HZ BOOKS
华章IT

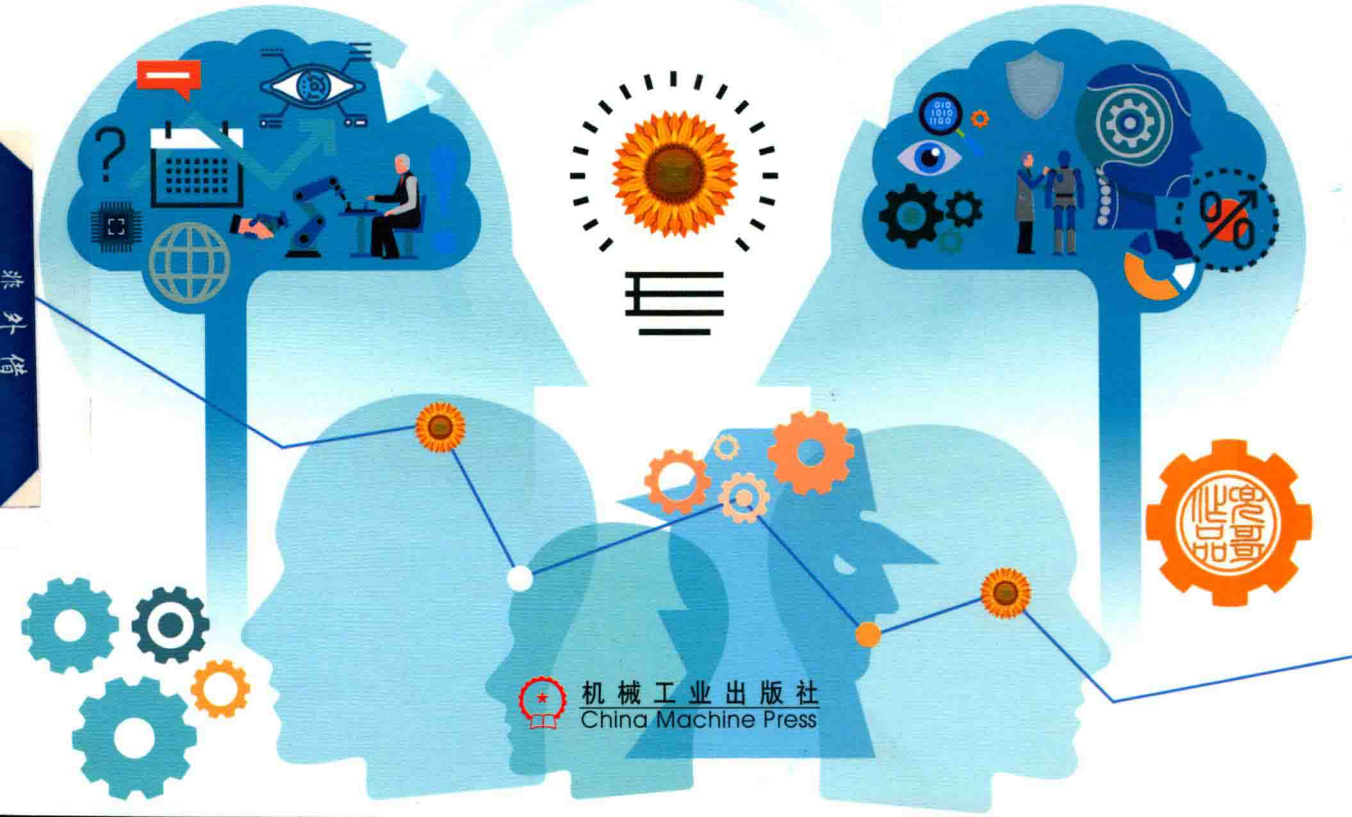
智能系统与技术丛书

百度安全专家撰写，零基础学习智能化Web安全技术，AI+安全

Deap Learning for the Web Security

Web安全之深度学习实战

刘焱 编著



机械工业出版社
China Machine Press

Web深度学习实战

智能系统与技术丛书

Deap Learning for the Web Security

Web安全之深度学习实战

刘焱 编著

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

Web 安全之深度学习实战 / 刘焱编著 . —北京：机械工业出版社，2017.10
（智能系统与技术丛书）

ISBN 978-7-111-58447-6

I. W… II. 刘… III. 机器学习—安全技术 IV. TP181

中国版本图书馆 CIP 数据核字（2017）第 279091 号

Web 安全之深度学习实战

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：吴 怡
印 刷：北京市荣盛彩色印刷有限公司
开 本：186mm × 240mm 1/16
书 号：ISBN 978-7-111-58447-6
版 次：2018 年 1 月第 1 版第 1 次印刷
印 张：16.5
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
购书热线：（010）68326294 88379649 68995259
投稿热线：（010）88379604
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

FOREWORD

序

当前正是一个技术变革引发各行业变革的时代，云计算、大数据、物联网等新技术在为各行业带来新机遇的同时，也带来了新的安全漏洞和风险。新的风险当然就需要新的网络安全技术来针对解决，机器学习技术无疑是新一代网络安全技术中的佼佼者。新一代网络安全产品和解决方案已经普遍将机器学习当做必要功能来进行开发，机器学习的概念如何落实于产品，落实于业务，落实在具体场景中，已成为国内网络安全产业普遍关注和探索的问题。本书是目前为止我所阅读过的国内关于机器学习应用于网络安全应用领域最详细和实用的参考读物。作者基于一线实际业务实践，不仅深入介绍了深度学习的相关算法和开源技术，而且提供了一系列有关机器学习的业务场景的具体案例。所以强烈推荐所有网络安全应用的从业人员和对机器学习领域感兴趣的爱好者们人手一本！

金湘宇，君源创投管理合伙人

2017 年 9 月于北京

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。但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级增长，各类新型攻击方式层出不穷，0day 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；如果让机器可以像阿尔法狗理解围棋一样理解网络威胁，那将是巨大进步。事情又回到最初的那个问题，如何能让机器可以真正学会识别安全威胁？机器学习可能是一个不错的答案。

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、架构师同样也是一本不错的科普书籍。如果读者看完本书，在工作学习中遇到问题时可以想起一到两种算法，那么我觉得就达到效果了；如果读者读完本书，可以像使用 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。正如霍金所言“多写 1 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确表现差别很大，很难说深度学习就一定比朴素贝叶斯好，也很难说支

持向量机就比不过卷积神经网络，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。

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，介绍了本书使用的 TensorFlow、TFLearn 等深度学习库的安装以及使用方法。第 2 章和第 3 章介绍了卷积神经网络和循环神经网络这两大深度学习算法的基础知识。第 4 章介绍在生产环境搭建机器学习平台需要使用的开源组件，包括 Logstash、Kafka、Storm、Spark 等，并且介绍了 GPU 和 TPU 的基础知识。第 5 章到第 15 章，介绍了 11 个使用机器学习技术解决实际安全问题的案例，包括验证码识别、垃圾邮件识别、负面评论识别、骚扰短信识别、Linux 后门检测、用户行为分析与恶意行为检测、WebShell 检测、智能扫描器、DGA 域名检测、恶意程序分类识别、反信用卡欺诈，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。

本书是我所著机器学习三部曲的第二部，第一部主要以机器学习常见算法为主线，利用生活中的例子和具体安全场景来介绍机器学习常见算法，是机器学习入门书籍，便于读者可以快速上手。全部代码都能在普通 PC 上运行。本书将重点介绍深度学习，并以具体的 11 个案例介绍机器学习的应用，定位是面向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，本书将重点放在问题的解决而不是算法的介绍。由于深度学习通常计算量已经超过了 PC 的能力，部分代码需要在服务器甚至 GPU 上运行，不过这不影响大家的阅读与学习。第三部将重点介绍强化学习和对抗网络，并利用若干虚构安全产品或者项目来介绍如何让机器真正具备阿尔法狗级别的智能。遗憾的是，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，这些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，尤其是照料孩子等繁杂事务。我很感谢我的女儿，写书这段时间几乎没有时间陪她玩，她很懂事地自己玩，我想用这本书作为她的生日礼物送给她。我还要感谢吴怡编辑对我的支持和鼓励，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、阿文 @ 丁

牛、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、王珉然@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、刘袁君@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、钱华钧、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”、知识星球（原名小密圈）“Web安全之机器学习”，欢迎大家关注并在线交流。之前没有使用过知识星球的读者可以在各类应用市场上搜索。

本书使用的代码和数据均在GitHub上发布，地址为：https://github.com/duoergun0729/2book，代码层面任何疑问可以在GitHub上直接反馈。

CONTENTS

目　录

CHAPTER 1

第 1 章

打造深度学习工具箱

在本系列图书的第一本《Web 安全之机器学习入门》中，我们以常见安全问题为背景介绍了常见的机器学习算法，主要以 KNN、SVM、朴素贝叶斯等浅层学习算法为主。以 Scikit-Learn 为代表的机器学习开发库帮助我们可以很便捷地在单机环境下验证我们的想法。近几年深度学习发展迅速，以 TensorFlow 为代表的一批优秀的深度学习开发库大大降低了大家学习使用深度学习技术的门槛。作为本书的第 1 章，本章将帮助大家打造自己的深度学习工具箱，并结合实际例子介绍 TensorFlow、TFLearn、PaddlePaddle 以及 Karas 的使用方法。

本章代码请参考本书配套 GitHub 中的 tools.py。

1.1 TensorFlow

TensorFlow 是谷歌的第二代人工智能学习系统，其名称来源于本身的运行原理。Tensor 意味着 N 维数组，Flow 意味着基于数据流图的计算，TensorFlow 为 Tensor 从流图的一端流动到另一端计算过程。所以也可以把 TensorFlow 当做将复杂的数据结构传输至人工智能神经网中进行分析和处理的系统。

TensorFlow 可用于语音识别或图像识别等多项机器深度学习领域，是对 2011 年开发的深度学习基础架构 DistBelief 进行了各方面的改进，它可在小到一部智能手机、大到数千台数据中心服务器的各种设备上运行。

1.1.1 安装

TensorFlow 支持非常丰富的安装方式[⊖]。

⊖ http://www.tensorfly.cn/tfdoc/get_started/os_setup.html

1. Ubuntu/Linux。

```
# 仅使用 CPU 的版本
$ pip install https://storage.googleapis.com/tensorflow/linux/cpu/tensorflow-
    0.5.0-cp27-none-linux_x86_64.whl
# 开启 GPU 支持的版本 (安装该版本的前提是已经安装了 CUDA sdk)
$ pip install https://storage.googleapis.com/tensorflow/linux/gpu/tensorflow-
    0.5.0-cp27-none-linux_x86_64.whl
```

2. Mac OS X。

在 Mac OS X 系统上，我们推荐先安装 homebrew，然后执行 brew install python，以便能够使用 homebrew 中的 Python 安装 TensorFlow：

```
# 当前版本只支持 CPU
$ pip install https://storage.googleapis.com/tensorflow/mac/tensorflow-0.5.0-
    py2-none-any.whl
```

3. 基于 Docker 的安装。

该命令将启动一个已经安装好 TensorFlow 及相关依赖的容器：

```
$ docker run -it b.gcr.io/tensorflow/tensorflow
```

4. 基于 VirtualEnv 的安装。

官方文档推荐使用 VirtualEnv 创建一个隔离的容器来安装 TensorFlow，这是可选的，但是这样做能使排查安装问题变得更容易。VirtualEnv 通过创建独立 Python 开发环境的工具，来解决依赖、版本以及间接权限问题。比如，一个项目依赖 Django1.3，而当前全局开发环境为 Django1.7，版本跨度过大，导致的不兼容使项目无法正常运行，使用 VirtualEnv 可以解决这些问题[⊖]。

首先，安装所有必备工具：

```
# 在 Linux 上：
$ sudo apt-get install python-pip python-dev python-virtualenv
# 在 Mac 上：
# 如果还没有安装 pip
$ sudo easy_install pip
$ sudo pip install --upgrade virtualenv
```

接下来，建立一个全新的 VirtualEnv 环境。为了将环境建在 ~/tensorflow 目录下，执行如下代码：

```
$ virtualenv --system-site-packages ~/tensorflow
$ cd ~/tensorflow
```

然后，激活 VirtualEnv：

⊖ http://www.jianshu.com/p/08c657bd34f1

```
$ source bin/activate
# 如果使用 bash $ source bin/activate.csh
# 如果使用 csh (tensorflow)$
# 终端提示符应该发生变化
```

在 VirtualEnv 内，安装 TensorFlow：

```
(tensorflow)$ pip install --upgrade <$url_to_binary.whl>
```

接下来，使用类似命令运行 TensorFlow 程序：

```
(tensorflow)$ cd tensorflow/models/image/mnist
(tensorflow)$ python convolutional.py
# 当使用完 TensorFlow
(tensorflow)$ deactivate
```

1.1.2 使用举例

TensorFlow 的 API 使用起来比较繁琐，通常直接使用针对其 API 的高层封装 TFLearn 即可，具体实现请参考下节 TFLearn 的使用举例。

1.2 TFLearn

TFLearn 是一个模块化和透明的深度学习库，构建在 TensorFlow 之上，它为 TensorFlow 提供高层次 API，目的是快速搭建试验环境，同时保持对 TensorFlow 的完全透明和兼容性。

TFLearn 具有以下特点[⊖]：

- 容易使用和易于理解的高层次 API 用于实现深度神经网络，附带教程和例子。
- 通过高度模块化的内置神经网络层、正则化器、优化器等进行快速原型设计。
- 强大的辅助函数，训练任意 TensorFlow 图，支持多输入、多输出和优化器。
- 简单而美观的图可视化，具有关于权值、梯度、特征图等细节。
- 无需人工干预，可使用多 CPU、多 GPU。

本书的主要代码都是基于 TFLearn 开发的。

1. 安装。

TFLearn 的安装推荐使用使用 pip 工具：

```
pip install tflearn
```

如果需要使用源码安装，可以直接从 GitHub 上下载对应源码：

⊖ https://zhuanlan.zhihu.com/p/25322066