的正确识别

干扰恶意软件的识别

干扰自动鉴黄系统的识别

干扰自动驾驶系统识别交通标识

…



对抗样本威胁模型

- 非定向攻击 Non-Targeted
- 定向攻击 Targeted

- 白盒 White box
- 黑盒无探查 Black box without probing
- 黑盒有探查 Black box with probing

- 数字攻击 Digital attack
- 物理攻击 Physical attack

- 允许的扰动范围 Perturbation
- 攻击目标单一或泛化

安全极客与AI安全

- AI安全是一个大课题
- 安全极客在AI安全领域的作用
- 极棒对AI安全的愿景

极棒对AI安全的支持

- AI安全问题的知识普及
- 鼓励AI安全相关的攻破项目在极棒大赛的展示（人脸识别，笔迹模仿…）
- 主办多项AI安全相关的专项赛事（CAAD线上赛，CAAD CTF，数据追踪挑战赛，GAN掉马赛克，仿声验声…）
- 奖金累计达200万人民币

CAAD 对抗攻击与防御的专项赛事

- Competition on Adversarial Attacks and Defenses
- CAAD CTF @ Las Vegas
- CAAD CTF @ Shanghai

CAAD

ROUND 1 RUNNING

GeekPwn 极棒

ROUND 1

00 : 20 : 57

RANKING

1	TSAIL	1030
2	YYZZ	1010
3	BLADE	1000
4	NORTHWESTSEC	990
5	UCNESL	990
6	JD-OMEGA	980

ATTACKS LOGS

- BLADE CLASSIFY AS: TOUCAN
- TSAIL ATTACK NORTHWESTSEC
- NORTHWESTSEC CLASSIFY AS: MUSHRC
- YYZZ ATTACK UCNESL
- UCNESL CLASSIFY AS: ARCTIC FOX
- YYZZ ATTACK TSAIL
- TSAIL CLASSIFY AS: ARCTIC FOX
- YYZZ ATTACK JD-OMEGA
- JD-OMEGA CLASSIFY AS: REDBONE
- YYZZ ATTACK NORTHWESTSEC
- NORTHWESTSEC CLASSIFY AS: ARCTIC I

CAAD 线上赛

- 三个子赛项 Non-targeted Attack, Targeted Attack, Defense
- Google Brain的Ian Goodfellow，清华大学的朱军教授为顾问；Google Brain的Alexey Kurakin，加州大学伯克利分校的Dawn Song教授为评审委员
- 5月10号开始提交到8月31号截止
- Non-targeted attack参赛队伍30支，Targeted attack参赛队伍23支，Defense参赛队伍33支
- 各个子项目第一名人民币100,000元，第二名人民币 55,000 元，第三名人民币 35,000元，第四，五名人民币 6,000元

极棒AI赛事展望

- 鼓励更多人工智能从业者关注AI安全
- 比赛将覆盖更多的安全模型
- 比赛规则和比赛流程将更加科学和公平
- 促进更多AI安全研究者和传统信息安全领域的交流

谢谢！