密码学报 ISSN 2095-7025 CN 10-1195/TN
Journal of Cryptologic Research, 2021, 8(2): 202–214

E-mail: jcr@cacrnet.org.cn
http://www.jcr.cacrnet.org.cn
Tel/Fax: +86-10-82789618

针对深度学习的对抗攻击综述*

刘 会, 赵 波, 郭嘉宝, 彭钺峰

武汉大学 国家网络安全学院, 武汉 430072
通信作者: 赵波, E-mail: zhaobo@whu.edu.cn

摘 要: 深度学习在图像、文本、语音等多种数据类型的处理上取得了显著进展. 然而, 深度学习的不可解释性导致深度学习的输出缺乏可信性, 致使其在许多安全领域的应用受到了严重的制约. 研究人员发现通过对原始样本加入微小扰动所生成的对抗样本能够有效欺骗深度学习模型, 并将生成对抗样本的方式称之为对抗攻击. 对抗攻击能够使深度学习以高置信度的方式给出错误的输出, 实现针对深度学习检测服务的逃逸攻击. 本文首先介绍了对抗攻击的基本原理, 并从扰动范围、攻击者掌握目标模型知识的情况、攻击目标的针对性、攻击频次等 4 个方面对对抗攻击进行分类. 然后, 总结了近年来计算机视觉领域中对抗攻击研究的代表性成果, 对比分析各种攻击方案的特点. 特别针对对抗攻击在自然语言处理、语音识别、恶意软件检测和可解释性对抗样本等 4 种典型场景下的应用进行了详细介绍, 进一步揭示了对抗样本对深度学习服务的安全威胁. 最后, 通过回顾对抗攻击的发展历程, 探究该技术面临的主要挑战并指出其未来潜在的发展方向.

关键词: 深度学习; 对抗样本; 对抗攻击; 逃逸攻击; 计算机视觉

中图分类号: TP309.7 **文献标识码:** A DOI: 10.13868/j.cnki.jcr.000431

中文引用格式: 刘会, 赵波, 郭嘉宝, 彭钺峰. 针对深度学习的对抗攻击综述[J]. 密码学报, 2021, 8(2): 202–214. [DOI: 10.13868/j.cnki.jcr.000431]

英文引用格式: LIU H, ZHAO B, GUO J B, PENG Y F. Survey on adversarial attacks towards deep learning[J]. Journal of Cryptologic Research, 2021, 8(2): 202–214. [DOI: 10.13868/j.cnki.jcr.000431]

Survey on Adversarial Attacks Towards Deep Learning

LIU Hui, ZHAO Bo, GUO Jia-Bao, PENG Yue-Feng

School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China
Corresponding author: ZHAO Bo, E-mail: zhaobo@whu.edu.cn

Abstract: Deep learning has made a significant progress in the processing of images, text, voice and other types of data. However, the uninterpretability of deep learning leads to the lack of credibility of its output, which severely limits its applications in security-sensitive systems. Researchers found that the adversarial sample generated by adding small perturbations to the original sample could effectively deceive the deep learning model. The way of generating the adversarial sample is called adversarial attack. Adversarial attacks can make the deep learning model give wrong output in a high confidence, and realize escape attacks on detection services based on deep learning. This paper first introduces the basic principle of adversarial attacks, and gives the taxonomy of attack schemes

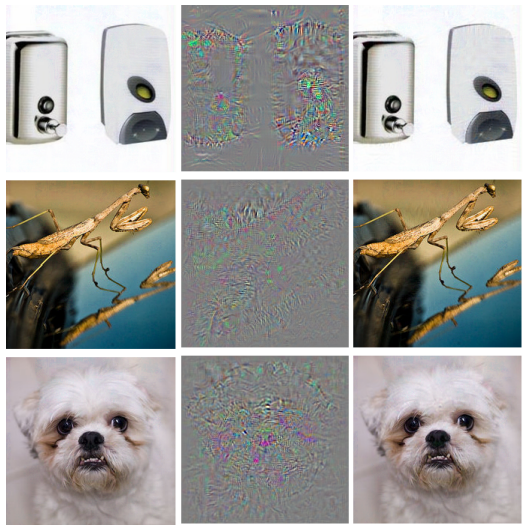
* 基金项目: 国家自然科学基金 (U1936122)
Foundation: National Natural Science Foundation of China (U1936122)
收稿日期: 2020-04-20 定稿日期: 2020-12-07

according to perturbation scope, adversary's knowledge, adversarial specificity and attack frequency. Then, the characteristics of various attack schemes in computer vision are compared and analyzed by summarizing the representative research achievements on adversarial attacks in recent years. In particular, the applications of adversarial attacks in natural language processing, speech recognition, malicious software detection and interpretable adversarial examples are introduced in detail, which further reveal the security threat of the adversarial example towards deep learning service. Finally, by reviewing the development of adversarial attacks, this paper discusses their major challenges and points out the potential development directions.

Key words: deep learning; adversarial example; adversarial attack; escape attacks; computer vision

1　引言

深度学习被广泛应用于计算机视觉、自然语言处理、语音识别等多个领域并取得了重大突破. 尤其在图像识别和图像分类的任务中, 深度学习具备非常高的准确度, 甚至表现出了超越人类的工作能力. 然而, 即使深度神经网络通过模拟人类大脑神经网络结构取得了显著的效果, 但是深度神经网络的理解方式与人类认知仍然存在较大差异, 深度学习的工作原理缺乏可解释性, 其输出结果的可信性难以得到有效的保障. 深度学习缺乏可解释性由其自身结构和运行机理决定, 具体表现在以下 3 个方面: (1) 深度神经网络的神经元数量大、参数多; (2) 神经网络采用分层结构, 层次之间连接方式多样; (3) 神经网络自主学习样本特征, 而许多特征人类无法理解. 在探索深度学习的可解释性、揭示深度学习的工作原理的过程中, Szegedy 等人 [1] 发现深度神经网络对加入特定扰动的图像样本表现出极高的脆弱性, 并将这种带有对抗扰动的样本称之为 "对抗样本"(见图 1).



(a) 原始图像　(b) 对抗扰动　(c) 对抗样本

图 1　对抗样本生成示例

Figure 1　Adversarial example

在计算机视觉领域, 对抗样本现象是指对输入图像加入人眼视觉难以感知的轻微扰动, 导致基于深度学习的图像分类器以高置信度的方式输出错误的结果. 在数字水印领域 [2], 对抗扰动会影响嵌入模式的完整性, 使得水印无法检测, 数字媒体的真实性无法得到有效验证. Sharif [3] 通过优化方法计算扰动并打印到眼镜框上, 攻击者只需要佩戴这种定制的眼镜便可以成功欺骗人脸识别系统, 获得合法的访问权限. 对抗样本的研究对解释深度学习工作原理具有显著的意义, 同时也极大促进了基于深度学习安全攻防的发展.

本文从对抗攻击的基本原理出发, 重点调研了其在计算机视觉领域的关键技术和代表性成果, 特别探讨了对抗攻击在具体场景下的应用价值, 进一步揭示了对抗样本对深度学习的安全威胁. 通过对对抗攻击的发展历程进行梳理, 探究该技术面临的主要挑战, 并指出未来的发展前景. 本文的主要贡献如下:

(1) 系统分析了计算机视觉领域中对抗攻击的典型算法, 并按扰动范围、模型知识、攻击目标的针对性、攻击频次对对抗攻击进行分类, 提供这类算法的整体概述;

(2) 调研了对抗攻击在具体场景下的实际应用, 包括自然语言处理、语音识别、恶意软件检测、对抗样本可解释性等, 进一步明确了对抗攻击的研究对于深度学习的价值;

(3) 按对抗攻击的发展历程对其进行详细梳理, 探讨了对抗攻击面临的主要挑战和未来可能的发展方向.

本文整体架构如下. 第 2 节主要介绍了对抗攻击的基本知识和概念, 以及计算机视觉领域中常用数据集. 第 3 节按照扰动范围、模型知识背景、攻击针对性和攻击频次提出相应的对抗攻击分类方法. 第 4 节重点分析了对抗攻击在计算机视觉领域中的 8 类关键技术. 第 5 节介绍了对抗攻击在诸如自然语言处理、语音识别、恶意软件检测和可解释性对抗样本等领域的应用. 第 6 节探讨了对抗样本攻击面临的主要挑战和未来的发展前景. 第 7 节总结全文.

2 背景知识

2.1 深度学习

深度学习是由大量带有激活函数的神经元组成的深层次的神经网络. 神经元接受上层输入信号后进行加权连接, 通过激活函数处理产生神经元的输出并进行信号的下层传递, 从而构建了深层次的神经网络结构. 深度神经网络能够在不依赖于专家知识的情况下自动学习原始数据的显隐性特征, 其形式化表达如公式(1)所示.

$$f(x,\theta) = f^{(k)}(\cdots f^{(2)}(f^{(1)}(x,\theta_1),\theta_2),\theta_k) \tag{1}$$

这里 $f^{(i)}(x,\theta_i)$ 是第 i 层网络的函数, $i = 1,2,\cdots,k$, 其中 k 是深度神经网络的层数. 在计算机视觉领域, 卷积神经网络 (CNN) 是最常用的神经网络结构之一. CNN 由输入层、卷积层、池化层和全连接层组成, 其中卷积层通过权重共享进行卷积运算, 池化层通过对主要特征采样调整信号规模. 手写字体识别模型 LeNet-5 [4] 诞生于 1998 年, 是最早的 CNN 之一. 近年来, 随着 ImageNet 大规模视觉识别挑战赛 (ILSVRC) 的兴起, 涌现了大量优秀的 CNN 模型, 代表性的研究成果包括:

(1) AlexNet [5]: 由 2012 年 ILSVRC 冠军获得者 Krizhevsky 等人提出, 总共有 8 个带权重的网络层, 其中前 5 层为卷积层, 后 3 层为全连接层;

(2) VGG [6]: 由 2014 年 ILSVRC 第二名获得者 Simonyan 等人提出, 以 VGG-16 和 VGG-19 为代表, 具有网络层次深、泛化能力强等特点;

(3) GoogLeNet [7]: 由 2014 年 ILSVRC 冠军获得者 Szegedy 等人提出, 通过引入 Inception 模块来提高网络内部计算资源的利用率;

(4) ResNet [8]: 由 2015 年 ILSVRC 冠军获得者 He 等人提出, 通过改变深度神经网络的连接方式简化网络训练, 单个模型在 ImageNet 数据集的准确率高达 95.51%;

(5) SeNet [9]: 由 2017 年 (最后一届)ILSVRC 冠军获得者 Hu 等人提出, 该模型关注通道之间的关系, 并提出 SE 模块对学习到的特征进行自适应重构.

2.2 对抗攻击

对抗样本是指通过对原始样本人为加入人眼视觉难以感知的细微扰动所形成的输入样本, 该样本能使深度学习模型以高置信度的方式给出错误的输出. 通过生成对抗样本以达成逃避基于深度学习的检测服务的攻击方式被称为对抗攻击. 对抗攻击的流程如图 2 所示.

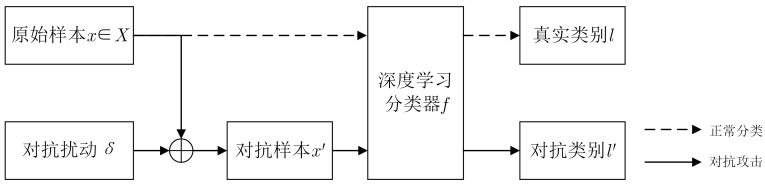


图 2　对抗攻击流程图

Figure 2　Flow chart of adversarial attacks

在图像分类任务中, 用户输入一张图像至已训练好的深度学习分类器中, 分类器会给出相应类别的预测结果. 当遭受对抗攻击后, 原始样本被加入人眼难以察觉的扰动, 形成的对抗样本能够误导分类器给出其他类别的预测结果. 对抗攻击的形式化表达如公式(2)所示.

$$\begin{aligned} \min \quad & \|\delta\| \\ \text{s.t.} \quad & f(x)=l \\ & f(x')=l' \\ & l\neq l' \\ & x'=x+\delta\in D \end{aligned} \tag{2}$$

这里, 深度学习分类器 f 分类原始样本 x 至类别 l, 分类加入扰动 δ 所形成的对抗样本 x' 至类别 l', D 是样本的取值范围. 对抗攻击是指在成功误导深度学习分类器的前提下求解扰动量的最小值, 本质是一个约束最优化问题.

2.3　样本距离度量

对抗攻击是指通过加入人眼难以察觉的扰动, 生成能够成功欺骗深度学习分类器以达到逃逸攻击的目的. 为了使对抗样本更具欺骗性, 对抗攻击中目标函数的定义显得尤为重要. 目标函数的定义涉及到对原始样本与对抗样本的距离度量, 以量化样本之间的相似性. 在许多经典的对抗攻击算法中, L_p 被广泛用于度量样本间的 p 范式距离, 其定义如公式(3)所示.

$$\|\delta\|_p=\left(\sum_{i=1}^{n}|\delta_i|^p\right)^{\frac{1}{p}} \tag{3}$$

这里 L_0, L_2 和 L_∞ 是 3 个最常用的样本距离度量指标. L_0 是指对抗扰动的数量, L_2 是指原始样本与对抗样本的欧几里得距离, L_∞ 表示对抗扰动的最大改变强度.

2.4　数据集

对抗攻击通常需要对相同数据集进行仿真实验, 以评估和对比攻击方法的性能. 在图像分类领域, ImageNet、MNIST 和 CIFAR-10 是 3 个应用非常广泛的开源数据集. ImageNet 是根据 WordNet 层次结构组织的图像数据集. 该数据集数量庞大、类别丰富, 是迄今为止最优秀的图像数据集, 著名的 ILSVRC 挑战赛基于此数据集展开对抗攻击和防御. MNIST 数据集来自美国国家标准与技术研究所, 是一个手写体数字 (0–9) 数据库. 该数据集包含 60 000 个训练样本和 10 000 个测试样本, 数字已标准化于大小为 28×28 的图像中. CIFAR-10 是由 Geoffrey Hinton 的学生 Alex Krizhevsky 和 Vinod Nair 等人整理搭建的小型数据集, 用于普通物体识别. 该数据集包含 60 000 张大小为 $32\times 32\times 3$ 的图像, 其中训练样本 50 000 张、测试样本 10 000 张. CIFAR-10 数据集总共分为 10 个类别, 分别是飞机、汽车、鸟、猫、鹿、狗、蛙、马、船和卡车. MNIST 和 CIFAR-10 内容简洁、尺寸小, 易于对抗攻击的实施, 因此经常被作为评价对抗攻击性能的图像数据集.

3 对抗攻击分类

攻击者通常会根据不同的场景设计相应的对抗攻击方案. 归纳对抗攻击方法的特性, 我们从扰动范围、攻击者掌握目标模型知识的情况、攻击目标的针对性、攻击实施的频次等 4 个方面对对抗攻击进行分类.

(1) 扰动范围

a. 全局像素扰动攻击

全局像素扰动攻击是指攻击者生成对抗样本过程中通过对图像所有像素增加合适的扰动, 以达到欺骗深度学习模型的目的. 利用梯度下降生成对抗样本的方法是一种典型的全局像素扰动攻击, 具有更强的可迁移性.

b. 部分像素扰动攻击

部分像素扰动攻击是指攻击者通过权衡各个像素的扰动优先级、并选择最佳的扰动组合生成对抗样本, 以达到欺骗深度学习模型的目的. 该方法有时仅须改变一个像素值, 但生成的对抗样本通常不具备可迁移性.

(2) 模型知识

a. 白盒攻击

白盒攻击是指攻击者在已知目标模型所有知识的情况下生成对抗样本的一种攻击手段, 这些知识包括网络结构、权重和超参、激活函数类型、训练数据等. 这种攻击方案实施起来较为容易, 但多数场景下攻击者难以获得深度学习模型的内部知识, 因此应用场景非常有限.

b. 黑盒攻击

黑盒攻击是指攻击者在不知道目标模型任何内部信息的情况下实施的攻击方案. 这类攻击者通常扮演普通用户获得基于深度学习的应用服务的分类结果, 通过应用服务提供的输出对该模型实施对抗攻击. 由于不需要掌握目标模型, 黑盒攻击更容易在低控制权场景下部署和实施.

(3) 针对性

a. 定向攻击

定向攻击旨在将深度学习分类器误导至攻击者指定的类别. 例如在人脸识别系统中, 攻击者需要将未授权的人脸伪装成已授权的人脸, 以实现非法授权. 定向攻击一方面需要降低深度学习对输入样本真实类别的置信度, 同时应尽可能提升攻击者指定类别的置信度, 因此攻击难度较大.

b. 非定向攻击

非定向攻击旨在将深度学习分类器误导至错误的类别即可, 而不指定具体的类别. 例如在监控系统中, 攻击者希望通过生成对抗样本实现逃逸攻击, 达到规避检测的目的. 非定向攻击仅需要尽可能降低深度学习对输入样本真实类别的置信度, 因此攻击难度相对较小.

(4) 攻击频次

a. 单次攻击

单次攻击是指攻击者只需要一次计算就能够生成成功欺骗深度学习模型的对抗样本, 即通过一次计算找到约束条件下的最优解. 一般情况下, 单次攻击的效率高, 但生成的对抗样本鲁棒性较差.

b. 迭代攻击

迭代攻击通常需要多次计算逼近约束条件下的最优解. 该攻击方案较单次攻击需要更长的运行时间, 但通常能得到误分类率更高、鲁棒性更强的对抗样本.

4 关键技术研究进展

对抗样本的发现源自于对深度学习可解释性的探索. Szegedy 等人 [1] 发现加入特定扰动的图像样本能够轻易欺骗深度神经网络, 并提出 “对抗样本” 这一概念. 这个有趣的发现促进了研究人员对对抗样本引发的安全问题的思考. 攻击者通过对输入样本加入少量扰动便能有效实施逃逸攻击, 轻易规避基于深度学习服务的安全检测. 这一节我们按各个技术提出的时间顺序介绍了近年来计算机视觉领域中对抗攻击研究的代表性成果, 并按照第 3 节所介绍的对抗攻击分类方法对这些攻击算法进行分类, 分类结果见表 1.

表 1　对抗攻击分类
Table 1　Taxonomy of adversarial attacks

对抗攻击	扰动范围	模型知识	针对性	攻击频次
L-BFGS	全局	白盒	定向	迭代
FGSM	全局	白盒	非定向	单次
JSMA	部分	白盒	定向	迭代
DeepFool	全局	白盒	非定向	迭代
Universal Perturbation	全局	白盒	非定向	迭代
One-Pixel	部分	黑盒	定向	迭代
C&W 攻击	全局	白盒	定向	迭代
Luo&Liu 攻击	部分	黑盒	定向	迭代

(1) L-BFGS 攻击

在探索深度学习可解释性的研究中, Szegedy [1] 等人证明了深度学习对加入特定扰动的输入样本表现出极强的脆弱性, 并由此发现了对抗样本的存在, 提出了第一个针对深度学习的对抗攻击方案 L-BFGS. L-BFGS 攻击的定义如公式(4)所示.

$$\begin{aligned} \min \quad & c\|\delta\| + J_\theta\left(x', l'\right) \\ \text{s.t.} \quad & x' \in [0,1] \end{aligned} \tag{4}$$

这里 c 是大于 0 的常量, x' 是对输入样本 x 增加扰动 δ 所形成的对抗样本, J_θ 为损失函数. L-BFGS 攻击所生成的对抗样本的质量严重依赖于参数 c 的选取, 因此该方法通常需要花费大量的时间寻找合适的参数 c 以求解约束最优化问题. 利用 L-BFGS 攻击生成的对抗样本具有良好的迁移性, 大多数情况下在不同类型的深度神经网络结构、不同数据集训练的模型中同样适用. 该方法的提出引起了学术界和工业界对深度学习可信性的思考, 开启了针对深度学习的对抗攻击和防御等安全问题的研究.

(2) FGSM 攻击

Goodfellow 等人 [10] 认为对抗样本的存在源自于深度神经网络的高维度线性特性, 高维度的线性模型必然存在对抗样本. 基于这一观点, Goodfellow 等人设计了 FGSM (fast gradient sign method) 对抗攻击, 该方法的形式化表达如公式(5)所示.

$$\delta = \varepsilon \operatorname{sign}\left(\nabla_x J_\theta(x, l)\right) \tag{5}$$

这里 ε 是常量, sign 表示符号函数, 对抗扰动为 δ. FGSM 攻击采用后向传播求解神经网络损失函数的梯度. 该方法仅需一次梯度更新得到对抗扰动, 属于单次攻击的类别, 因此对抗攻击实施的效率非常高, 但对抗样本的不可见性难以保证. 在此基础上, 许多改进的方案相继提出. 考虑到一次梯度更新生成的对抗样本扰动强度较大, Kurakin 等人 [11] 提出了基础迭代法 I-FGSM, 通过多个小步梯度更新优化扰动强度. Rozsa 等人 [12] 提出 FGVM(fast gradient value method) 攻击, 直接利用损失函数的梯度值 $\delta = \nabla_x J_\theta(x, l)$ 作为扰动强度生成对抗样本, 为每张图像提供多个可能的对抗性扰动.

(3) JSMA 攻击

JSMA (Jacobian-based saliency map attack) 攻击 [13] 由 Papernot 等人于 2016 年提出. 不同于 FGSM, JSMA 攻击是一种利用前向传播计算输入扰动对神经网络输出结果的影响. JSMA 攻击包括计算深度神经网络的雅可比矩阵、构建对抗显著映射和选择扰动像素三个步骤. 深度神经网络 f 对输入样本

x 的雅可比矩阵计算方法如公式(6)所示.

$$\nabla f(x)=\frac{\partial f(x)}{\partial x}=\left[\frac{\partial f_j(x)}{\partial x_i}\right]_{i\times j} \tag{6}$$

为量化像素值的改变对目标分类器的影响, JSMA 攻击提出了利用雅可比矩阵构建对抗显著映射, 其表达式如公式(7)所示.

$$S(x,t)[i]=\begin{cases} 0, \frac{\partial f_t(x)}{\partial x_i}<0 \text{ or } \sum_{j\neq t}\frac{\partial f_j(x)}{\partial x_i}>0 \\ \left(\frac{\partial f_t(x)}{\partial x_i}\right)\left|\sum_{j\neq t}\frac{\partial f_j(x)}{\partial x_i}\right|, \text{ otherwise} \end{cases} \tag{7}$$

这里 i 是输入特征. 输入特征值越大, 说明基于深度学习的目标分类器对该特征的扰动越敏感. 因此, 在扰动像素的选择阶段, JSMA 攻击选择对抗显著值最大的像素加入扰动, 以此生成对抗样本欺骗深度学习分类器. 实验证明, 当改变原始样本中 4.02% 的特征时, JSMA 攻击可以获得 97% 的攻击成功率. 由于雅可比矩阵计算依赖于神经网络的输入样本, 不同的输入样本计算得到的雅可比矩阵差异较大, 因此 JSMA 攻击通常不具备可迁移性.

(4) DeepFool 攻击

考虑到深度神经网络对对抗样本表现出极强的不稳定性, Moosavi-Dezfooli 等人 [14] 提出了一种计算对抗扰动的方法 DeepFool, 通过计算原始样本与对抗样本的决策边界的最小距离来量化深度学习分类器面向对抗攻击的鲁棒性. 给定一个反射分类器 $f(x)=\omega^{\mathrm{T}}x+b$, 其对应的仿射平面 $\Gamma=\left\{x:\omega^{\mathrm{T}}x+b=0\right\}$, 那么, 改变分类器对原始样本 x_0 分类结果的最小扰动 δ 等于 x_0 到仿射平面 Γ 的正交投影, 最小扰动 δ 的计算方式如公式所示.

$$\begin{aligned} &\delta_*\left(x_0\right):=\arg\min_{\delta}\|\delta\|_2 \\ &\text{s.t. } \operatorname{sign}\left(f\left(x_0+\delta\right)\right)\neq\operatorname{sign}\left(f\left(x_0\right)\right)=-\frac{f(x_0)}{\|\omega\|_2^2} \end{aligned} \tag{8}$$

通过整体迭代, DeepFool 攻击能够得到对抗扰动的近似最小值 δ, 其表达如公式(9)所示.

$$\begin{aligned} &\arg\min_{\delta_i}\left\|\delta_i\right\|_2 \\ &\text{s.t. } f\left(x_i\right)+\nabla f\left(x_i\right)^{\mathrm{T}}\delta_i=0 \end{aligned} \tag{9}$$

这里 δ_i 是指第 i 轮迭代中加入的对抗扰动, 可由公式(8)计算得到. DeepFool 攻击通过每一轮的像素修改将原始样本推向决策边界, 直至跨越决策边界形成对抗样本. 相比于 FGSM 攻击、JSMA 攻击, DeepFool 攻击生成的对抗样本平均扰动最小. 然而, DeepFool 攻击是以最小距离使原始样本跨越决策边界形成对抗样本, 因此无法将深度学习分类器误导至指定的类别, 即不具备定向攻击的能力.

(5) Universal Perturbation 攻击

不同于针对单个图像的对抗攻击, Universal Perturbation 攻击 [15] 提出了一种通用的对抗扰动计算方法. 该方法生成的扰动具有很强的泛化能力, 能够跨数据集、跨模型实施对抗攻击. Universal Perturbation 攻击借鉴了 DeepFool 攻击的思想, 利用对抗扰动将大多数原始样本推出决策边界, 其定义如公式(10)所示.

$$f(x+\delta)\neq f(x) \text{ for “most” } x\in X \tag{10}$$

这里通用的对抗扰动 δ 必须满以下约束:

$$\begin{aligned} &\|\delta\|_p\leq\varepsilon \\ &\underset{x\in X}{P}(f(x+\delta)\neq f(x))\geq 1-\alpha \end{aligned} \tag{11}$$

这里参数 ε 控制对抗扰动 δ 的扰动强度, α 控制对图像库 X 实施对抗攻击的失败率. 在计算 Universal Perturbation 的整体迭代过程中, Moosavi-Dezfooli 等人采用 DeepFool 算法计算每一个输入样本的最小扰动并更新对抗扰动 δ, 直至大多数 $(P \geq 1-\alpha)$ 的对抗样本能够成功欺骗深度学习分类器. Universal Perturbation 攻击的存在揭示了深度神经网络决策边界之间的几何关联. 攻击者不需要直接攻击目标模型, 而是在本地生成泛化能力强的对抗扰动, 以此迁移至目标模型实施对抗攻击, 实现在低控制权场景下对抗攻击的部署和开展.

(6) One-Pixel 攻击

One-Pixel 攻击 [16] 通过仅改变原始图像中一个像素点实现针对深度神经网络的对抗攻击, 是一种基于前向传播的攻击方案. 该方法对扰动像素的位置信息和扰动强度进行编码, 基于差分进化算法利用深度神经网络的反馈结果引导对抗扰动的进化方向, 直至对抗扰动收敛至稳定的状态. One-Pixel 攻击的定义如公式(12)所示.

$$\begin{aligned} &\text{maximize } f_{l'}(x+\delta) \\ &\text{s.t. } \|\delta\|_0 \leq d \end{aligned} \tag{12}$$

这里 $f_{l'}(x')$ 表示深度学习分类器 f 识别对抗样本 x' 为类别 l' 的概率, $d=1$ 表示仅改变一个像素值. 该问题本质上是一个单约束的优化问题. 考虑到暴力求解该优化问题的时间代价高, Su 等人 [16] 引入差分进化算法求解最优的对抗扰动. 这里采用的差分进化算法不包括交叉算子, 其变异算子如公式(13)所示.

$$x_i(g+1) = x_{r1}(g) + F\left(x_{r2}(g) + x_{r3}(g)\right), \quad r1 \neq r2 \neq r3 \tag{13}$$

这里 $x_i(g+1)$ 表示第 $g+1$ 代的候选解, 该候选解是由对抗扰动的坐标 $x-y$ 和改变强度 RGB 组成. $F=0.5$ 表示缩放因子, $r1$, $r2$, $r3$ 是随机数. 由于差分进化算法不需要知道深度神经网路的梯度信息, 仅依赖输出类别的概率分布引导进化方向, 因此 One-Pixel 攻击属于黑盒攻击. 此外, 该攻击方案不要求深度神经网络可微分, 适用于多种深度学习分类器. 基于进化算法的优化问题计算依赖于种群规模和迭代次数, 为尽可能获得全局最优解, 种群规模和迭代次数的设定相对较大. One-Pixel 攻击通常需要在较大的种群规模中通过多轮迭代寻求对抗扰动的最优解, 因此攻击效率较低.

(7) C&W 攻击

防御性蒸馏 [17] 是利用知识蒸馏将复杂模型所学的 “知识” 迁移到结构简单的神经网络中, 通过避免攻击者直接接触原始神经网络达到防御对抗攻击的目的. 针对防御性蒸馏, Carlini 和 Wagner 联合提出了在 L_0, L_2 和 L_∞ 范式下的一组有效的攻击方法 C&W 攻击 [18]. 通过对比这 3 种范式下的实验结果, 他们认为, L_2 范式下的 C&W 攻击具备最强的攻击能力. L_2 范式的 C&W 攻击如公式(14)所示.

$$\begin{aligned} &\min \quad \|\delta\|_2 + c \cdot f(x') \\ &s.t. \quad x' = x + \delta \in D \end{aligned} \tag{14}$$

这里, f 的定义如下:

$$f(x') = \max(\max\{Z(x')_i : i \neq t\} - Z(x')_t, -k) \tag{15}$$

这里 $Z(x)$ 是神经网络 softmax 层的输出, 超参 k 约束该攻击找到具有高置信度且错误类别为 t 的对抗样本 x'. C&W 攻击表现出了较 L-BFGS、FGSM、JSMA 和 DeepFool 更好的攻击效果, 同时具备破坏防御性蒸馏的能力. 然而, C&W 攻击需要花费大量的时间寻找合适的参数以约束扰动的可见性, 属于迭代攻击的类别, 因此攻击效率相对较低.

(8) Luo&Liu 攻击

对抗攻击的加入应尽可能不被人眼视觉察觉, 因此样本距离度量需要充分考虑人眼视觉系统. Luo 等人 [19] 通过调研人眼视觉系统发现, 人眼对平坦区域的扰动较纹理区域更敏感, 在加入对抗扰动时应考虑

扰动像素周围的纹理特征. 而图像的纹理特征可以通过方差量化, 量化方法如公式(16)所示.

$$\mathrm{SD}\left(p_i\right)=\sqrt{\frac{\sum_{p_k\in S_i}\left(p_k-\mu\right)^2}{n^2}} \tag{16}$$

这里 $\mathrm{SD}\left(p_i\right)$ 表示以扰动像素 p_i 为中心的 $n\times n$ 窗口区域内的方差. 考虑到扰动强度对人眼视觉的影响, Luo 等人 [19] 建立了与扰动强度和纹理特征有关的样本距离度量方法, 如公式(17)所示.

$$D\left(x,x'\right)=\sum_{i=1}^{m}\frac{\chi_i}{\mathrm{SD}\left(p_i\right)} \tag{17}$$

这里 m 表示加入扰动的数量, χ 是扰动强度, $D\left(x,x'\right)$ 表示原始样本与对抗样本的视觉距离. 为了增强对抗样本的鲁棒性, Luo&Liu 攻击构建了可微分的目标函数如公式(18)所示.

$$\mathrm{Gap}\left(x'\right)\approx P_{l'}-\log\left(\sum e^{kP_i}\right)/k \tag{18}$$

这里 i 表示除目标类别之外的其他类别, P_i 是指深度学习分类器将样本识别为类别 i 的概率. 通过重构约束函数和目标函数, Luo 等人提出了一种新的对抗攻击方案, 如公式(19)所示.

$$\begin{aligned}&\underset{x'}{\mathrm{argmax}}\quad \mathrm{Gap}\left(x'\right)\\&\mathrm{s.t.}\quad\ D\left(x,x'\right)\le D_{\max}\end{aligned} \tag{19}$$

这里 $D_{\max}$ 表示样本间的最大距离, 由攻击者手动输入. Luo&Liu 攻击通过评估每个像素值的扰动优先级, 利用贪婪算法得到近似最优的像素扰动组合, 形成对抗扰动. 相比于 L_p 范式距离度量, 该方案的约束函数考虑了扰动强度、数量和纹理特征, 生成的对抗样本具有更强的隐蔽性. 然而, 由于像素优先级的评估需要遍历原始样本的所有像素及其对应的扰动强度, 大量的时间开销使得该攻击方案效率低下.

5 对抗攻击应用

对抗攻击的提出源自于深度学习在计算机视觉领域的应用和探索. 事实上, 对抗攻击充分利用了深度神经网络与人类在理解输入样本时先天存在的差异, 通过加入扰动扩大这种差异, 从而导致深度学习模型给出不同于人类感知的判断. 因此, 对抗攻击不仅仅局限于计算机视觉领域, 文本、音频、代码等其他数据类型也同样面临对抗攻击的安全威胁. 第 5 节介绍了对抗攻击在诸如自然语言处理 [20]、语音识别 [21,22]、恶意软件检测 [23,24] 等领域的应用, 特别介绍了对抗攻击技术对深度神经网络工作原理的探索, 强调该技术在揭示深度学习可解释性的应用价值 [1,25].

5.1 自然语言处理

对抗攻击对深度学习的威胁在自然语言处理领域广泛存在. 不同于图像相邻像素相关性高, 文本的离散性使其难以优化, 而且文本对某些单词非常敏感, 一个单词的简单替换也可能导致整体语义发生变化. Li 等人 [20] 充分利用文本的离散性研制了一套高效生成对抗文本的工具 TextBugger, 其攻击流程主要包括选择单词扰动位置、添加扰动两个阶段.

在选择单词扰动位置时, TextBugger 针对白盒攻击和黑盒攻击的场景分别提出了单词扰动优先级的评估方法. 在白盒攻击的场景下, TextBugger 利用雅克比矩阵评估文本扰动对目标分类器的影响. 在黑盒攻击的场景下, TextBugger 通过计算移除文本种某一该单词后检测器对文本预测结果置信度的变化, 以此寻找对分类结果影响最大的单词.

在添加扰动过程中, TextBugger 定义了 5 种常用的文本扰动方法: (1) 插入: 在单词中随机插入一个空格; (2) 删除: 随机删除单词中的一个字母; (3) 交换: 随机交换邻近的两个字母; (4) Sub-C: 将单词中的一个字母替换为外观相似的字符 (比如 o 替换成 0, l 替换成数字 1); (5) Sub-W: 在词空间中寻找邻近的

单词替代 (比如 foolish 替换成 silly). TextBugger 尝试对扰动优先级高的单词选择不同的扰动策略, 直到成功欺骗基于深度学习的文本分类器.

TextBugger 攻击针对特定白盒模型, 如卷积神经网络 (CNN)、长短期记忆网络 (LSTM) 等表现出良好效果. 在针对如 Google Perspective、IBM Classifier、Facebook fastText 等 5 类基于深度学习的文本分类的应用中, TextBugger 同样具有非常高的误分类率, 进而证明了对抗攻击在自然语言处理领域具有显著的现实威胁.

5.2 语音识别

相比于视觉, 人类对语音领域的扰动更为敏感, 而且这些扰动在无线传输信道中容易受到干扰 [21, 22]. 因此, 对抗攻击在语音识别领域的不可察觉性和鲁棒性面临严峻挑战. Yao 等人 [21] 利用听觉掩码的心理声学模型生成音频对抗样本, 从鲁棒性和不可察觉性两方面定义的对抗攻击的目标函数如公式(20)所示.

$$J\left(x', l'\right) = \underset{t\sim \mathrm{T}}{E}\left[J_{\mathrm{net}}\left(t(x+\delta), l'\right) + \alpha J_{\theta}(x, \delta)\right] \tag{20}$$

这里声学模拟器的变换函数 t 能够将输入音频 $x+\delta$ 转化为混响音频 $t(x+\delta)$. $\underset{t\sim \mathrm{T}}{E}[\cdot]$ 表示 t 满足 T 分布的期望值. α 为常数. $J_{\mathrm{net}}(\cdot)$ 是深度神经网络的损失函数. $J_{\theta}(\cdot)$ 表示不可察觉性损失函数, 其定义如公式(21)所示.

$$J_{\theta}(x, \delta) = \frac{1}{\left\lfloor \frac{N}{2} \right\rfloor + 1} \sum_{k=0}^{\left\lfloor \frac{N}{2} \right\rfloor} \max\left\{\bar{p}_{\delta}(k) - \theta_x(k), 0\right\} \tag{21}$$

这里 N 是预定义窗口大小, $\left\lfloor \frac{N}{2} \right\rfloor$ 输出不大于 $\frac{N}{2}$ 的最大整数. $\bar{p}_{\delta}(k)$ 表示扰动的功率谱密度, $\theta_x(k)$ 表示听觉掩码阈值. 当 $\bar{p}_{\delta}(k)$ 小于 $\theta_x(k)$ 时, 人类听觉无法感知扰动 δ. 该方法充分考虑人类听觉系统, 成功攻击了谷歌推出的 Lingvo 语音识别系统, 保证对任意完整语音定向攻击的成功率高达 100%.

5.3 恶意软件

恶意软件检测系统本质上执行二分类任务, 即区别恶意软件和非恶意软件. 基于深度学习的恶意软件检测系统 [23, 24] 通常需要完成恶意代码特征提取和特征学习两个过程. 对抗攻击通过对代码特征人为加入少量扰动以误导恶意软件检测系统, 使其按攻击者的意图给出分类结果. 因此, 生成恶意软件的对抗样本主要有以下两步: 评估扰动对分类器的影响和选择扰动生成对抗样本. 基于此, Grosse 等人 [23] 借鉴了 JSMA 攻击的算法, 利用雅可比矩阵计算代码特征的扰动优先级, 然后加入当前扰动优先级最大的特征直至成功规避检测系统. 为保证对抗性样本的可用性, 该方案对扰动做了如下限制: (1) 对抗扰动仅影响应用程序中单行代码的特性; (2) 仅扰动与 AndroidManifest.xml 清单文件相关的特性. 这种攻击方法通过两个限制条件确保了恶意软件的对抗样本保留其原始功能, 同时能有效规避检测系统. 由于雅可比矩阵的计算需要知道目标系统的神经网络结构, 因此这种攻击方法属于白盒攻击.

5.4 可解释性对抗样本

对抗样本的提出源自于对深度学习可解释性的探索, 而后发展成针对基于深度学习检测服务的逃逸攻击及其防御. 而在文献 [25] 中, Ilyas 等人将对输入样本的特征划分为鲁棒性特征和非鲁棒性特征. 鲁棒性特征是指人类视觉可以理解的特征 (例如鼻子、眼睛、嘴等), 其他特征为非鲁棒性特征. 基于这一划分, 他们利用对抗攻击解释了对抗样本存在的原因: 对抗脆弱性由非鲁棒性特性引起. 为证明这一结论, Ilyas 等人 [25] 构建了以下两个实验.

实验 1: 构建由鲁棒性特征组成的训练样本并训练深度学习模型. 如果该模型的鲁棒性更强, 说明通过删除非鲁棒性特征能够提升模型鲁棒性.

实验 2: 构建由非鲁棒性特征组成的训练样本并训练深度学习模型. 如果该模型有效, 说明深度学习模型使用到了非鲁棒性特征, 对抗样本是有价值的特征.

为构建由鲁棒性特征组成的训练样本, 实验 1 对输入样本集 (X, L) 展开对抗攻击, 得到对抗样本集 (X', L'). 由于对抗攻击破坏了输入样本 X 的非鲁棒性特征、保留其鲁棒性特征, 因此可以构建鲁棒性特征样本集 (X', L). 实验分别利用鲁棒性特征样本集 (X', L) 和原始样本集 (X, L) 训练同一深度学习模

型. 结果表明, 利用鲁棒性特征样本集 (X', L) 训练的模型鲁棒性更强, 从而证明删除样本中的非鲁棒性特征能够有效提升深度学习模型的鲁棒性.

为构建由非鲁棒性特征组成的训练样本, 实验 2 利用对抗攻击方法得到对抗样本集 (X', L'), 并以此作为非鲁棒性特征样本集, 此时可以认为对抗样本 X' 保留了类别 L 的鲁棒性特征和类别 L' 的非鲁棒性特征. 实验利用非鲁棒性特征样本集 (X', L') 训练深度学习模型, 并在原始测试样本集进行精确度测试. 结果发现在 CIFAR 数据集的精确度高达 43%, 说明深度神经网络学习到了输入样本的非鲁棒性特征.

该方法巧妙地利用对抗攻击从正反两方面证明了输入样本的非鲁棒性特征是有价值的特征, 而不是有限样本过拟合的产物, 进而解释了对抗样本现象是由输入样本的非鲁棒性特性引起. 如果不明确地阻止深度神经网络学习输入样本的非鲁棒性特征, 深度学习将始终面临对抗攻击的安全威胁.

6 挑战与展望

对抗样本现象吸引了学术界和产业界的共同关注. 随着对抗攻击技术的发展, 对抗样本的防御手段得到重视. 主流的对抗样本防御技术包括蒸馏神经网络 [17]、对抗训练 [26,27]、梯度掩盖 [28,29] 等, 但这些防御手段大多最终都被成功破解或被证明无效 [30,31]. 对抗样本的检测技术作为一种防御手段的补充应运而生, 例如, 利用深度神经网络对合法样本与对抗样本进行二分类 [32], 在深度神经网络中加入检测机制 [33], 打造平台以评估样本面向对抗攻击的鲁棒性等 [34]. 但这些检测技术无法完全区分合法样本和对抗样本, 对抗攻击的安全隐患仍然存在. 因此, 在深度学习的可解释性没有完全揭晓之前, 针对深度学习的对抗攻击仍然值得重视.

对抗攻击技术在计算机视觉领域取得了显著的成绩, 其未来可能的发展方向至少包括以下 3 个方面. 一是生成具备高鲁棒性 [19,35,36]、隐蔽性 [19,37] 和可迁移性 [15,38] 的对抗样本, 保障对抗攻击在复杂环境和低控制权场景下的部署和实施; 二是注重对抗攻击的应用价值 [20–24,39], 例如, 生成病毒的对抗样本规避恶意软件检测系统 [23,24], 生成恶意文件的对抗样本攻击文件分类器 [39] 等; 三是利用对抗攻击技术探索深度神经网络的工作原理, 从本质上揭示深度学习的可解释性 [1,25].

7 结语

本文介绍了针对深度学习的对抗攻击基本原理和分类方法, 重点选取了计算机视觉领域中 8 类主流的对抗攻击算法进行详细解读和分析, 并着重介绍了对抗攻击在自然语言处理、语音识别、恶意软件检测和解释对抗样本现象等 4 个方面的应用. 回顾其发展历程, 对抗样本现象的发现源于对深度学习可解释性的探索, 初期的对抗攻击方法重点关注攻击行为的可用性, 而后向可迁移性、隐蔽性、鲁棒性等多方面发展, 同时注重对抗攻击在各种复杂环境和低控制权场景下的实际应用, 并逐步延申至针对深度学习检测服务的逃逸攻击. 在总结其发展历程的基础上, 本文分析了对抗攻击面临的主要挑战, 并指出了该技术未来 3 个重要的研究方向.

参考文献

[1] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[C]. In: Proceedings of 2014 International Conference on Learning Representations (ICLR Poster). Banff, AB, Canada. April 14–16, 2014.

[2] QUIRING E, ARP D, RIECK K, et al. Unifying attacks on machine learning and digital watermarking[C]. In: Proceedings of 2018 IEEE European Symposium on Security and Privacy (EuroS&P). IEEE, 2018: 488–502. [DOI: 10.1109/EuroSP.2018.00041]

[3] SHARIF M, BHAGAVATULA S, BAUER L, et al. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition[C]. In: Proceedings of 2016 ACM Conference on Computer and Communications Security (CCS). ACM, 2016: 1528–1540. [DOI: 10.1145/2976749.2978392]

[4] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278–2324. [DOI: 10.1109/5.726791]

[5] KRIZHEVSKY A, SUTSKEVER I, HINTON G. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84–90. [DOI: 10.1145/3065386]

[6] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]. In: Proceedings of 2015 International Conference on Learning Representations (ICLR). San Diego, CA, USA. May 7–9, 2015: 1–14.

[7] SZEGEDY C, LIU W, YANG Q J, et al. Going deeper with convolutions[C]. In: Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2015: 1–9. [DOI: 10.1109/CVPR.2015.7298594]

[8] HE K M, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]. In: Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016: 770–778. [DOI: 10.1109/CVPR.2016.90]

[9] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]. In: Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2018: 7132–7141. [DOI: 10.1109/CVPR.2018.00745]

[10] GOODFELLOW I, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[C]. In: Proceedings of 2015 International Conference on Learning Representations (ICLR Poster). San Diego, CA, USA. May 7–9, 2015.

[11] KURAKIN A, GOODFELLOW I, BENGIO S, et al. Adversarial examples in the physical world[C]. In: Proceedings of 2017 International Conference on Learning Representations (ICLR). Toulon, France. April 24–26, 2017: 1–14.

[12] ROZSA A, RUDD E M, BOULT T E, et al. Adversarial diversity and hard positive generation[C]. In: Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016: 410–417. [DOI: 10.1109/CVPRW.2016.58]

[13] PAPERNOT N, MCDANIEL P, JHA S, et al. The limitations of deep learning in adversarial settings[C]. In: Proceedings of 2016 IEEE European Symposium on Security and Privacy (EuroS&P). IEEE, 2016: 372–387. [DOI: 10.1109/EuroSP.2016.36]

[14] MOOSAVI-DEZFOOLI S, FAWZI A, FROSSARD P, et al. DeepFool: A simple and accurate method to fool deep neural network[C]. In: Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016: 2574–2582. [DOI: 10.1109/CVPR.2016.282]

[15] MOOSAVI-DEZFOOLI S, FAWZI A, FAWZI O, et al. Universal adversarial perturbations[C]. In: Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017: 86–94. [DOI: 10.1109/CVPR.2017.17]

[16] SU J, VARGAS D, SAKURAI K, et al. One pixel attack for fooling deep neural networks[J]. IEEE Transactions on Evolutionary Computation, 2019, 23(5): 828–841. [DOI: 10.1109/TEVC.2019.2890858]

[17] PAPERNOT N, MCDANIEL P, WU X, et al. Distillation as a defense to adversarial perturbations against deep neural networks[C]. In: Proceedings of 2016 IEEE Symposium on Security and Privacy (S&P). IEEE, 2016: 582–597. [DOI: 10.1109/SP.2016.41]

[18] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]. In: Proceedings of 2017 IEEE Symposium on Security and Privacy (S&P). IEEE, 2017: 39–57. [DOI: 10.1109/SP.2017.49]

[19] LUO B, LIU Y N, WEI L X, et al. Towards imperceptible and robust adversarial example attacks against neural networks[C]. In: Proceedings of 2018 AAAI Conference on Artificial Intelligence (AAAI). New Orleans, LA, USA. 2018: 1652–1659.

[20] LI J F, JI S L, DU T Y, et al. Textbugger: Generating adversarial text against real-world applications[C]. In: Proceedings of 2019 Network and Distributed System Security Symposium (NDSS). San Diego, CA, USA. 2019: 1–15. [DOI: 10.14722/ndss.2019.23138]

[21] QIN Y, CARLINI N, GOODFELLOW I, et al. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition[C]. In: Proceedings of 2019 36th International Conference on Machine Learning (PMLR). Long Beach, CA, USA. 2019, 1–13.

[22] KWON H, KIM Y C, YOON H S, et al. Selective audio adversarial example in evasion attack on speech recognition system[J]. IEEE Transactions on Information Forensics and Security, 2019, 15: 526–538. [DOI: 10.1109/TIFS.2019.2925452]

[23] GROSSE K, PAPERNOT N, MANOHARAN P, et al. Adversarial examples for malware detection[C]. In: Computer Security—ESORICS 2017, Part II. Springer Cham, 2017: 62–79. [DOI: 10.1007/978-3-319-66399-9_4]

[24] LI H, ZHOU S Y, YUAN W, et al. Adversarial-example attacks toward Android malware detection system[J]. IEEE Systems Journal, 2020, 14(1): 653–656. [DOI: 10.1109/JSYST.2019.2906120]

[25] ILYAS A, SANTURKAR S, TSIPRAS D, et al. Adversarial examples are not bugs, they are features[C]. In: Proceedings of 2019 Conference on Neural Information Processing Systems (NeurIPS). Vancouver, BC, Canada. 2019: 1–12. [DOI: 10.23915/distill.00019]

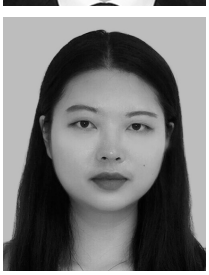
[26] GANI Y, USTINOVA E, AJAKAN H, et al. Domain-adversarial training of neural networks[J]. Journal of Machine

Learning Research, 2016, 17(1): 2096–2030. [DOI: 10.5555/2946645.2946704]
[27] TRAMÈR F, KURAKIN A, PAPERNOT N, et al. Ensemble adversarial training: Attacks and defenses[C]. In: Proceedings of 2018 International Conference on Learning Representations (ICLR). Vancouver, BC, Canada. April 30–May 3, 2018: 1–19.
[28] ATHALYE A, CARLINI N, WAGNER D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples[EB/OL]. arXiv preprint arXiv:1802.00420, 2018.
[29] ZANTEDESCHI V, NICOLAE M, RAWAT A. Efficient defenses against adversarial attacks[C]. In: Proceedings of 2017 ACM Workshop on Artificial Intelligence and Security. ACM, 2017: 39–49. [DOI: 10.1145/3128572.3140449]
[30] CARLINI N, WAGNER D. Defensive distillation is not robust to adversarial examples[EB/OL]. arXiv preprint arXiv:1607.04311, 2016.
[31] CARLINI N, WAGNER D. MagNet and "efficient defenses against adversarial attacks" are not robust to adversarial examples[EB/OL]. arXiv preprint arXiv:1711.08478, 2017.
[32] LU J J, ISSARANON T, FORSYTH D. SafetyNet: Detecting and rejecting adversarial examples robustly[C]. In: Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). IEEE, 2017: 1–9. [DOI: 10.1109/ICCV.2017.56]
[33] FROSST N, SABOUR S, HINTON G. DARCCC: Detecting adversaries by reconstruction from class conditional capsules[EB/OL]. arXiv preprint arXiv:1811.06969, 2018.
[34] LIU H, ZHAO B, HUANG L Q, et al. FoolChecker: A platform to evaluate the robustness of images against adversarial attacks[J]. Neurocomputing, 2020, 412: 216–225. [DOI: 10.1016/j.neucom.2020.05.062]
[35] EYKHOLT K, EVTIMOV I, FERNANDES E, et al. Robust physical-world attacks on deep learning visual classification[C]. In: Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2018: 1625–1634. [DOI: 10.1109/CVPR.2018.00175]
[36] JAN S T K, MESSOU J, LIN Y C, et al. Connecting the digital and physical world: Improving the robustness of adversarial attacks[J]. Proceedings of 2019 AAAI Conference on Artificial Intelligence (AAAI), 2019, 33(1): 962–969. [DOI: 10.1609/aaai.v33i01.3301962]
[37] WANG Z B, SONG M K, ZHENG S Y, et al. Invisible adversarial attack against deep neural networks: An adaptive penalization approach[J]. IEEE Transactions on Dependable and Secure Computing, 2019. [DOI: 10.1109/TDSC.2019.2929047]
[38] WEI X X, LIANG S Y, CHEN N, et al. Transferable adversarial attacks for image and video object detection[C]. In: Proceedings of 2019 International Joint Conference on Artificial Intelligence (IJCAI). Macao, China. 2019: 954–960. [DOI: 10.24963/ijcai.2019/134]
[39] XU W L, QI Y J, EVANS D. Automatically evading classifiers: A case study on PDF malware classifiers[C]. In: Proceedings of 2016 Network and Distributed System Security Symposium (NDSS). 2016: 1–15.

作者信息

刘会 (1992–), 湖北荆州人, 博士生在读. 主要研究领域为可信 AI.
liuhui@whu.edu.cn

赵波 (1972–), 湖北武汉人, 博士, 教授, 博士生导师, 中国密码学会理事, 中国计算机学会高级会员. 主要研究领域为可信计算和人工智能安全.
zhaobo@whu.edu.cn



郭嘉宝 (1993–), 湖北武汉人, 博士生在读. 主要研究领域为人工智能安全.
garbo_guo@whu.edu.cn

彭钺峰 (1998–), 河南信阳人, 硕士生在读. 主要研究领域为人工智能安全.
yuefengpeng@whu.edu.cn