HZ BOOKS
华章IT

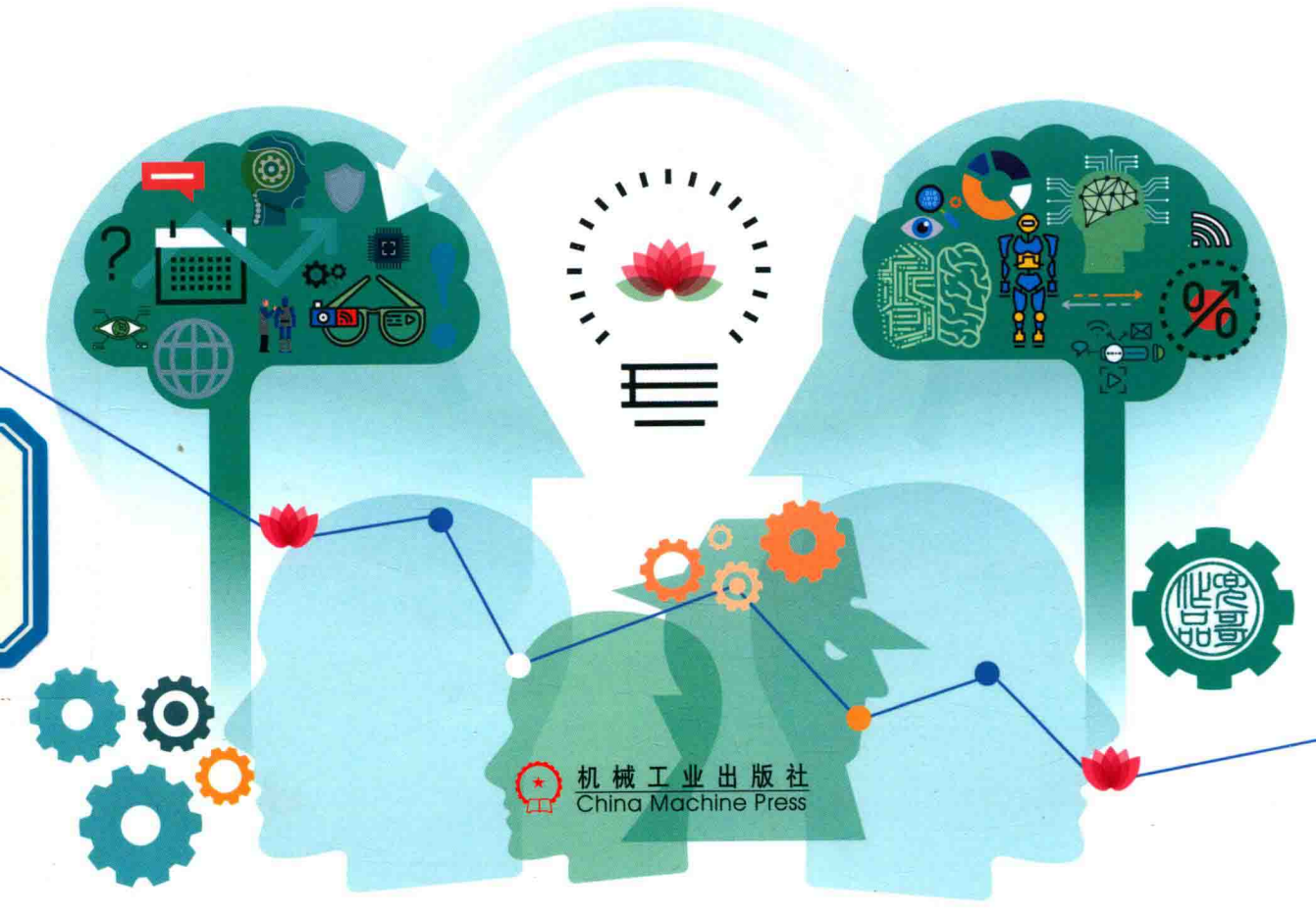
智能系统与技术丛书

百度安全专家撰写，Web安全工具智能化升级

Reinforcement Learning and Generative Adversarial Networks for the Web Security

Web安全之强化学习与GAN

刘 焱 编著



机械工业出版社
China Machine Press

Reinforcement Learning and Generative Adversarial Networks for the Web Security

Web安全之强化学习与GAN

刘焱 编著

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

Web 安全之强化学习与 GAN/ 刘焱编著 . —北京：机械工业出版社，2018.3
（智能系统与技术丛书）

ISBN 978-7-111-59345-4

I. W… II. 刘… III. 机器学习—研究 IV. TP181

中国版本图书馆 CIP 数据核字（2018）第 045513 号

Web 安全之强化学习与 GAN

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）
责任编辑：陈佳媛 责任校对：李秋荣
印 刷：北京诚信伟业印刷有限公司 版 次：2018 年 4 月第 1 版第 1 次印刷
开 本：186mm × 240mm 1/16 印 张：17
书 号：ISBN 978-7-111-59345-4 定 价：79.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88379426 88361066 投稿热线：（010）88379604
购书热线：（010）68326294 88379649 68995259 读者信箱：hzit@hzbook.com

PRAISE

对本书的赞誉

此亦笃信之年，此亦大惑之年。此亦多丽之阳春，此亦绝念之穷冬。人或万事俱备，人或一事无成。我辈其青云直上，我辈其黄泉永坠。——《双城记》狄更斯 著，魏易 译

如今是一个人工智能兴起的年代，也是一个黑产猖獗的年代；是一个机器学习算法百花齐放的年代，也是一个隐私泄露、恶意代码传播、网络攻击肆虐的年代。AlphaGo碾压柯洁之后，不少人担心AI会抢了人类的工作，然而信息安全领域专业人才严重匮乏，极其需要AI来补充专业缺口。

兜哥的这本书展示了丰富多彩的机器学习算法在错综复杂的Web安全中的应用，是一本非常及时的人工智能在信息安全领域的入门读物。正如书中所述，没有最好的算法，只有最合适的算法。虽然这几年深度学习呼声很高，但各种机器学习算法依然在形形色色的应用场景中有着各自独特的价值，熟悉并用好这些算法在安全领域的实战中会起到重要的作用。

——Lenx，百度首席安全科学家，安全实验室负责人

存储和计算能力的爆发式增长，让我们获得了比以往更全面、更实时获取以及分析数据的潜在能力，但面对产生的海量信息，如何快速准确地将其转化为业务需求，则需要依赖一些非传统的手段。就安全领域来说，原先依赖于规则的问题解法过于受限于编写规则的安全专家自身知识领域的广度和深度，以及对于问题本质的理解能力。但我们都知道，安全漏洞层出不穷，攻击利用的方式多种多样，仅仅依赖于规则来发现问题在现阶段的威胁形势下慢慢变得捉襟见肘。面对威胁，企业安全人员需要打造这样一种能力，它能够让我们脱离单纯的点对点的竞争，case by case的对抗，转而从更高的维度上来审视业务，发现潜在的异常事件，而这些异常事件可能会作为安全人员深入调查的起点。这种能力能让我们找到原有安全能力盲区以及发现新威胁，促使我们的技能水平以及对威胁的响应速度持续提升。同时这种能力和防御体系结合，也有可能让我们在面对某些未知威胁时，达到以不变应万变、获得天然免疫的理想状态。兜哥的这本书或许是开启我

们这种能力的一把钥匙。本书用通俗易懂的语言介绍了机器学习原理，结合实际企业中的安全业务需求场景，让广大安全人员能够感受到这种“如日中天”的技术在传统安全领域内如何大放异彩。最后，May the force be with you。

——王宇，蚂蚁金服安全总监

百度是拥有海量互联网数据的几家公司之一，兜哥是百度前IT安全负责人，现Web安全产品负责人，研发的产品不仅应用于百度公司内部检测网络攻击，也应用在多个百度的商业安全产品中，服务于数万站长。兜哥的团队是国内最早一批将机器学习算法应用于网络安全场景的团队之一，本书聚集了兜哥及其团队多年的安全实践经验，覆盖了互联网公司可能会遇到的多个安全场景，比如用图算法检测WebShell等，非常好地解决了百度商业安全客户被入侵留后门的问题。兜哥将自己的技术选型、算法、代码倾囊相授，我相信本书的出版将会大大降低安全研发工程师转型安全数据分析专家的难度，值得推荐。

——黄正，百度安全实验室X-Team负责人，MSRC 2016中国区第一

伴随着互联网的爆炸式发展，网络安全已上升到国家战略层面，能直接看到效果的安全能力建设得到高度重视。与此同时，安全团队却又不得不面对百花齐放的业务场景、大规模的数据中心，以及愈加剧烈、复杂和不确定性的网络攻击。如何在传统攻防对抗之外寻找更有效、可落地的对抗方式，已成为各大企业安全团队思考的重点。所幸，近些年来，计算和存储资源已不再是安全团队的瓶颈，安全团队自身在工程能力上也已非昔日吴下阿蒙。机器学习成为近些年来安全领域里第一批从学术走向工业的应用方向，并已有很多阶段性的实践成果。很欣喜地看到兜哥一直在推进机器学习系列的文章并编写了此书。此书重点讲解了常见机器学习算法在不同场景下的潜在应用和实践，非常适合初学者入门。希望此书能够启发更多的同行继续实践和深耕机器学习应用这个方向，并给安全行业带来更多的反馈和讨论。

——程岩，京东安全首席架构师

网络安全是信息时代的重大挑战和核心课题之一，而机器学习是迄今为止人工智能大厦最坚实稳固的基石。本书从基本原理出发，通过实际案例深入介绍和分析机器学习技术和算法在网络安全领域的应用与实践，是一本不可多得的入门指南和参考手册。

——姚聪博士，北京旷视科技（Face++）有限公司高级研究员

PREFACE

前　言

网络安全一直和 AI 相伴相生，从网络安全诞生的那一天起，人们就一直试图使用自动化的方式去解决安全问题。网络安全专家一直试图把自己对网络威胁的理解转换成机器可以理解的方式，比如黑白名单、正则表达式，然后利用机器强大的计算能力，夜以继日地从流量、日志、文件中寻找似曾相识的各类威胁。似乎这一切就是那么天经地义并无懈可击。事情似乎又没有那么简单，机器其实并没有完全学到人的经验，网络安全专家一眼就可以识破的变形，对于机器却难以理解；更可怕的是，恶意程序数量呈指数增长，各类新型攻击方式层出不穷，0day（零日攻击）的出现早已超过一线明星出现在新闻头条的频率，依靠极其有限的网络专家总结的经验和几个安全厂商所谓的样本交换，已经难以应付现在的网络安全威胁。如果安全专家一眼就可以识破的威胁，机器也能够自动化发现甚至做出相应的响应，这已经是很大的进步；如果让机器可以像 AlphaGo 理解围棋一样，能够理解网络威胁，那将是巨大进步。事情又回到最初的那个问题，如何能让机器真正学会识别安全威胁？机器学习可能是一个不错的答案。

目标读者

本书面向信息安全从业人员、大专院校计算机相关专业学生以及信息安全爱好者、机器学习爱好者，对于想了解人工智能的 CTO、运维总监、架构师，本书同样也是一本不错的科普书籍。如果看完本书，可以让读者在工作学习中遇到问题时想起一到两种算法，那么我觉得就达到效果了；如果可以让读者像使用 printf 一样使用 SVM、朴素贝叶斯等算法，那么这本书就相当成功了。

我写本书的初衷是帮助安全爱好者以及信息安全从业者了解机器学习，可以动手使用简单的机器学习算法解决实际问题。在写作中尽量避免生硬的说教，能用文字描述的尽量不用冷冰冰的公式，能用图和代码说明的尽量不用多余的文字，正如霍金说言，“多

写一个公式，少一半读者”，希望反之亦然。

机器学习应用于安全领域遇到的最大问题就是缺乏大量的黑样本，即所谓的攻击样本，尤其相对于大量的正常业务访问，攻击行为尤其是成功的攻击行为是非常少的，这就给机器学习带来了很大挑战。本书很少对不同算法进行横向比较，也是因为在不同场景下不同算法的表现差别的确很大，很难说深度学习就一定比朴素贝叶斯好，也很难说支持向量机就不如卷积神经网络，拿某个具体场景进行横评意义不大，毕竟选择算法不像购买 SUV，可以拿几十个参数评头论足，最后还是需要大家结合实际问题去选择。

如何使用本书

本书的第 1 章主要介绍了如何打造自己的深度学习工具箱，介绍了 AI 安全的攻与防，介绍了针对 AI 设备和 AI 模型的攻击，以及使用 AI 进行安全建设和攻击。第 2 章介绍了如何打造深度学习的工具箱。第 3 章介绍了如何衡量机器学习算法的性能以及集成学习的基本知识。第 4 章介绍了 Keras 的基本知识以及使用方法，这章是后面章节学习开发的基础。第 5 章介绍了强化学习，重点介绍了单智力体的强化学习。第 6 章介绍了 Keras 下强化学习算法的一种实现 Keras-rl。第 7 章介绍了强化学习领域经常使用的 OpenAI Gym 环境。第 8 章～第 10 章，介绍了基于机器学习的恶意程序识别技术以及常见的恶意程序免杀方法，最后介绍了如何使用强化学习生成免杀程序，并进一步提升杀毒软件的检测能力。第 11 章介绍如何使用强化学习提升 WAF 的防护能力，第 12 章介绍如何使用强化学习提升反垃圾邮件的检测能力。第 13 章介绍了对抗生成网络的基础知识，第 14 章介绍了针对机器学习模型的几种攻击方式，包括如何欺骗图像识别模型让其指鹿为马。每个案例都使用互联网公开的数据集并配有基于 Python 的代码，代码和数据集可以在本书配套的 GitHub 下载。

本书是我机器学习三部曲的第三部，在第一部中，主要以机器学习常见算法为主线，以生活中的例子和具体安全场景介绍机器学习常见算法，定位为机器学习入门书籍，便于大家快速上手。全部代码都可以在普通 PC 电脑上运行。在第二部中，重点介绍深度学习，并以具体的 11 个案例介绍机器学习的应用，面向的是具有一定机器学习基础或者致力于使用机器学习解决工作中问题的读者。本书重点介绍强化学习和对抗网络，并介绍了 AI 安全的攻与防。一直有个遗憾的地方：深度学习的优势发挥需要大量精准标注的训练样本，但是由于各种各样的原因，我只能在书中使用互联网上已经公开的数据集，数据量级往往很难发挥深度学习的优势，对于真正想在生产环境中验证想法的读者需要搜集更多样本。

致谢

这里我要感谢我的家人对我的支持，本来工作就很忙，没有太多时间处理家务，写

书以后更是花费了我大量的休息时间，我的妻子无条件承担起了全部家务，尤其是照料孩子方面的繁杂事务。我很感谢我的女儿，写书这段时间几乎没有时间陪她玩，她也很懂事地自己玩，我也想用这本书作为生日礼物送给她。我还要感谢编辑吴怡对我的支持和鼓励，让我可以坚持把这本书写完。最后还要感谢各位业内好友尤其是我 boss 对我的支持，排名不分先后：马杰 @ 百度安全、冯景辉 @ 百度安全、Tony@ 京东安全、程岩 @ 京东安全、简单 @ 京东安全、林晓东 @ 百度基础架构、黄颖 @ 百度 IT、李振宇 @ 百度 AI、Lenx@ 百度安全、黄正 @ 百度安全、郝轶 @ 百度云、云鹏 @ 百度无人车、赵林林 @ 微步在线、张宇平 @ 数盟、谢忱 @Freebuf、李新 @Freebuf、李琦 @ 清华、徐恪 @ 清华、王宇 @ 蚂蚁金服、王泯然 @ 蚂蚁金服、王龙 @ 蚂蚁金服、周涛 @ 启明星辰、姚志武 @ 借贷宝、刘静 @ 安天、刘元军 @ 医渡云、廖威 @ 易宝支付、尹毅 @sobug、宋文宽 @ 联想、团长 @ 宜人贷、齐鲁 @ 搜狐安全、吴圣 @58 安全、康宇 @ 新浪安全、幻泉 @i 春秋、雅驰 @i 春秋、王庆双 @i 春秋、张亚同 @i 春秋、王禾 @ 微软、李臻 @paloalto、西瓜 @ 四叶草、郑伟 @ 四叶草、朱利军 @ 四叶草、土夫子 @XSRC、英雄马 @ 乐视云、sbilly@360、侯曼 @360、高磊 @ 滴滴、高磊 @ 爱加密、高渐离 @ 华为、刘洪善 @ 华为云、宋柏林 @ 一亩田、张昊 @ 一亩田、张开 @ 安恒、李硕 @ 智联、阿杜 @ 优信拍、李斌 @ 房多多、李程 @ 搜狗、姚聪 @face+、李鸣雷 @ 金山云、吴鲁加 @ 小密圈，最后我还要感谢我的亲密战友陈燕、康亮亮、蔡奇、哲超、新宇、子奇、月升、王磊、碳基体、刘璇、钱华沟、刘超、王胄、吴梅、冯侦探、冯永校。

我平时在 Freebuf 专栏以及“i 春秋”分享企业安全建设以及人工智能相关经验与最新话题，同时也运营我的微信公众号“兜哥带你学安全”，欢迎大家关注并在线交流。

本书使用的代码和数据均在 GitHub 上发布，地址为：https://github.com/duoergun0729/3book，代码层面任何疑问可以在 GitHub 上直接反馈。

CONTENTS

目　录

对本书的赞誉
前言

第 1 章　AI 安全之攻与防 ····· 1

1.1　AI 设备的安全 ····· 2
1.2　AI 模型的安全 ····· 3
1.3　使用 AI 进行安全建设 ····· 4
1.4　使用 AI 进行攻击 ····· 9
1.5　本章小结 ····· 9

第 2 章　打造机器学习工具箱 ····· 11

2.1　TensorFlow ····· 11
2.2　Keras ····· 13
2.3　Anaconda ····· 14
2.4　OpenAI Gym ····· 19
2.5　Keras-rl ····· 19
2.6　XGBoost ····· 19
2.7　GPU 服务器 ····· 20
2.8　本章小结 ····· 23

第 3 章　性能衡量与集成学习 ····· 24

3.1　常见性能衡量指标 ····· 24
3.1.1　测试数据 ····· 24
3.1.2　混淆矩阵 ····· 25
3.1.3　准确率与召回率 ····· 25
3.1.4　准确度与 F1-Score ····· 26
3.1.5　ROC 与 AUC ····· 27
3.2　集成学习 ····· 28
3.2.1　Boosting 算法 ····· 29
3.2.2　Bagging 算法 ····· 31
3.3　本章小结 ····· 32

第 4 章　Keras 基础知识 ····· 34

4.1　Keras 简介 ····· 34
4.2　Keras 常用模型 ····· 35
4.2.1　序列模型 ····· 35
4.2.2　函数式模型 ····· 35
4.3　Keras 的网络层 ····· 36
4.3.1　模型可视化 ····· 36
4.3.2　常用层 ····· 38
4.3.3　损失函数 ····· 44
4.3.4　优化器 ····· 44
4.3.5　模型的保存与加载 ····· 45
4.3.6　基于全连接识别 MNIST ····· 45
4.3.7　卷积层和池化层 ····· 47
4.3.8　基于卷积识别 MNIST ····· 49
4.3.9　循环层 ····· 49
4.3.10　基于 LSTM 进行 IMDB 情感分类 ····· 52

4.4 本章小结 …… 54

第 5 章 单智力体强化学习 …… 55

5.1 马尔可夫决策过程 …… 55
5.2 Q 函数 …… 56
5.3 贪婪算法与ϵ- 贪婪算法 …… 57
5.4 Sarsa 算法 …… 59
案例 5-1：使用 Sarsa 算法处理金币问题 …… 60
5.5 Q Learning 算法 …… 62
案例 5-2：使用 Q Learning 算法处理金币问题 …… 63
5.6 Deep Q Network 算法 …… 64
案例 5-3：使用 DQN 算法处理 CartPole 问题 …… 65
5.7 本章小结 …… 71

第 6 章 Keras-rl 简介 …… 72

6.1 Keras-rl 智能体介绍 …… 73
6.2 Keras-rl 智能体通用 API …… 73
6.3 Keras-rl 常用对象 …… 75
案例 6-1：在 Keras-rl 下使用 SARSA 算法处理 CartPole 问题 …… 75
案例 6-2：在 Keras-rl 下使用 DQN 算法处理 CartPole 问题 …… 77
案例 6-3：在 Keras-rl 下使用 DQN 算法玩 Atari 游戏 … 78
6.4 本章小结 …… 86

第 7 章 OpenAI Gym 简介 …… 87

7.1 OpenAI …… 87
7.2 OpenAI Gym …… 88
7.3 Hello World！ OpenAI Gym …… 89
7.4 编写 OpenAI Gym 环境 …… 92
7.5 本章小结 …… 98

第 8 章 恶意程序检测 …… 99

8.1 PE 文件格式概述 …… 100
8.2 PE 文件的节 …… 104
8.3 PE 文件特征提取 …… 107
8.4 PE 文件节的特征提取 …… 119
8.5 检测模型 …… 121
8.6 本章小结 …… 129

第 9 章 恶意程序免杀技术 …… 130

9.1 LIEF 库简介 …… 130
9.2 文件末尾追加随机内容 …… 132
9.3 追加导入表 …… 132
9.4 改变节名称 …… 133
9.5 增加节 …… 134
9.6 节内追加内容 …… 135
9.7 UPX 加壳 …… 135
9.8 删除签名 …… 137
9.9 删除 debug 信息 …… 138
9.10 置空可选头的交验和 …… 138
9.11 本章小结 …… 138

第 10 章 智能提升恶意程序检测能力 …… 139

10.1 Gym-Malware 简介 …… 139
10.2 Gym-Malware 架构 …… 141
10.2.1 PEFeatureExtractor …… 141
10.2.2 Interface …… 143
10.2.3 MalwareManipulator …… 143
10.2.4 DQNAgent …… 144
10.2.5 MalwareEnv …… 145

10.3 恶意程序样本 …… 148
10.4 本章小结 …… 149

第 11 章 智能提升 WAF 的防护能力 …… 150

11.1 常见 XSS 攻击方式 …… 151
11.2 常见 XSS 防御方式 …… 152
11.3 常见 XSS 绕过方式 …… 153
11.4 Gym-WAF 架构 …… 155
11.4.1 Features 类 …… 156
11.4.2 Xss_Manipulator 类 …… 156
11.4.3 DQNAgent 类 …… 160
11.4.4 WafEnv_v0 类 …… 161
11.4.5 Waf_Check 类 …… 162
11.5 效果验证 …… 163
11.6 本章小结 …… 164

第 12 章 智能提升垃圾邮件检测能力 …… 165

12.1 垃圾邮件检测技术 …… 166
12.1.1 数据集 …… 166
12.1.2 特征提取 …… 168
12.1.3 模型训练与效果验证 …… 171
12.1.4 模型的使用 …… 172
12.2 垃圾邮件检测绕过技术 …… 173
12.2.1 随机增加 TAB …… 174
12.2.2 随机增加回车 …… 174
12.2.3 大小写混淆 …… 175
12.2.4 随机增加换行符 …… 175
12.2.5 随机增加连字符 …… 176
12.2.6 使用错别字 …… 176
12.3 Gym-Spam 架构 …… 177
12.3.1 Features 类 …… 178
12.3.2 Spam_Manipulator 类 …… 178
12.3.3 DQNAgent 类 …… 179
12.3.4 SpamEnv_v0 类 …… 181
12.4 效果验证 …… 182
12.5 本章小结 …… 183

第 13 章 生成对抗网络 …… 184

13.1 GAN 基本原理 …… 184
13.2 GAN 系统架构 …… 185
13.2.1 噪音源 …… 185
13.2.2 Generator …… 186
13.2.3 Discriminator …… 187
13.2.4 对抗模型 …… 188
13.3 GAN …… 188
13.4 DCGAN …… 194
13.5 ACGAN …… 202
13.6 WGAN …… 210
13.7 本章小结 …… 217

第 14 章 攻击机器学习模型 …… 218

14.1 攻击图像分类模型 …… 218
14.1.1 常见图像分类模型 …… 219
14.1.2 梯度算法和损失函数 …… 222
14.1.3 基于梯度上升的攻击原理 …… 224
14.1.4 基于梯度上升的算法实现 …… 226
14.1.5 基于 FGSM 的攻击原理 …… 228
14.1.6 基于 FGSM 攻击的算法实现 …… 229
14.2 攻击其他模型 …… 231
案例 14-1：攻击手写数字识别模型 …… 233
案例 14-2：攻击自编码器 …… 240
案例 14-3：攻击差分自编码器 …… 249
14.3 本章小结 …… 262

CHAPTER 1

第 1 章

AI 安全之攻与防

大概一年前我看到下面这张漫画（见图 1-1），当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，但是它和我的生活并没有太大关系。

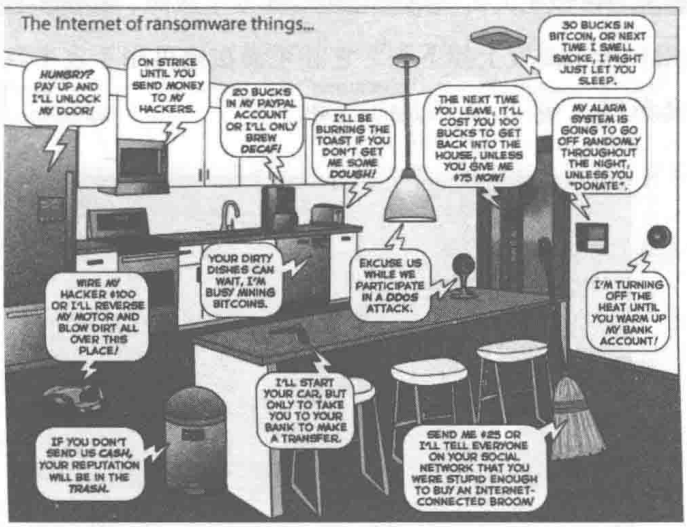


图 1-1　智能家居漫画图

这个观念很快就被打破了，我所在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，连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，我那不到三岁的女儿也学会了用智能音箱听小猪佩奇。AI 设备润物细无声，双十一时智能音箱已经不到 100 元了。

1.1 AI 设备的安全

一次偶然的机会，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 家居设备时刻与云连通，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它们，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？

在 2017 年的 BlackHat 安全会议上，阿里巴巴安全部门的研究人员演示了用声音和超声攻击依赖于陀螺仪、加速度计等微机电系统传感器输入信号的智能设备（见图 1-2）。这种声音武器在理论上可以让无人机坠落，让机器人发生故障，让虚拟现实和增强现实软件失去方向感，让用户从平衡板上摔下来，它甚至潜在地可用于攻击自主驾驶汽车或干扰汽车的安全气囊传感器。

图 1-2 阿里巴巴安全研究人员演示了用声音和超声攻击智能设备[⊖]

⊖ http://www.cnbeta.com/articles/tech/636609.htm

因此，AI 设备的安全显然是 AI 安全的一个重要领域。

1.2 AI 模型的安全

AI 算法听起来好像遥不可及，但是在图像分类、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（见图 1-3）。

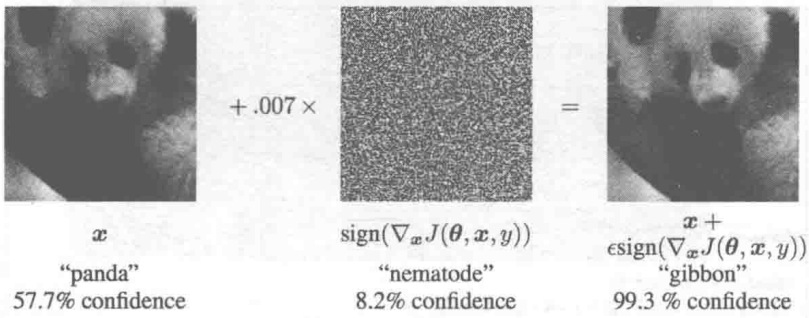


图 1-3 精心处理的熊猫图片可以被机器学习模型识别为长臂猿

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（见图 1-4）。

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，精心构造的微小调整也可以对分类结果产生显著影响。以图 1-5 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。

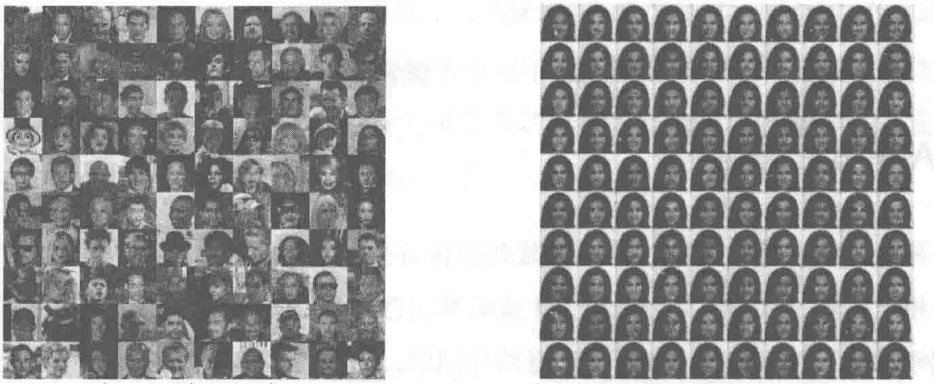


图 1-4 微小的改变可以把众人都识别为同一个人[⊖]

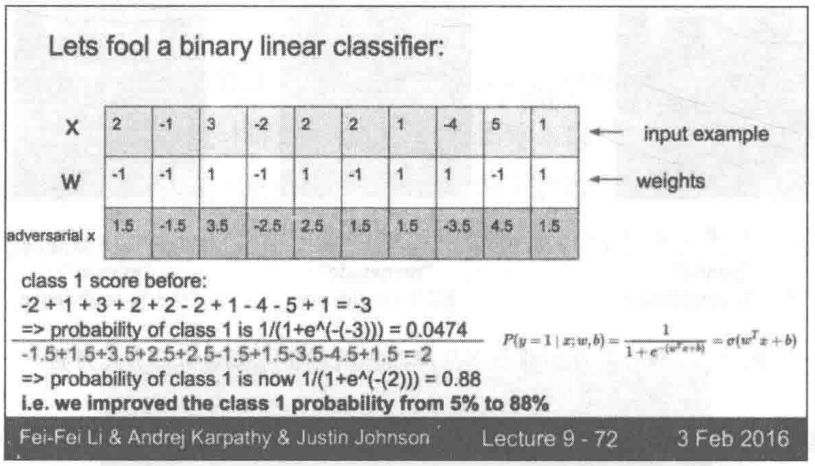


图 1-5 特征向量微小变化也可以对分类结果产生巨大影响（图片来自 Stanford CS231n 2016）

1.3 使用 AI 进行安全建设

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

⊖ https://arxiv.org/abs/1702.06832

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，大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，人没想到的肯定防不住”。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（见图 1-6）。

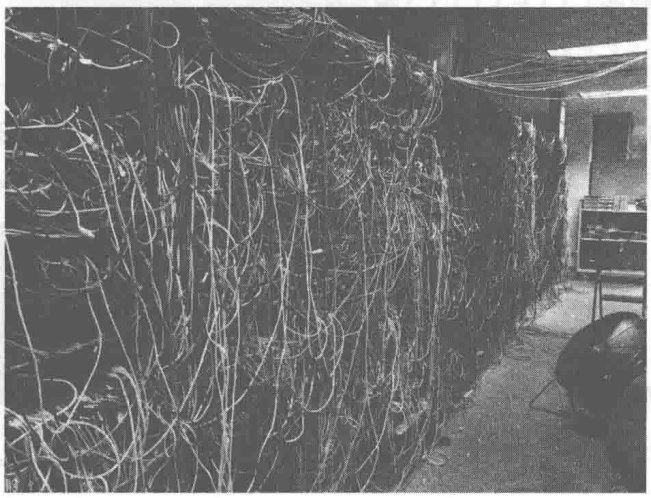


图 1-6　一片混乱的机房

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，并提供了超过 500 G 的原始数据。有意思的是，取得第一名的队所采用的方法与我们常见的方法存在很大不同，展现了 AI 在安全领域的巨大潜力。早期的反病毒软件大都单一地采用特征匹配的方法，简单地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获得了冠军。

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?》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发展，其数量在近几年呈几何级增长（见图 1-7），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。

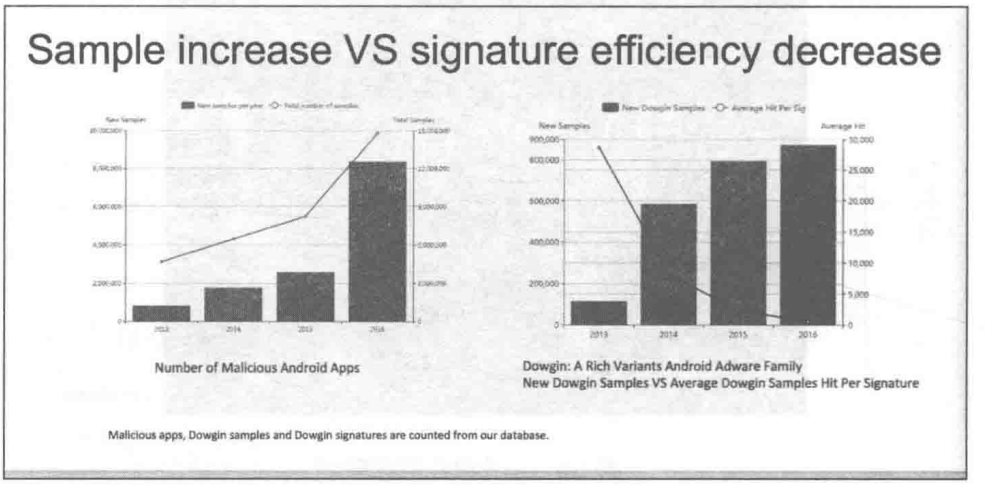


图 1-7 恶意 APK 数量猛增[⊖]

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（见图 1-8）。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？

对 APK 的特征提取主要分为三大类：

- 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。

⊖ 图 1-7 至图 1-11 均引自网址 http://www.blackhat.com/eu-16/briefings.html#ai-based-antivirus-can-alphaav-win-the-battle-in-which-man-has-failed。