HZ BOOKS
华章IT

智能系统与技术丛书

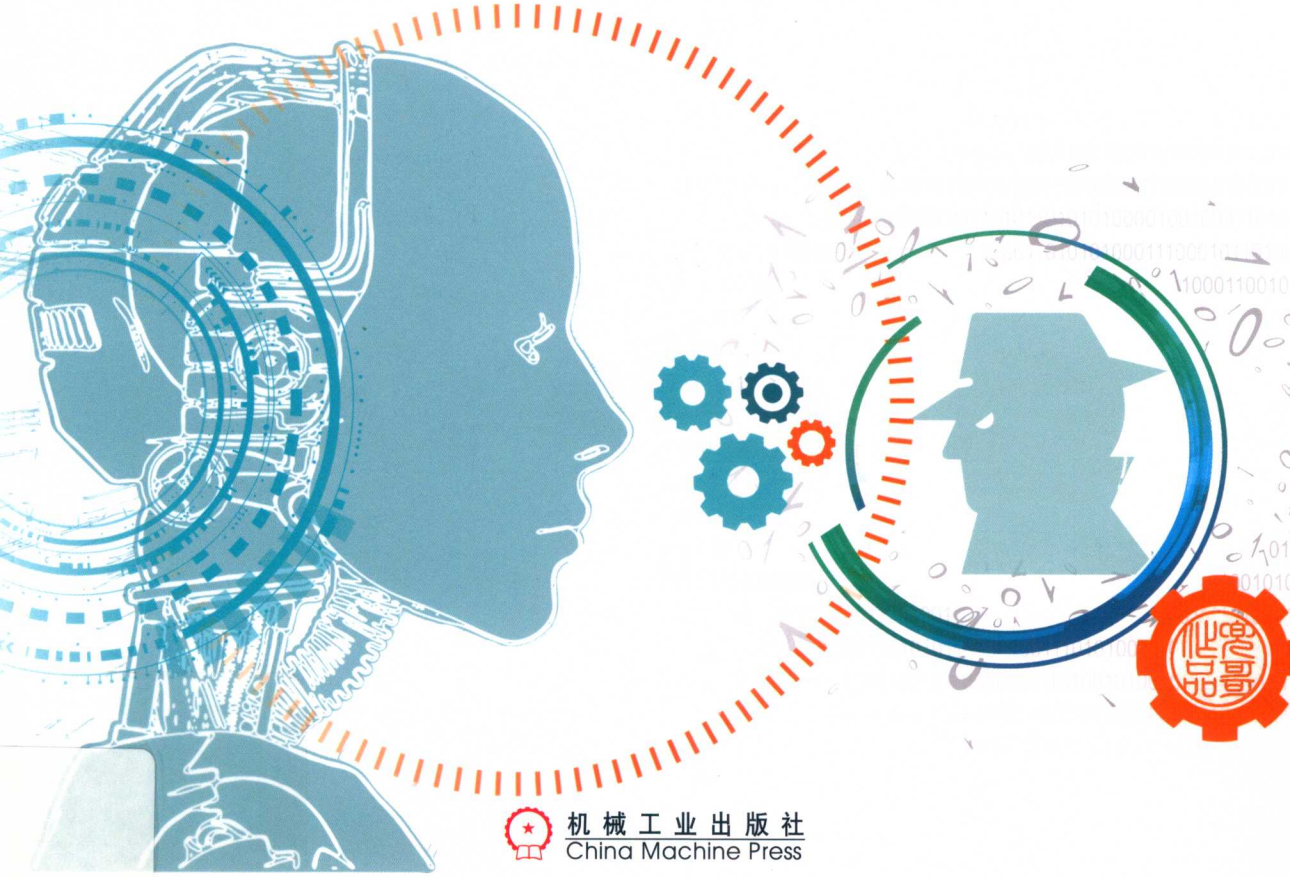
对抗样本的入门知识，AI安全必备

四色印刷

Introduction to Adversarial Examples

AI安全之对抗样本入门

兜哥 编著



机械工业出版社
China Machine Press

智能系统与技术丛书

Introduction to Adversarial Examples

AI安全之对抗样本入门

兜哥 编著

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

AI 安全之对抗样本入门 / 兜哥编著 . —北京：机械工业出版社，2019.5
（智能系统与技术丛书）

ISBN 978-7-111-62682-4

I. A… II. 兜… III. 人工智能 – 研究 IV. TP18

中国版本图书馆 CIP 数据核字（2019）第 087188 号

AI 安全之对抗样本入门

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：殷　虹
印　　刷：中国电影出版社印刷厂　　版　　次：2019 年 6 月第 1 版第 1 次印刷
开　　本：186mm × 240mm　1/16　　印　　张：19
书　　号：ISBN 978-7-111-62682-4　　定　　价：129.00 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　　读者信箱：hzit@hzbook.com

PRAISE

对本书的赞誉

机器学习在安全领域的应用越来越广泛，特别是近几年来，深度学习在安全漏洞检测、Web 应用防火墙、病毒检测等领域都有工业级的落地应用；但是黑客和黑产相应的入侵手法也发生了变化，其中一个手段就是从之前尝试绕过深度学习模型，变为攻击深度学习模型本身。兜哥在人工智能安全领域的实战和学术造诣深厚，本书从深度学习自身的脆弱性和遭受的一些攻击场景入手，讨论了如何加固深度学习模型和防范类似的攻击，对企业的安全工程师和从事安全人工智能的同仁，都有很好的指导和借鉴意义。

——施亮，顶象技术首席科学家 & 合伙人

人工智能已经被证明在越来越多的细分领域达到甚至超过了人类的平均水平，中国、美国、俄罗斯等许多国家也把发展人工智能提升到国家战略层面。人们在大力发展人工智能的同时，对于人工智能自身的安全问题的研究却相对滞后，这将严重制约其在重要领域的应用。兜哥的这本书很好地介绍了 AI 安全领域非常基础且重要的对抗样本的基本原理，帮助大家了解人工智能自身的安全问题，以便开发出更加安全的 AI 应用。

——胡影博士，中国电子技术研究院信息安全研究中心数据安全部主任

认识兜哥是从他的著作《企业安全建设入门》开始，在传统行业基于稳定性不断对商业软件进行深度改造时，他描述了互联网如何将开源用到了输出阶段。当传统安全遇到 ABCD 的时候，兜哥选择沉下来做 AI 工程师并分享了大数据实践下的核心对抗样本调参思路，为他的工匠精神和分享精神点赞。

——吕毅博士，中国人民银行金融信息中心信息安全部副主任

本书结合了作者在安全领域的多年实践经验，对对抗样本分析所面临的挑战进行了系统阐述，对业界常见的方式方法做了系统的归纳总结，有其独到的见解和主张。与其他机器学习系列丛书中的内容不同，本书针对的对象是人工智能本身，从对抗样本这一

维度入手，深入浅出地叙述了对抗样本的基本原理、攻击方式和常见防御算法等内容。本书对信息安全和人工智能的从业者来说，都具有一流的参考价值。

——王亿韬，北美互联网金融公司 Affirm 信息安全主管，CISSP/CSSLP/OSCP

作者的系列书——AI 安全三部曲，去年我已一一拜读，每本书都深入浅出，层层递进，读后大呼过瘾。前时应兜哥之邀为本书写荐语，欣然应允，不料今年工作繁忙拖了很久，甚为汗颜。近期终能挤出时间仔细研读。本书秉承兜哥的一贯风格，虽然锚定的是非常前沿的课题，仔细去看，依旧浅显易懂，分攻防两端，递进列举各类算法，并暖心提供基于不同 AI 框架的实现和评估。大巧若拙，递进的脉络在读者脑中自然形成，读后必然深受启发。我在此感谢兜哥勤奋执笔，又让我有先阅之乐。

——王新刚，北美安全公司 Shape Security 数据平台负责人

PREFACE

序　一

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。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

PREFACE

序　二

兜哥写的“AI+ 安全”一系列书，每本我都让朋友从国内带来硅谷拜读。我和 Lenx 等大佬混迹硅谷安全圈，一致很钦佩百度安全实验室对于“AI+ 安全”领域的贡献。

执笔为这本书写序的这天（2019 年 1 月 25 日）恰逢人类 AI 历史又一伟大时刻：AlphaStar 战胜了 StarCraft2 人类高手玩家 MaNa。我对这次战役的关注，甚至超过了 AlphaGo。StarCraft 即时战略游戏比围棋这种规则明确、棋盘信息透明的游戏，更加接近人类生存的环境：不完备的环境信息。安全领域一直是不完备的环境信息，在我看来，这意味着道高一尺魔高一丈强对抗的网络安全行业，也即将迎来一波新的方法论。

我从 2012 年起在美国硅谷的安全行业从事“AI+ 大数据”在安全和风控的企业级应用，有幸在 EMC/ RSA、FireEye / Mandiant 等公司，与世界级的白帽一起成功地使用 AI 技术防御过若干个 APT 黑客军团（比如 2017 年在 FireEye 开发了 NLP 算法抵抗了越南 APT32 ）。最近刚创办了 AnChain.ai 公司，志在将 AI 用于区块链安全，并在历史上首次检测到了 BAPT（Blockchain APT）黑客军团。

我也有幸直接参与了几次和黑客正面交锋的案例，比如 2018 年的 BAPT-FOMO3D 黑客，他们攻击 FOMO3D 游戏 DApp，居然启用了上万个攻击合约，短时间内盗走了两百万美元，而项目方束手无策。

“战略上藐视敌人，战术上重视敌人。”作为安全行业的多年从业者，我坚信我们可以利用技术的力量形成战略的优势。

对于广大安全产品研发人员，我建议在“战术上重视敌人”，学习 AI 技术将如虎添翼。增强学习、对抗学习、深度学习等领域近年来涌现了许多新算法，并且在 ASIC、GPU 这些强大的芯片基础上，实现了高效的人工智能落地应用，比如 AlphaStar、AlphaGO 和硅谷满地跑的无人驾驶车。

我建议大家仔细研读兜哥的这本新书，因为黑客也在学习进化。

——Victor Fang 博士，AnChain.ai 创始人和 CEO

PREFACE

自　序

这不是我第一次写序了，但是真拿起笔还是觉得挺难写的。记得我写第一本书时，女儿还在蹒跚学步，拿着我的书时还分不清书的前后正反，但是现在她已经在幼儿园学会做月饼了。据说现在人工智能的知识已经写入高中教材了，这方面我倒是对闺女比较有信心，至少在人工智能方面她启蒙得比较早。

机缘巧合，我负责了一个AI模型安全的开源项目AdvBox[⊖]。虽然我之前接触过AI安全，但是主要集中在AI赋能传统安全的领域，简单讲就是我之前写的《Web安全之机器学习入门》这类书里介绍的如何将AI技术应用到传统安全领域。相对传统安全领域，AI模型的安全是一个陌生但更加有趣的领域。从简单的把猫变成狗，到稍微复杂点的用奇怪的“鸟叫虫鸣”就可以唤醒智能音箱、点外卖，再到更加酷炫的换脸和欺骗无人车，这都是对抗样本的典型应用。开发和运营AdvBox是我工作的一部分，也是我生活的一部分，我写了个小脚本每天爬GitHub上的排名并发送邮件给我，每天早上我都会刷刷邮箱看一下最新排名。AdvBox的关注数也从零一直慢慢爬到业内关注的一个排名。为了更好地运营AdvBox，我组建了QQ群和微信群，在回答用户问题的过程中，我发现其实很多问题非常基础：从TensorFlow/PyTorch的使用到深度学习的基础知识，从对抗样本的基本原理到图像处理中的一些小技巧。于是我想到了可以写一本书来介绍这方面的基础知识。

很多人在网上抱怨，2018年是比较艰难的一年，我也有同样的感受。工作繁忙了许多，而且换了新环境，许多东西需要再去适应。这本书从构思到基本写完，花了将近一年的时间，几乎占据了我全部的休息时间，确实挺不容易。其实我也反思过几次，为啥不去做些轻车熟路的事，非要去折腾些BIG4的论文里介绍的东西，非要在一个新成果刚

⊖ https://github.com/baidu/AdvBox

发布就可能要过时的赛道去追逐。如果非要找个理由的话，我个人的感受是：容易的事轮不到我，早已是千军万马过独木桥了；困难的事坚持做，没准就走下去了。

本书的定位是学习对抗样本的入门书籍，因此也简单介绍了相关领域的背景知识便于读者理解，使用的数据集和场景是比较典型的图像分类、目标识别等领域。第 1 章介绍了深度学习的基础知识，重点介绍了与对抗样本相关的梯度、优化器、反向传递等知识点。第 2 章介绍了如何搭建学习对抗样本的软硬件环境，虽然 GPU 不是必需的，但是使用 GPU 可以更加快速地验证你的想法。第 3 章概括介绍了常见的深度学习框架，从 TensorFlow、Keras、PyTorch 到 MXNet。第 4 章介绍了图像处理领域的基础知识，这部分知识对于理解对抗样本领域的一些常见图像处理技巧非常有帮助。第 5 章介绍了常见的白盒攻击算法，从最基础的 FGSM、DeepFool 到经典的 JSMA 和 CW。第 6 章介绍了常见的黑盒攻击算法。第 7 章介绍了对抗样本在目标识别领域的应用。第 8 章介绍了对抗样本的常见抵御算法，与对抗样本一样，抵御对抗样本的技术也非常有趣。第 9 章介绍了常见的对抗样本工具以及如何搭建 NIPS 2017 对抗防御环境和轻量级攻防对抗环境 robust-ml，通过这章读者可以了解如何站在巨人的肩膀上，快速生成自己的对抗样本，进行攻防对抗。

本书适合人工智能领域的从业人员、大专院校计算机相关专业学生，而不仅仅是信息安全领域的学生和从业人员。因为对抗样本的知识对于人工智能领域的从业人员非常重要，伴随着人工智能的遍地开花和逐步落地，从智能驾驶到人脸支付，从智能家居到智能安防，人工智能从一个学术名词变成了真正影响大家生活的技术，直接关系到大家的人身安全、财产安全还有个人隐私。对抗样本带来的问题，也是传统安全技术几乎难以解决的。了解对抗样本的基本原理，对于人工智能领域的从业人员开发出更加安全的应用是非常有帮助的。

我要感谢家人对我的支持，本来工作就很忙，没有太多时间处理家务，写书以后更是侵占了我大量的休息时间，我的妻子无条件地承担起了全部家务，尤其是照料孩子。我还要感谢我的女儿，写书这段时间几乎没有时间陪她玩，我也想用这本书作为她的生日礼物。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。Lenx 对于我写这本书帮助很大，他对 AI 安全的深刻理解，积极但又务实，让我在研究的道路上既信心满满又不至于过于狂热而迷失道路。还要感谢 AdvBox 团队以及在 GitHub 上给 AdvBox 提交过代码的同学们。

最后还要感谢各位业内好友对我的支持，以下排名不分先后：

马杰 @ 百度安全、Lenx@ 百度安全、黄正 @ 百度安全、包沉浮 @ 百度安全、海棠姐 @ 百度安全、Edward@ 百度安全、贾云瀚 @ 百度安全、云鹏 @ 百度无人车、施亮 @ 顶象、Victor Fang@AnChain、谢忱 @Freebuf、大路 @ 天际友盟、郭伟 @ 数字观星、周涛 @ 阿里、姚志武 @ 借贷宝、刘静 @ 安天、高磊 @ 阿里、尹毅 @sobug、吴圣 @58、

康宇@新浪、幻泉@i春秋、田老师@阳光保险、ReadOnly@易宝支付、樊春亮@泰康、聂君@360、林伟@360、白教主@360、李滨@腾讯、张维垚@饿了吗、阿杜@优信、高磊@安巽科技、王延辉@平安、吕毅@人行、雷诚@武汉大学、鸟哥@阿里云、咸鱼@京东、梁知音@京东、小马哥@京东、张超@清华大学、徐恪@清华大学、李勇@清华大学、李琦@清华大学。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。

计算机是一门非常强调理论联系实践的学科，我经常在和读者交流时说：安全会议的PPT就像电影，浓缩、精彩、有趣；博客、公众号上的文章就像电视剧，有细节、有思路；技术书籍更像原著，系统、详细、原汁原味。但是就像看再多武侠小说也不一定真能耍出一招半式一样，真正掌握一门武艺还是要勤于练习。因此我在GitHub上也把书里提到的示例代码开源出来，读者可以根据情况一边修改一边验证自己的理解，对应的地址为：https://github.com/duoergun0729/adversarial_examples。

PREFACE

前　言

生活中的深度学习

深度学习自 2006 年产生之后就受到科研机构、工业界的高度关注。最初，深度学习主要用于图像和语音领域。从 2011 年开始，谷歌研究院和微软研究院的研究人员先后将深度学习应用到语音识别，使识别错误率下降了 20% ~ 30%[⊖]。2012 年 6 月，谷歌首席架构师 Jeff Dean 和斯坦福大学教授 Andrew Ng 主导著名的 Google Brain 项目，采用 16 万个 CPU 来构建一个深层神经网络，并将其应用于图像和语音的识别，最终大获成功。

2016 年 3 月，AlphaGo 与围棋世界冠军、职业九段棋手李世石进行围棋人机大战，以 4 比 1 的总比分获胜；2016 年年末 2017 年年初，该程序在中国棋类网站上以“大师”（Master）为注册账号与中日韩数十位围棋高手进行快棋对决，连续 60 局无一败绩；2017 年 5 月，在中国乌镇围棋峰会上，它与排名世界第一的围棋世界冠军柯洁对战，以 3 比 0 的总比分获胜。AlphaGo 的成功更是把深度学习的热潮推向了全球，成为男女老少茶余饭后关注的热点话题。

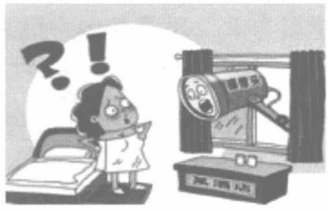
现在，深度学习已经遍地开花，在方方面面影响和改变着人们的生活，比较典型的应用包括智能家居、智能驾驶、人脸支付和智能安防。

深度学习的脆弱性

深度学习作为一个非常复杂的软件系统，同样会面对各种黑客攻击。黑客通过攻击深度学习系统，也可以威胁到财产安全、个人隐私、交通安全和公共安全（见图 0-1）。针对深度学习系统的攻击，通常包括以下几种。

⊖ G Dahl, D Yu, L Deng. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012,20(1):30-42.

财产安全



个人隐私

交通安全

公共安全

图 0-1　深度学习的脆弱性

1. 偷取模型

各大公司通过高薪聘请 AI 专家设计模型，花费大量资金、人力搜集训练数据，又花费大量资金购买 GPU 设备用于训练模型，最后得到深度学习模型。深度学习模型的最终形式也就是从几百 KB 到几百 MB 不等的一个模型文件。深度学习模型对外提供服务的形式也主要分为云模式的 API，或者私有部署到用户的移动设备或数据中心的服务器上。针对云模式的 API，黑客通过一定的遍历算法，在调用云模式的 API 后，可以在本地还原出一个与原始模型功能相同或者类似的模型[⊖]；针对私有部署到用户的移动设备或数据中心的服务器上，黑客通过逆向等传统安全技术，可以把模型文件直接还原出来供其使用。偷取深度学习模型的过程如图 0-2 所示。

2. 数据投毒

针对深度学习的数据投毒主要是指向深度学习的训练样本中加入异常数据，导致模型在遇到某些条件时会产生分类错误。如图 0-3 所示。早期的数据投毒都存在于实验室环境，假设可以通过在离线训练数据中添加精心构造的异常数据进行攻击。这一攻击方式需要接触到模型的训练数据，而在实际环境中，绝大多数情况都是公司内部在离线数

⊖ Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, Thomas Ristenpart.Stealing Machine Learning Models via Prediction APIs [C]. USENIX Security, 2016.

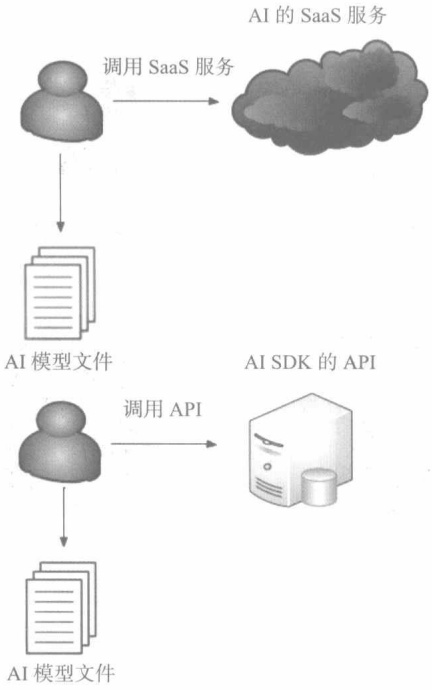


图 0-2　偷取深度学习模型

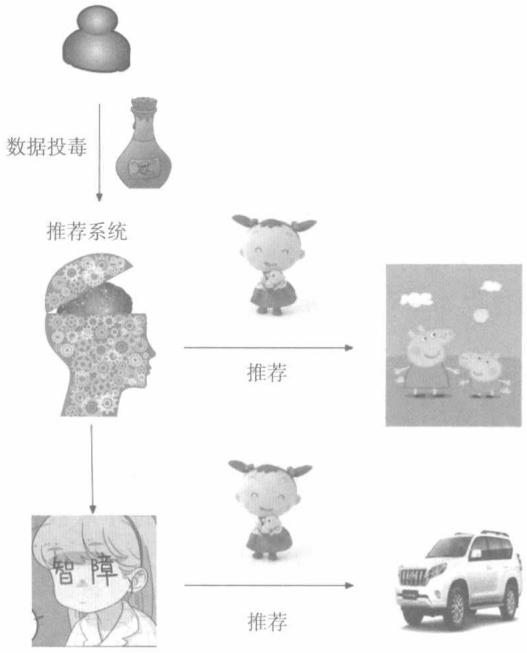


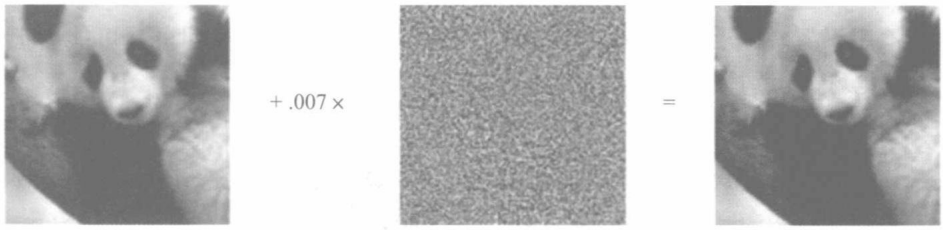
图 0-3　针对深度学习的数据投毒

据中训练好模型再打包对外发布服务，攻击者难以接触到训练数据，攻击难以发生。于是攻击者把重点放到了在线学习的场景，即模型是利用在线的数据，几乎是实时学习的，比较典型的场景就是推荐系统。推荐系统会结合用户的历史数据以及实时的访问数据，共同进行学习和判断，最终得到推荐结果。黑客正是利用这一可以接触到训练数据的机会，通过一定的算法策略，发起访问行为，最终导致推荐系统产生错误。

3. 对抗样本

对抗样本由 Christian Szegedy 等人提出，是指在数据集中通过故意添加细微的干扰所形成的输入样本，这种样本导致模型以高置信度给出一个错误的输出。在正则化背景下，通过对抗训练减少原有独立同分布的测试集的错误率，在对抗扰动的训练集样本上训练网络[⊖]。

简单地讲，对抗样本通过在原始数据上叠加精心构造的人类难以察觉的扰动，使深度学习模型产生分类错误。以图像分类模型为例，如图 0-4 所示，通过在原始图像上叠加扰动，对于肉眼来说，扰动非常细微，图像看起来还是熊猫，但是图像分类模型却会以很大的概率识别为长臂猿。



x
“熊猫”
57.7% 的置信度

$\text{sign}(\nabla_x J(\theta, x, y))$
“线虫”
8.2% 的置信度

$x+$
$\varepsilon\text{sign}(\nabla_x J(\theta, x, y))$
“长臂猿”
99.3% 的置信度

图 0-4　针对图像分类模型的对抗样本[⊜]

下面以一个图像分类模型为例，更加直接地解释对抗样本的基本原理。通过在训练样本上学习，学到一个分割平面，在分割平面一侧的为绿球，在分割平面另外一侧的为红球。生成攻击样本的过程，就是在数据上添加一定的扰动，让其跨越分割平面，从而把分割平面一侧的红球识别为绿球，如图 0-5 所示。

⊖ Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[J]. Computer Science, 2013.

⊜ 图片源于文献 Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy, Explaining and Harnessing Adversarial Examples, ICLR 2015。

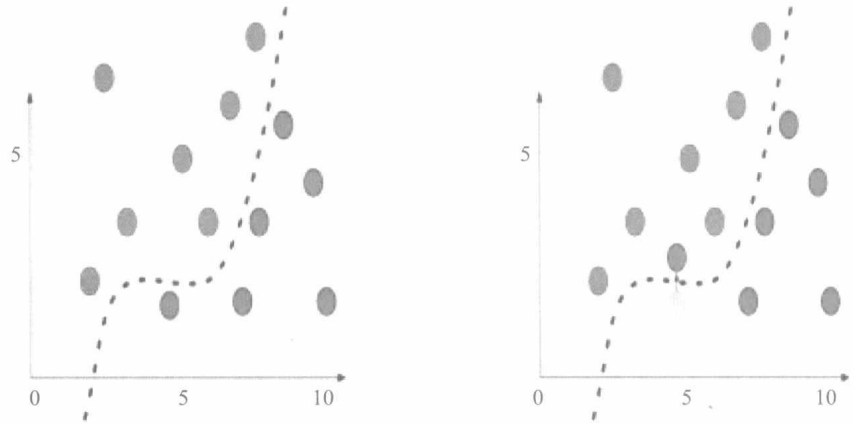


图 0-5　对抗样本的基本原理

对抗样本按照攻击后的效果分为 Targeted Attack（定性攻击）和 Non-Targeted Attack（无定向攻击）。区别在于 Targeted Attack 在攻击前会设置攻击的目标，比如把红球识别为绿球，或者把面包识别为熊猫，也就是说在攻击后的效果是确定的；Non-Targeted Attack 在攻击前不用设置攻击目标，只要攻击后，识别的结果发生改变即可，可能会把面包识别为熊猫，也可能识别为小猪佩琪或者小猪乔治，如图 0-6 所示。

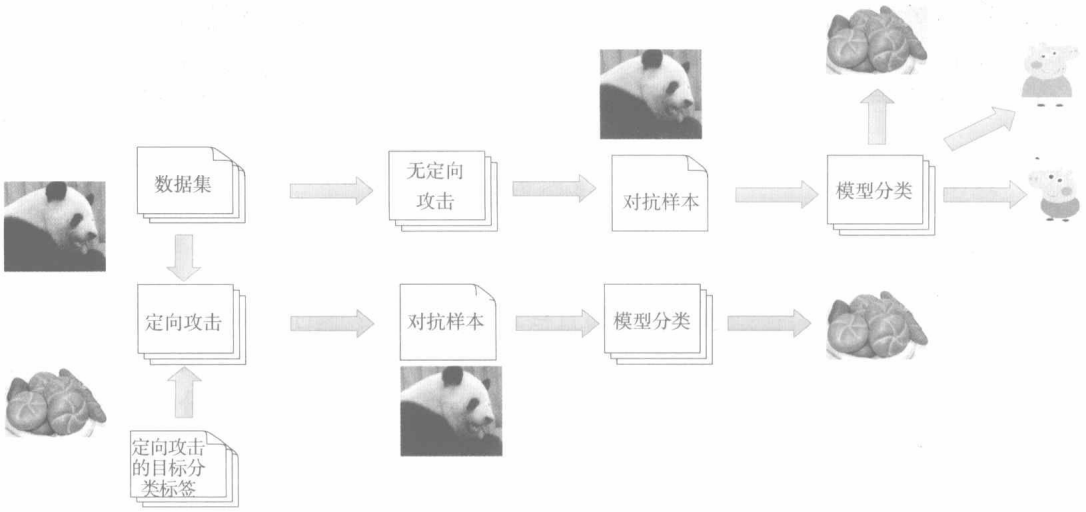


图 0-6　Targeted Attack 和 Non-Targeted Attack

对抗样本按照攻击成本分为 White-Box Attack(白盒攻击)、Black-Box Attack(黑盒攻击)和 Real-World Attack/Physical Attack（真实世界 / 物理攻击）。

White-Box Attack（见图 0-7）是其中攻击难度最低的一种，前提是能够完整获取模型的结构，包括模型的组成以及隔层的参数情况，并且可以完整控制模型的输入，对输入的控制粒度甚至可以到比特级别。由于 White-Box Attack 前置条件过于苛刻，通常作为实验室的学术研究或者作为发起 Black-Box Attack 和 Real-World Attack/Physical Attack 的基础。

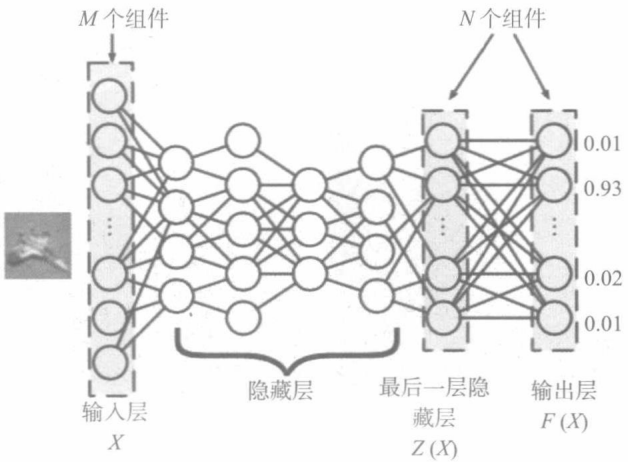


图 0-7　White-Box Attack

Black-Box Attack 相对 White-Box Attack 攻击难度具有很大提高，Black-Box Attack 完全把被攻击模型当成一个黑盒，对模型的结构没有了解，只能控制输入，通过比对输入和输出的反馈来进行下一步攻击，见图 0-8。

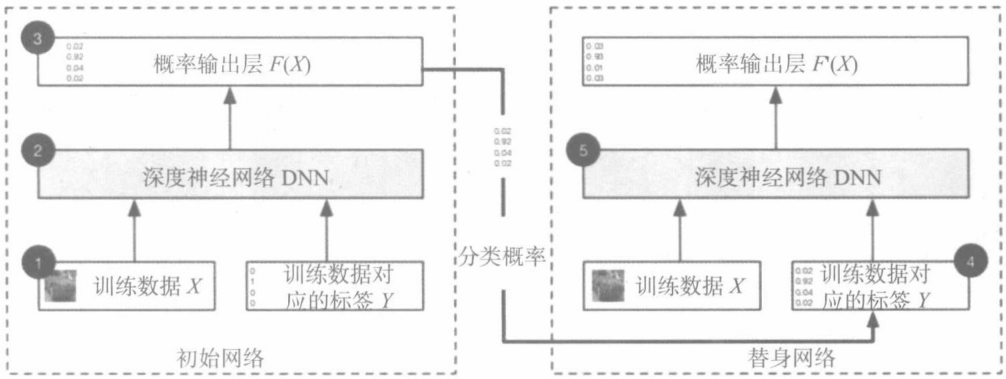


图 0-8　Black-Box Attack

Real-World Attack/Physical Attack（见图 0-9）是这三种攻击中难度最大的，除了不了解模型的结构，甚至对于输入的控制也很弱。以攻击图像分类模型为例（见图 0-10），生成的攻击样本要通过相机或者摄像头采集，然后经过一系列未知的预处理后再输入模型进行预测。攻击中对抗样本会发生缩放、扭转、光照变化、旋转等。

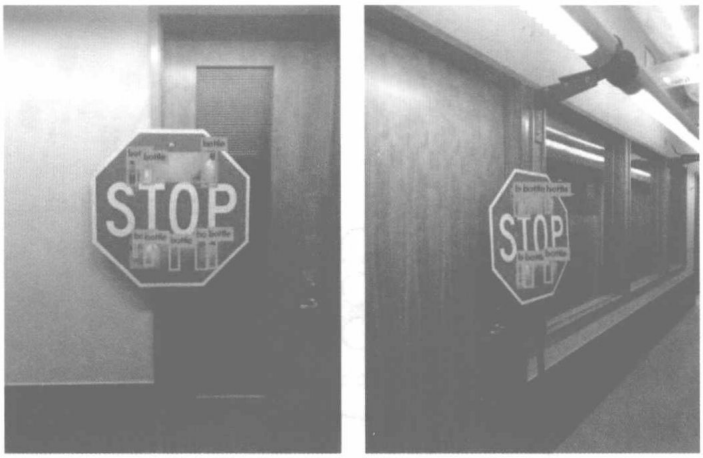


图 0-9　Real-World Attack/Physical Attack

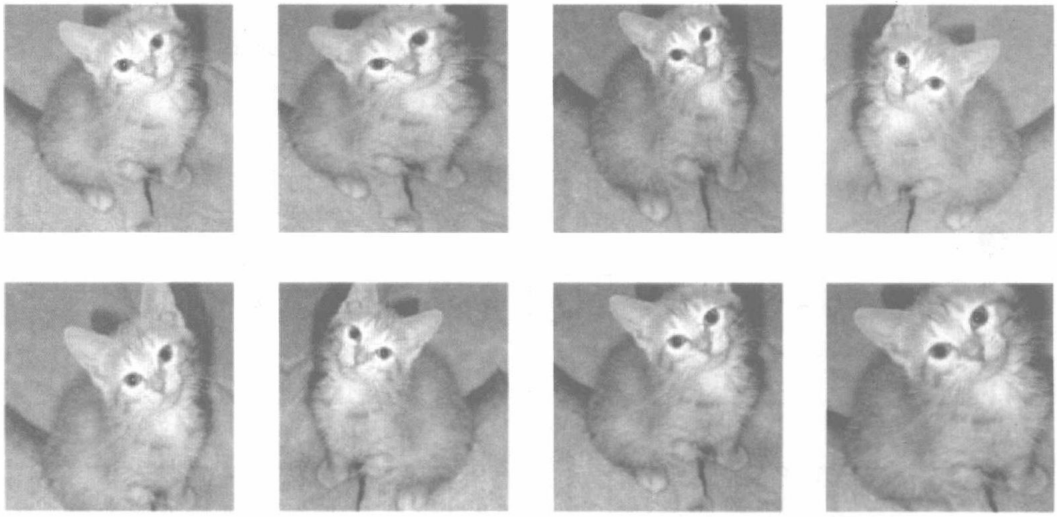


图 0-10　图像分类模型的真实世界 / 物理攻击