**AI时代的攻与防**

# 生活中的AI

大概是一年前，当我看到这张漫画时，我家里除了苹果系列的手机和MacBook，几乎再难以找到一个与云或者说AI沾边的产品。AI也只是我研究的一个方向但是和我的生活没有太大关系。



智能家居漫画图

这个观点很快就被打破了，我所在的公司的门禁和消费系统可以使用人脸识别，真正实现了刷脸上班吃鸡翅。前不久我也赶时髦买了智能音箱，非常意外的是我家的加湿器也居然可以被智能音箱控制，我家的网络电视机顶盒安装一个小软件后也可以和它联动。经过简单调试后，智能手机都不愿意用的老父亲，已经学会使用音箱在电视上选电影看，我的不到三岁的娃娃也学会用智能音箱听小猪佩琪。AI设备已经润物细无声，双十一的智能音箱已经不到100块了。



（亚马逊的智能音箱）



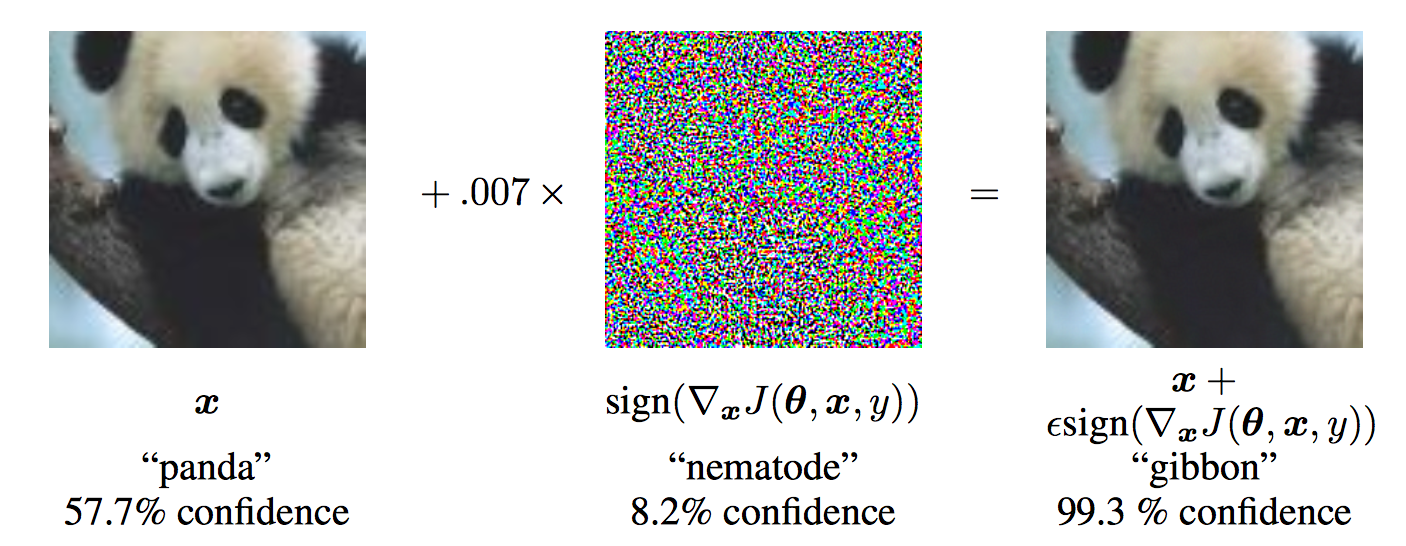
（JD的智能音箱）

# AI设备的安全

一次偶尔的机会，我在城铁上发现也可以远程管理我家的加湿器和智能音箱，我突然意识到，这些AI家居设备时刻与云联通，同时也与家里其他网络终端共享一个局域网，如果存在安全问题，黑客是否也可以远程控制他们，也可以时刻像音箱一样监听我们的谈话，嗅探我家网络上发生的一切呢？AI设备的安全显然是AI安全的一个重要领域。

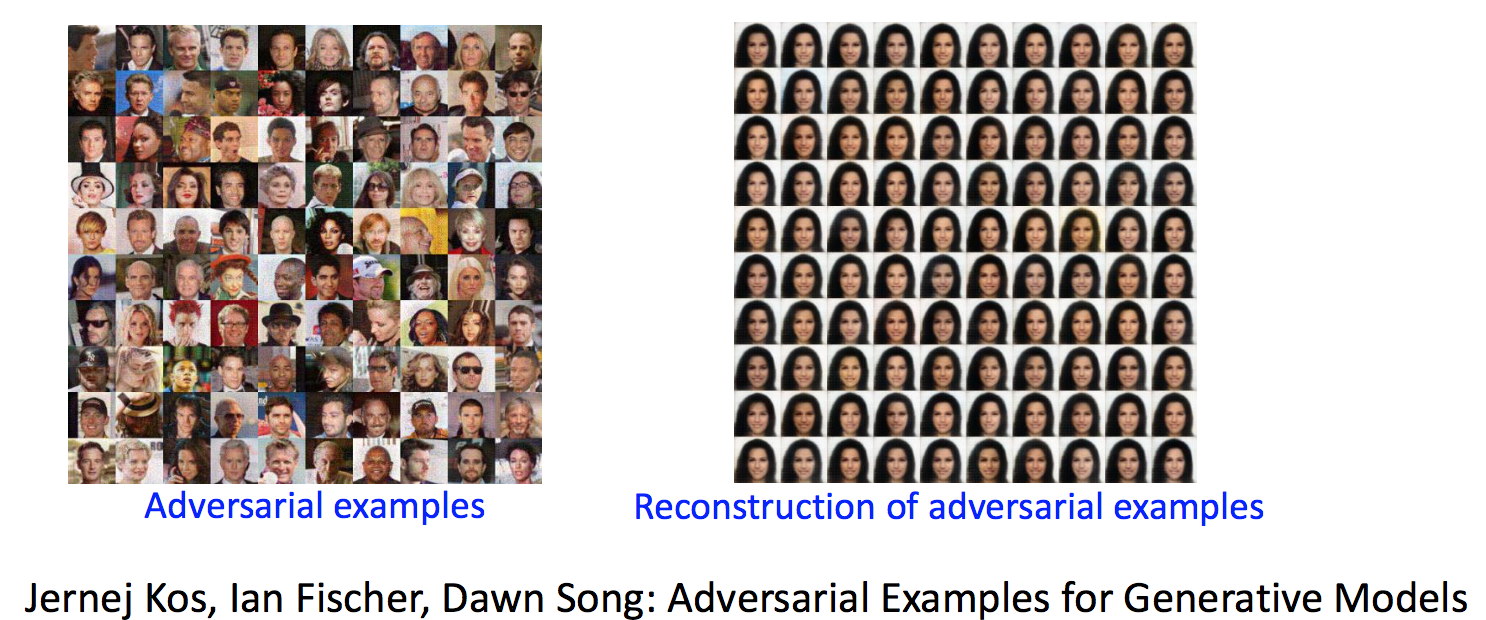
# AI模型的安全

AI算法听起来好像遥不可及，但是图像分类，语音识别和自然语言处理等领域，AI已经相当成熟。以图像分类来说，主流的算法已经可以达到99%以上的准确率。退役的美国网军司令曾经说过，世界上只有两种网络，一种是已经被攻破的，一种是不知道自己已经被攻破的。作为一个软件系统，AI算法或者说机器学习模型也是可以被欺骗的。一个经典案例就是针对图像分类模型的攻击，通过对熊猫照片的微小修改，人的肉眼几乎察觉不出任何变化，但是机器却会被欺骗，误判为长臂猿。



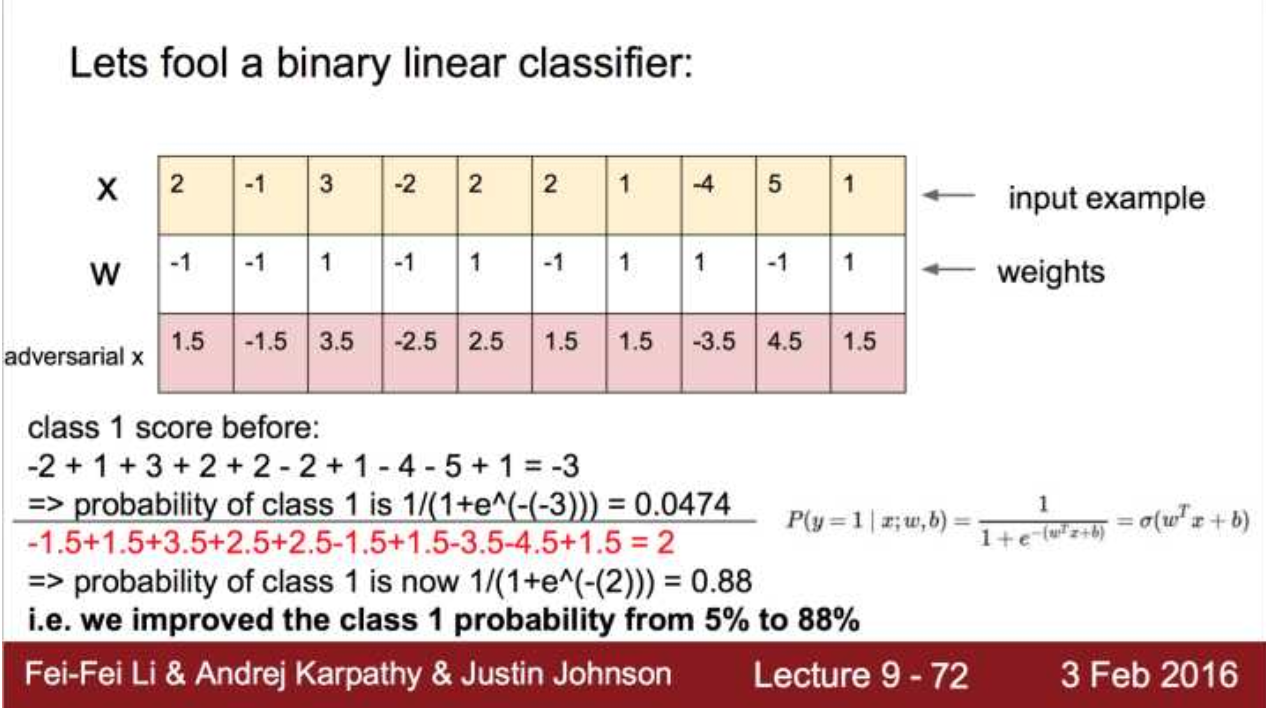
（精心处理的熊猫图片可以被机器学习模型识别为长臂猿）

获得美国麦克阿瑟天才奖的Dawn Song教授及其团队在这方面做了非常深入的研究，她们可以通过人眼无法识别的微小修改欺骗机器学习模型，把众人都识别为同一个人。



（微小的改变可以把众人都识别为同一个人）

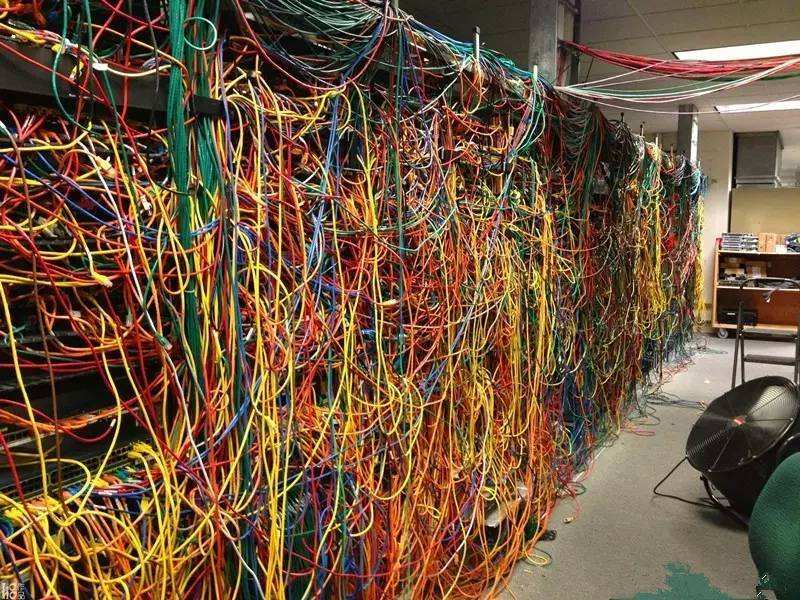
本质上机器学习模型是在多维特征向量层面，通过样本学习，迭代计算出分类结果，适当的微小调整也可以对分类结果产生显著影响。以下图为例，X的取值为（2，-1，3，-2，2，2，1，-4，5，1），但是只要稍微修改成（1.5，-1.5，3.5，-2.5，1.5，1.5，-3.5，4.5，1.5），分类结果为1的概率就可以从5%提升为88%。诸如这类针对AI模型的攻击也是AI安全的重要领域。



（特征向量微小变化也可以对机器学习模型分类结果产生巨大影响）

# 使用AI做安全

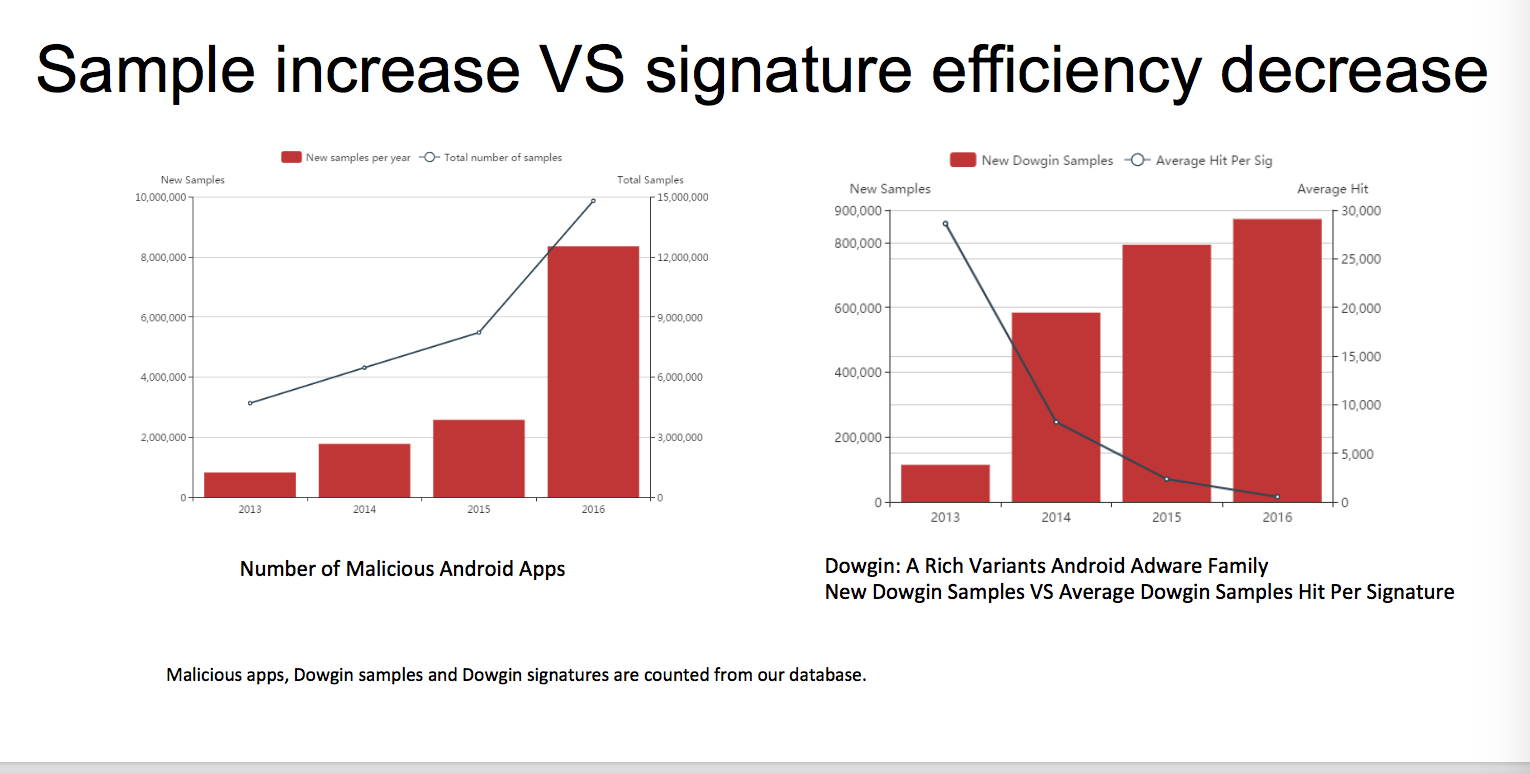
传统安全领域，无论是主动的威胁发现和安全防护还是被动的入侵检测，基本都是基于以往经验总结的静态检测规则和黑白名单。这些技术在过去很长一段时间已经被证明难以抵抗黑产以及针对性极强的商业间谍，基于经验积累的静态规则技术，总是处于被动挨打的境地，无论是精心的加密混淆还是没有补丁的零日攻击，多可以轻松绕过现有的防护体系。安全圈有句戏言，这种安全设备是人多聪明它多聪明，人已经想到的，没准能防住，没想到的肯定放不住。另外，这种基于规则的安全技术，现实中的最大问题是规则的难以维护，规则之间的重复与冲突更让这些问题雪上加霜，堪比一片混乱的机房。



（一片混乱的机房）

是否可以使用AI技术给安全领域带来一股新的力量呢？2015年，微软在Kaggle上发起了一个[恶意代码分类的比赛](https://www.kaggle.com/c/malware-classification/" \t "_blank)，并提供了超过500G的原始数据。有意思的是，取得第一名的队伍三个人所采用的方法与我们常见的方法存在很大不同，展现了[AI](http://lib.csdn.net/base/machinelearning" \o "机器学习知识库" \t "_blank)在安全领域的巨大潜力。早期的反病毒软件大都单一的采用特征匹配的方法，简单的利用特征串完成检测。随着恶意代码技术的发展，恶意代码开始在传播过程中进行变形以躲避查杀，此时同一个恶意代码的变种数量急剧提升，形态较本体也发生了较大的变化，反病毒软件已经很难提取出一段代码作为恶意代码的特征码。Kaggle比赛中最重要的环节就是特征工程，特征的好坏直接决定了比赛成绩。在这次Kaggle的比赛中冠军队伍选取了三个黄金特征，恶意代码图像、 OpCode n-gram 和 Headers 个数，其他一些特征包括 ByteCode n-gram ，指令频数等。机器学习部分采用了随机森林算法，并用到了 xgboost 和 pypy 加快训练速度，最终他们检测的效果超过了常见传统检测方式取得了冠军。

在移动领域，使用类似的思路也取得了不错的成绩，百度安全使用深度学习识别恶意APK，准确率达到99.96%，召回率达到了80%，2016年，反映该研究成果的论文[AI BASED ANTIVIRUS: CAN ALPHAAV WIN THE BATTLE IN WHICH MAN HAS FAILED?](https://www.blackhat.com/eu-16/briefings.html" \l "ai-based-antivirus-can-alphaav-win-the-battle-in-which-man-has-failed" \o "AI based Antivirus: Can AlphaAV Win the Battle in which Man Has Failed?)被blackhat会议收录并做了相关演讲。恶意APK伴随移动互联网井喷式的发展，其数量在近几年几何级别增长，传统的基于规则的检测技术已经无法覆盖如此大量的恶意程序。



（恶意APK数量猛增）

在大量的人工分析恶意APK的工作中发现，人工分析的过程很类似图像识别的过程。深度学习在图像识别领域有着成熟的应用，是否可以通过提取APK特征，通过深度学习的方法来自动化识别恶意APK呢？



（恶意APK分析与图像识别）

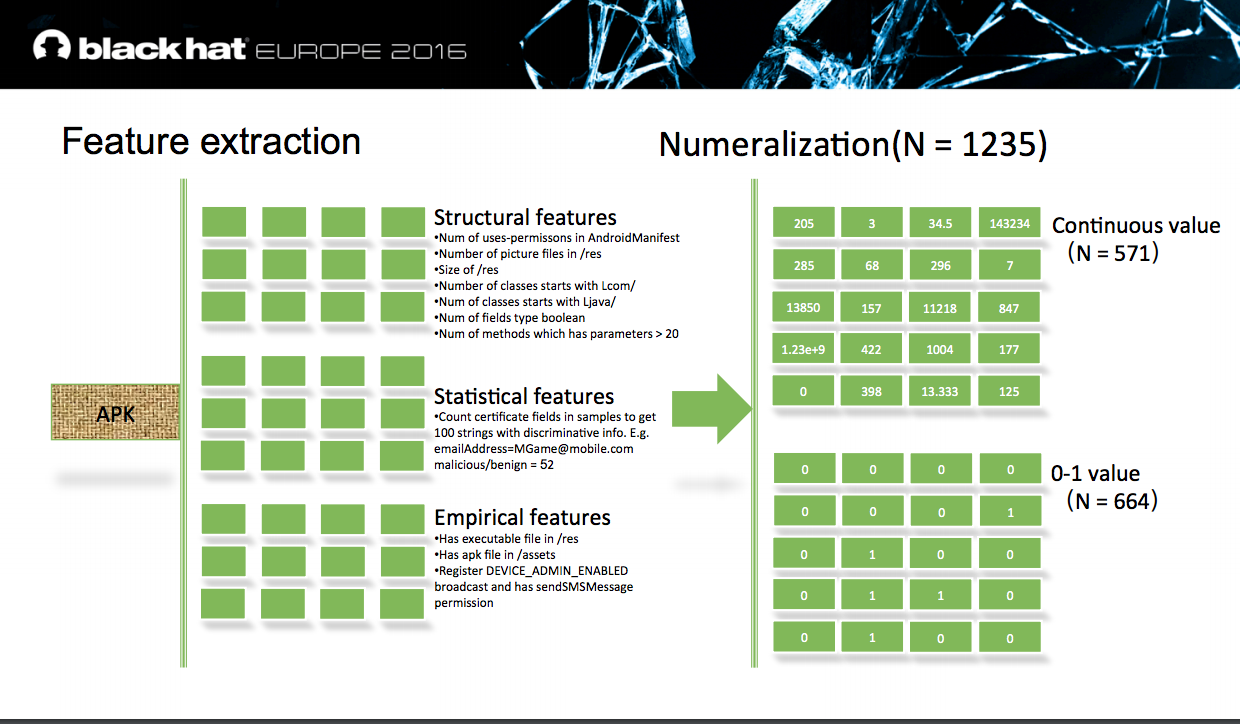
对APK的特征提取主要分为三大类特征：

第一类是结构化特征，包括APK申请的权限的个数，资源文件中包含的图像文件个数和参数大于20的函数的个数等。

第二类是统计类特征，包括近千条统计特征。

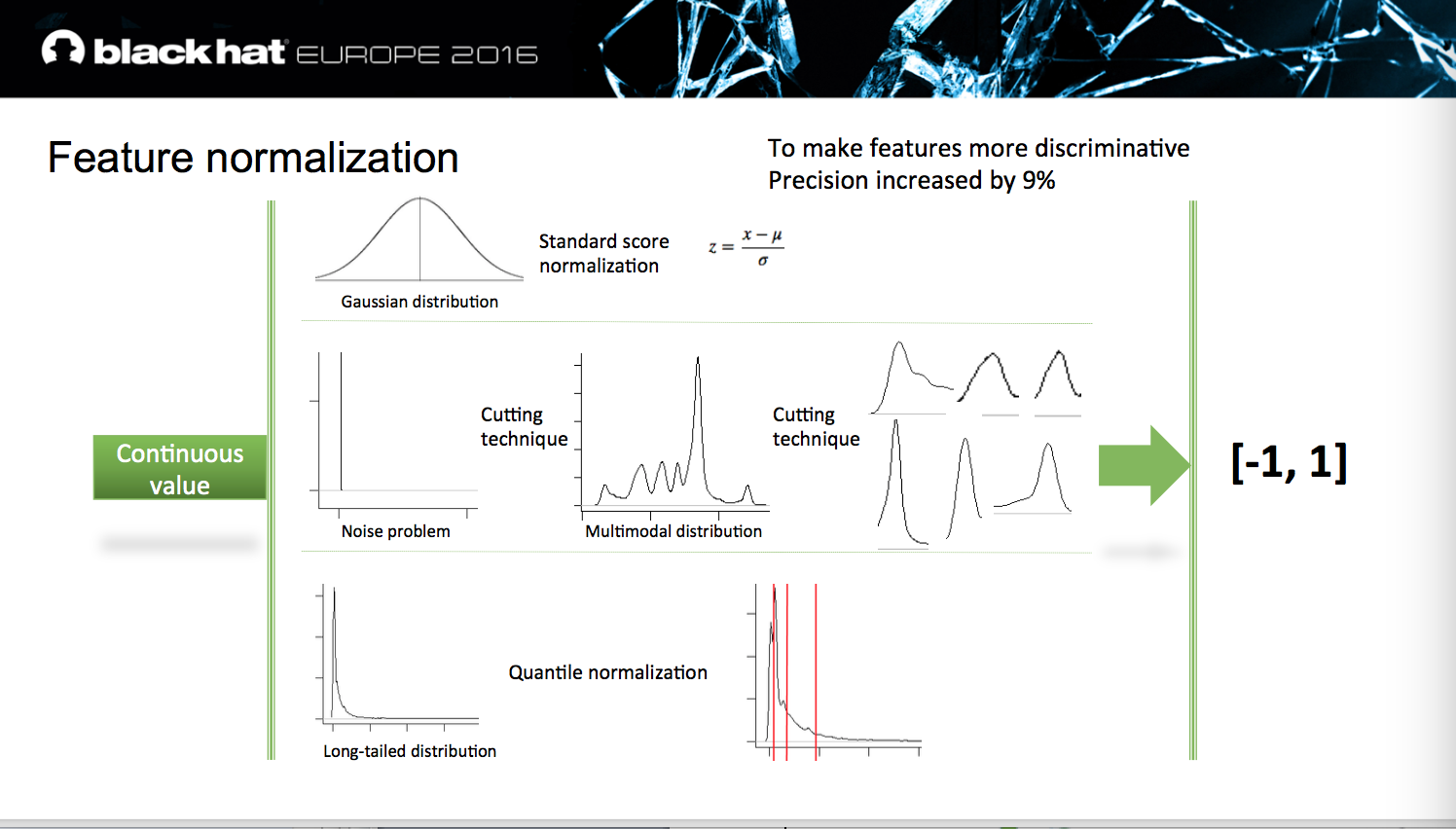
第三类是长期恶意APK检测的经验总结的特征，包括资源文件中是否包含可执行文件，assets文件夹中是否包含APK文件等。

特征提取过程如下图所示。



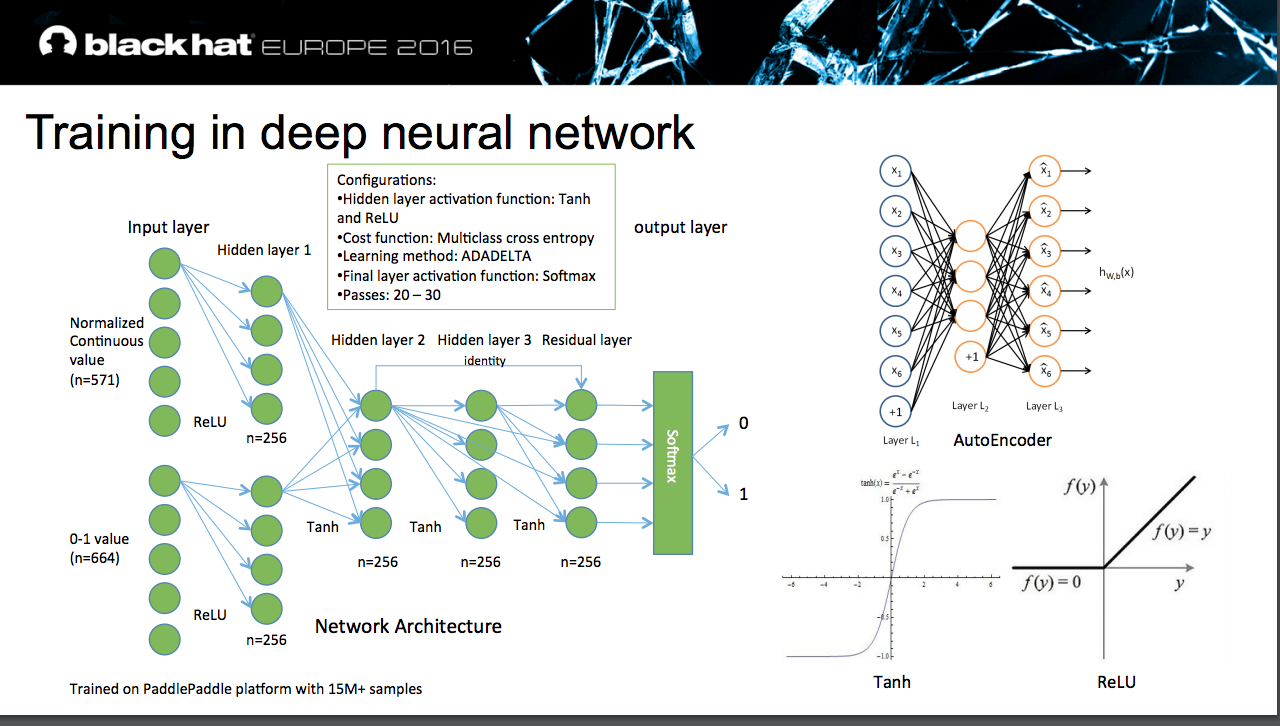
（APK特征提取过程）

归一化处理一千多个特征，取值控制在-1和1之间。



（归一化APK特征）

使用深度学习网络训练，训练过程如下图所示。



（APK深度学习训练过程）

整个训练过程中使用超过了15万个样本，训练使用了百度开源的深度学习框架Paddle。

事实上，基于AI技术，在安全领域的常见分类问题上，都取得了不错的进展，比如垃圾邮件，恶意网址，黄反鉴别，病毒检测等。这块有兴趣多了解的可以看我的《Web安全之机器学习入门》。

# 使用AI做攻击

利益驱动的黑产，往往在新技术的使用上超过大家的想象。2017年，腾讯守护者计划安全团队协助警方打掉市面上最大打码平台“快啊答题”，挖掘出一条从撞库盗号、破解验证码到贩卖公民信息、实施网络诈骗的全链条黑产。而在识别验证码这一关键环节，黑产运用AI技术训练机器，极大提升了单位时间内识别验证码的数量，2017年一季度打码量达到259亿次，且识别验证码的精准度超过80%。

# 小结

AI安全大体可以归纳为四类，细分为AI设备的安全，AI模型的安全，使用AI做正向安全建设以及使用AI发起攻击。哈佛大学发布《人工智能与国家安全》报告，指出AI的发展将通过变革军事优势、信息优势和经济优势等三方面影响国家安全，建议美国政府保持美国的技术领先地位，应资助多种形式的、长期的关于AI技术及其应用的战略分析，应有限考虑支持能带来持续利益和降低关键风险的AI研究与开发工作，重点投资在攻防两端“反AI”的能力。而最近，俄罗斯总统普京就AI领域的竞赛警告：“谁成为了这个领域的领导者，谁就将成为世界的统治者”。在2014年、2016年召开的两院院士大会，习近平总书记的讲话中都谈到了人工智能的迅猛发展，强调“我们不仅要把我国机器人水平提高上去，而且要尽可能多地占领市场”，并且还亲自担任中央网络安全和信息化领导小组组长。可见AI安全已经得到大国国家层面的重视。

# 参考文献

* <http://www.freebuf.com/news/140151.html>
* 《[AI BASED ANTIVIRUS: CAN ALPHAAV WIN THE BATTLE IN WHICH MAN HAS FAILED?](https://www.blackhat.com/eu-16/briefings.html#ai-based-antivirus-can-alphaav-win-the-battle-in-which-man-has-failed)》
* 《Adversarial Examples for Generative Models》
* 《Web安全之机器学习入门》
* 《人工智能与国家安全》