深度学习中对抗样本的构造及防御研究

段广晗[1]，马春光[2]，宋蕾[1]，武朋[2]
（1. 哈尔滨工程大学计算机科学与技术学院，黑龙江 哈尔滨 150001；
2. 山东科技大学计算机科学与工程学院，山东 青岛 266590）

摘　要：随着深度学习技术在计算机视觉、网络安全、自然语言处理等领域的进一步发展，深度学习技术逐渐暴露了一定的安全隐患。现有的深度学习算法无法有效描述数据本质特征，导致算法面对恶意输入时可能无法给出正确结果。以当前深度学习面临的安全威胁为出发点，介绍了深度学习中的对抗样本问题，梳理了现有的对抗样本存在性解释，回顾了经典的对抗样本构造方法并对其进行了分类，简述了近年来部分对抗样本在不同场景中的应用实例，对比了若干对抗样本防御技术，最后归纳对抗样本研究领域存在的问题并对这一领域的发展趋势进行了展望。

关键词：对抗样本；深度学习；安全威胁；防御技术
中图分类号：TP309
文献标识码：A
doi: 10.11959/j.issn.2096−109x.2020016

Research on structure and defense of adversarial example in deep learning

DUAN Guanghan[1], MA Chunguang[2], SONG Lei[1], WU Peng[2]
1. College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China
2. College of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao 266590, China

Abstract: With the further promotion of deep learning technology in the fields of computer vision, network security and natural language processing, which has gradually exposed certain security risks. Existing deep learning algorithms can not effectively describe the essential characteristics of data or its inherent causal relationship. When the algorithm faces malicious input, it often fails to give correct judgment results. Based on the current security threats of deep learning, the adversarial example problem and its characteristics in deep learning applications were introduced, hypotheses on the existence of adversarial examples were summarized, classic adversarial example construction methods were re-

收稿日期：2019−05−16；**修回日期：**2019−08−20
通信作者：马春光，machunguang@sdust.edu.cn
基金项目：国家自然科学基金资助项目（No.61472097，No.61932005，No.U1936112）；黑龙江省自然科学基金资助项目（No.JJ2019LH1770）
Foundation Items: The National Natural Science Foundation of China (No.61472097, No.61932005, No.U1936112), The Natural Science Foundation of Heilongjiang Province (No.JJ2019LH1770)

论文引用格式：段广晗，马春光，宋蕾，等. 深度学习中对抗样本的构造及防御研究[J]. 网络与信息安全学报, 2020, 6(2): 1-11.
DUAN G H, MA C G, SONG L, et al. Research on structure and defense of adversarial example in deep learning[J]. Chinese Journal of Network and Information Security, 2020, 6(2): 1-11.

viewed and recent research status in different scenarios were summarized, several defense techniques in different processes were compared, and finally the development trend of adversarial example research were forecasted.

Key words: adversarial example, deep learning, security threat, defense technology

1　引言

随着算力的提升和数据量的增长，深度学习技术在计算机视觉[1]、网络分析[2]、自然语言处理[3]等领域获得了广泛的应用。自动驾驶[4]、药物分析[5]等深度学习应用的发展，将给人类社会带来巨大变革。新技术在带来新机遇的同时带来了新的挑战，深度学习应用的安全性和可用性逐渐引起了研究者的关注。以往深度学习的研究与实现对应用场景和相应数据集有一定的前提假设，应用场景比较纯粹，相应的数据集也经过一定的预处理，缺乏对恶意场景和恶意数据的考虑。

2013 年，Szegedy 等[6]首次对输入添加刻意构造的扰动，使特定的深度学习模型产生错误的分类，他们将其构造的输入称为对抗样本（adversarial example）。在此基础上，研究者围绕不同的深度学习应用展开了对抗样本的研究。2016 年，Kurakin 等[7]将手机摄像头拍摄到的对抗样本输入 Inception 分类器，展示对抗样本在物理世界场景中的潜在威胁。自然语言处理领域中，替换单个词汇往往可以极大地改变文档的语义，2018 年，Alzantot 等[8]仅通过少量词汇的变动就以 97%和 70%的成功率攻击了情感分析和文本蕴含模型；2019 年，Qin 等[9]构造了人耳无法辨别的音频对抗样本，通过对真实环境的模拟，证明了对抗样本对现有无线音频技术的潜在威胁。深度学习所应用的各个领域，如计算机视觉、自然语言处理、语音识别等容易受到对抗样本的影响。

鉴于对抗样本研究的重要性及其在现实生活中的潜在威胁，本文给出对抗样本这一领域的全景展望，本文的内容包括：① 针对日益增多的对抗样本攻防研究，为对抗样本构造、防御技术给出了相应的分类方法；② 提供全面的对抗样本研究概述，整理了现有对抗样本存在原理的解释，给出了代表性对抗样本构造方法和防御技术的详细描述，同时展示了部分对抗样本典型实例，并对相应研究进行了必要的比较与总结；③ 总结分析了现有研究的局限性，并指出这一热点领域可能的发展方向。

2　对抗样本相关概念及敌手模型

2.1　深度学习

深度学习[10]是机器学习的一个分支，其主要目的是从数据中自动学习有效的特征表示。深度学习模型通过训练学习不同的神经网络，借助神经网络内部层级之间的特征转换，把原始数据抽象转化为更高层次的特征表示。图 1 给出了深度学习的数据处理流程。现有的深度神经网络主要有以下几种：深度神经网络（DNN，deep neural network）、卷积神经网络（CNN，convolutional neural network）、对抗生成网络（GAN，generative adversarial network）、循环神经网络（RNN，recurrent neural network）、自动编码器（AE，auto encoder）等。

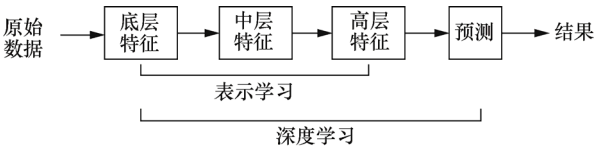

图 1　深度学习的数据处理流程
Figure 1　Data process for deep learning

2.2　对抗样本的正式定义及其性质

参考之前研究人员的工作，对抗样本是敌手设计导致深度学习模型产生错误的输入。经过训练的神经网络模型 f 可以将原始输入样本 x 正确分类为标签 l，敌手对原始输入样本添加扰动 η，使原始输入样本 x 成为对抗样本 x'，其中，$x'=x+\eta$，即

$$\|x'-x\|_P<\varepsilon \tag{1}$$

方程式必须满足的条件为

$$f(x')=l';f(x)=l;l\neq l'$$

由于深度学习模型参数众多，最小扰动 η 的计算过程极为烦琐，因此研究人员提出了很多近似方法构造对抗样本。图 2 给出一个构造对抗样本的过程，在输入中添加扰动后输入数据将移动到其他区域。

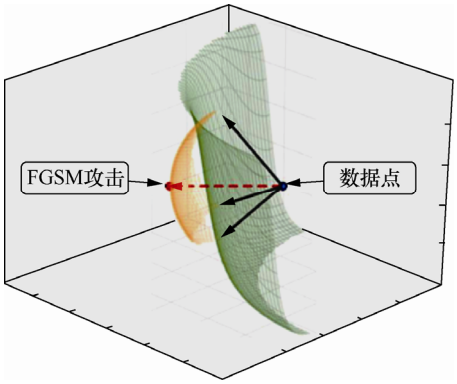

图 2　构造对抗样本的过程
Figure 2　The process of constructs adversarial samples

对抗样本同时展示出了很强的迁移性（transferability）[11]。迁移性主要体现在两点：一部分对抗样本在其他结构的神经网络上会被错误分类；同时会被同一数据集不相交子集训练得到的网络错误分类。这一特性意味着深度学习模型普遍存在被黑盒攻击的风险，Papernot 等[12]在目标神经网络结构、参数未知的情况下，首次对深度神经网络进行了黑盒攻击。他们利用目标深度神经网络的输入以及输出标签训练了替代神经网络，然后针对替代神经网络生成对抗样本，并成功地利用这些样本攻击了目标深度神经网络。

2.3　敌手模型

对抗样本的构造以及防御过程中需要考虑敌手的知识和目标等，本文通过对敌手目标、能力、知识以及策略的描述，建立一般性的对抗样本敌手模型[13]。

敌手目标往往是破坏深度学习模型的完整性与可用性。敌手构造对抗样本，从而导致模型输出错误的结果以达到其目标。在实践中，敌手可能会导致置信度降低（confidence reduction）、无目标攻击（non-targeted attack）或有目标攻击（targeted attack）。其中，置信度降低主要指敌手构造的样本以较低的置信度被正确分类；无目标攻击指敌手构造的样本被预测为与正确结果不同的任何分类；有目标攻击则指敌手构造的样本被预测为敌手指定的特定分类。

在对抗样本研究中，敌手往往具有最基本的数据操纵能力，即对预测数据进行一定修改并输入特定深度学习模型的能力。由于深度学习模型的构建与强大的硬件紧密结合，攻击者的能力可能受到无法训练大型模型的限制。本文假设排除了此限制，即敌手不受硬件限制约束。

深度神经网络中，敌手的可用信息主要包括训练数据、预测数据、网络结构及参数等。根据敌手的可用信息划分攻击场景为白盒场景、灰盒场景以及黑盒场景。在白盒场景中，敌手完全了解所使用的 DNN（架构、超参数、权重等），可以访问训练数据，敌手有能力完全复制受攻击的模型。在灰盒场景中，敌手可以收集有关网络架构的部分信息（如了解基准模型使用哪种开源架构），了解受攻击模型使用某种开源数据集进行训练。信息既不完整也不确定，攻击者具有部分模拟受攻击模型的能力。黑盒场景中，攻击者不知道受攻击的模型，此时模型对于敌手相当于谕言机。敌手可以提供有限的输入并收集输出信息。

敌手策略指敌手根据自身的目标、能力以及知识，采取合适具体的方法构造对抗样本，如利用梯度信息或使用生成对抗网络等。

3　存在性解释

自对抗样本发现以来，针对对抗攻击的生成机理的研究成为人工智能领域研究的热点和难点。对抗攻击的生成机理缺乏共识，目前针对对抗样本生成机理主要有几种假说，即盲区（pockets）假说、线性假说、边界倾斜假说、决策面假说、流形假说。

2014 年，Szegedy 等[6]提出，对抗样本存在于数据流中访问较少的盲区，然而，采样的数据不足以覆盖这些盲区，分类器无法有效处理处于盲区的数据样本，因此导致分类器泛化能力较差，出现错误分类的现象。图 3 中的对抗样本可能存在于某些低概率访问的区域。2014 年，Gu 等[14]研究了这类盲区的范围，发现这类盲区普遍存在于输入空间，并具有局部连续性。他们认为对抗样本的存在与训练过程和目标函数有关，与模型结构无关。

Goodfellow 等[15]反驳了上述观点，并给出了线性假说，他们认为尽管深度学习模型具有大量的非线性转换，但仍有许多线性行为。因此，对具有多维特征的数据输入叠加微小扰动可能会使分类器得出错误结果。基于这一理论，Goodfellow

等提出了可以有效生成对抗样本的 FGSM 方法。文献[16-18]给出了线性假说的经验证据，同时文献[17]中指出，对抗样本跨越了多维度连续子空间。不同模型的子空间部分重叠，使对抗样本具有迁移性。2015 年，Luo 等[19]提出了线性假说的变体，深度神经网络在输入流形的某些范围内存在线性行为，在其他范围则存在非线性行为。

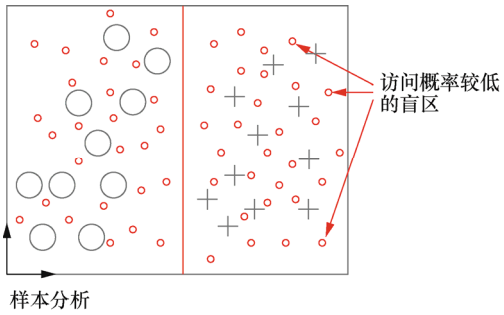

图 3　盲区假说中的对抗样本分布
Figure 3　Adversarial example distribution in the blind zone hypothesis

Tanay 等[20]认为线性行为不足以解释对抗样本现象，并建立了对对抗样本不敏感的线性模型。他们提出边界倾斜假说，即对抗样本存在于采样数据子流形的分类边界。由于该边界无法完全与实际数据流形边界保持一致，所以可能存在导致错误输出的对抗样本。图 4 中的对抗样本存在于实际分类边界与采样数据子流形的分类边界之间。他们认为对抗样本可能存在于数据方差分布较小的方向上，因此推测对抗样本是一种局部过拟合的现象。

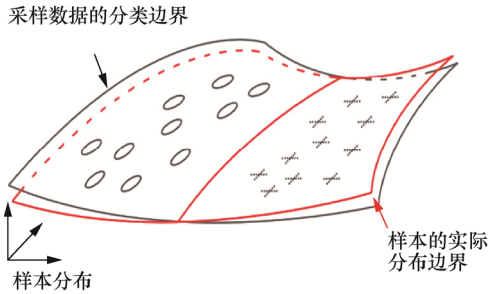

图 4　边界倾斜示意
Figure 4　Boundary tilt example

Moosavi-Dezfooli 等[21-22]发现存在可应用于所有输入的通用性扰动并提出了决策面假说，他们假设可能存在一个低维子空间，它包含决策边界的大多数法向量，并利用该子空间检验了决策边界曲率和对抗样本的相关性。文献[23]给出了决策面假说的实验证据，2018 年 Moosavi-Dezfooli 在文献[24]中对决策面假说给出了进一步的理论分析。

流形假说主要分为两大类。文献[25-28]认为，对抗样本偏离正常的数据流形，基于这一假说，上述文献分别提出了不同的对抗样本检测方法。2018 年，Gilmer 等[29-30]则否认了对抗样本偏离数据流形的假设，他们认为对抗样本由数据流形高维几何结构产生。Gilmer 在文献[29]中构造了实验性的合成数据集，在文献[30]中对对抗样本与数据流形高维几何结构的关系进行了分析。

目前，各种假说对于对抗样本的生成机理缺乏共识，由于深度学习模型的不可解释性以及数据流形几何结构的高度复杂性，不同假说对于对抗样本的生成机理研究具有不同的侧重点，缺乏数理完备的统一理论解释。理论解释的相对缺失意味着无法对现有深度学习应用提供完备有效的验证和检测手段，也制约了如自动驾驶等安全敏感应用的进一步发展。

4　对抗样本构造方法及相关研究

本节将介绍几种典型的对抗样本的构造方法及其实际应用实例。本文提取了对抗样本生成方法的 5 类属性，即生成特征、攻击目标、迭代次数、先验知识以及适用范围。其中生成特征有 3 类，即利用优化求解技术在输入空间中搜索对抗样本；利用敏感特征（如梯度信息）构造对抗样本；利用生成模型直接生成对抗样本。攻击目标则根据是否导致模型出现特定类型的错误来区分，即有目标攻击、无目标攻击。根据算法计算的迭代过程可分为单次迭代和多次迭代。单次迭代方法往往可以快速生成对抗样本，可以用于对抗训练以提高模型稳健性；多次迭代方法则需要更多的处理时间，但攻击效果好且难以防范。根据对抗样本的适用范围，将攻击分为针对特定模型的特定攻击和针对多种模型的通用攻击。表 1 给出了几类典型对抗样本构造方法。

(1) L-BFGS 方法

Szegedy 等[6]首次提出并定义了对抗样本，提出了以下形式的解析式，并使用限制内存 BFGS

表 1　典型对抗样本构造方法

Table 1　Typical adversarial examples construction methods

攻击名称	生成特征	攻击目标	迭代次数	先验知识	适用范围
L-BFGS	优化搜索	有目标	多次	白盒	特异攻击
Deep Fool	优化搜索	无目标	多次	白盒	特异攻击
UAP	优化搜索	无目标	多次	白盒	通用攻击
FGSM	特征构造	无目标	单次	白盒	特异攻击
BIM	特征构造	无目标	多次	白盒	特异攻击
LLC	特征构造	无目标	多次	白盒	特异攻击
JSMA	特征构造	有目标	多次	白盒	特异攻击
PBA	特征构造	有目标&无目标	多次	黑盒	特异攻击
ATN	生成模型	有目标&无目标	多次	白盒&黑盒	特异攻击
AdvGAN	生成模型	有目标	多次	白盒	特异攻击

（L-BFGS）方法求解最小扰动。特定分类器 f 的损失函数为 $loss_f$，对于给定超参 c 以及目标标签 l'，求解使模型分类为 l'的最小扰动 η 如式(2)所示。

$$\text{Min}c\|\eta\|_p + loss_f(x+\eta, l) \tag{2}$$

方程式必须满足的约束条件为

$$x+\eta \in [0,1]^m;\quad f(x+\eta)=l'$$

L-BFGS 方法求解出满足条件的最小扰动 η 构造对抗样本，对抗样本为 $x'=x+\eta$。L-BFGS 方法的实验结果中，平均最小失真率低至 0.062，这意味着对抗样本与正常输入几乎相同。但是 L-BFGS 方法的求解过程较为缓慢，很难应用于大型数据集。L-BFGS 方法可以针对特定图片生成针对性的对抗样本，2018 年，Eykholt 等[31]使用类似的方法攻击了无人驾驶任务中的路牌识别模型，对特定的路牌展开了针对性的对抗攻击。

(2) Deep Fool 方法

Moosavi-Dezfooli 等[32]提出无目标的 Deep Fool 方法，他们认为深度模型存在可分割不同类别数据的超平面。Deep Fool 由二分类模型出发计算最小扰动，最小扰动即当前输入点到分割超平面的最短距离，从而推导出二分类任务下的扰动生成方法，并从二分类推广至多分类。如图 5 所示，在线性二分类的模型 $f(x)$中，改变分类器决策的最小扰动为样本点 x_0 到分割超平面 $F=\{x: w^{\mathrm{T}}\cdot x+b=0\}$ 的正交投影。

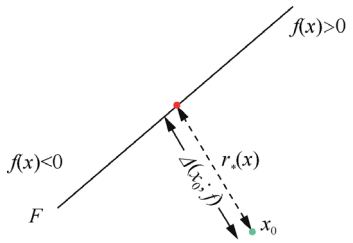

图 5　二分类模型中的最小扰动距离

Figure 5　Minimum disturbance distance in a binary classification model

具体解析式为

$$\text{sign}f(x_0+r)\neq \text{sign}f(x_0) \tag{3}$$

方程式必须满足的约束条件为

$$r_*(x_0)=\arg\min\|r\|_2=-\frac{f(x_0)}{\|w\|_2^2}w$$

其中 $r_*(x_0)$ 为当前点 x_0 到分割超平面的距离。Deep Fool 方法用迭代过程求解最小扰动并将其推广为更一般的二元分类。在此基础上进一步将算法推广至多分类模型，他们认为输入被映射到由多个决策面围成的超平面 P 中。和二分类的情况类似，他们选择最接近 P 的超平面并在其表面上投影求解最小扰动。Deep Fool 方法对原始输入的改动相对较少，同时生成的对抗样本具有较好的攻击效果，计算量也相对较高。

(3) 通用对抗扰动攻击（UAP，universal adversarial perturbations）

Moosavi-Dezfooli 等[21]进一步证明了存在跨越数据以及网络架构的通用对抗扰动。这种扰动

可以导致不同图片产生错误分类并具有跨模型的泛化特性。该方法对训练集中所有图片进行迭代版的 Deep Fool 攻击，直到找到一个可以欺骗大部分训练集的扰动。形式上对于给定输入 x 满足分布 μ，算法搜索上限为 ξ 的通用扰动 η 的计算如式(4)所示。

$$\eta:\ \|\eta_P\| \leqslant \xi \tag{4}$$

方程必须满足的约束条件为

$$\underset{x\sim\mu}{P}\left(f(x+\eta)\neq f(x)\right)\geqslant 1-\delta$$

其中，δ 量化从分布 μ 采样的所有图像期望的愚弄率，算法针对输入分布 μ生成一组采样，并对采样的数据点进行迭代。通过对训练数据集进行多次遍历，该算法能够发现多种通用性的扰动，并能对深度神经网络进行高精度的欺骗。

(4) FGSM（fast gradient sign method）

FGSM 基于 Goodfellow 的线性假说[15]，文献认为在梯度方向上累加较小的扰动会产生对抗样本。FGSM 克服了早期 L-BFGS 计算求解过程繁杂的缺点，添加的扰动形式为 $\eta=\varepsilon\text{sign}(\nabla xJ(\theta,x,y))$，其中，模型的参数值为 θ，模型的输入是 x，模型输出 x 的对应标签 y，$J(\theta,x,y)$ 为神经网络的损失函数，ε 为扰动步长，构造的对抗样本 $x'=x+\eta$。

在非指定目标攻击场景中，沿着梯度方向添加像素值将会使原始类别标签的损失值增大，从而降低模型判定对抗样本为原始类别的概率。文献[15]在 CIFAR-10 数据集上训练卷积网络，得到了错误率为 87.15%的对抗样本，图 6 展示了 FGSM 构造的对抗样本，原本以 57.7%置信度识别出的“熊猫”图片添加扰动后，以 99.3%置信度被识别为“长臂猿”。

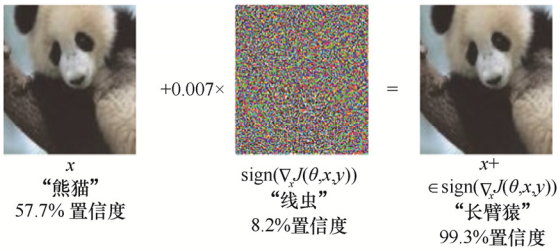

图 6　FGSM 攻击实例
Figure 6　FGSM attack example

FGSM 能简单有效地生成大量对抗样本并可应用于多种深度学习模型。2016 年，Papernot 等[33]利用改进后的 FGSM 方法构造了针对循环神经网络模型的对抗样本，将对抗样本从原有的连续网格数据推广到序列数据。2017 年，Papernot 等[34]在上述工作的基础上进一步构建了针对恶意代码检测模型的对抗样本。

(5) BIM（basic iterative method）方法

2017 年，Google Brain 的 Kurakin 等[7]首先在现实应用场景中构造了对抗样本，然后在 FGSM 的基础上提出迭代 FGSM 用于快速生成对抗样本。文献定义了一个上限为 255 的裁剪函数，将对抗样本限制在原始图片的 L_p 邻域中。其形式化定义如下：对于一个给定由 FGSM 构造的对抗样本 $x'=x+\varepsilon\text{sign}(\nabla xJ(\theta,x,y))$，其裁剪函数 $Clip_{x,\varepsilon}$ 为

$$\begin{aligned}Clip_{x,\varepsilon}\{x'\}(x,y,z)=&\min\{255,x(x,y,z)+\\&\varepsilon,\max\{0,x(x,y,z)-\varepsilon,\ x'(x,y,x)\}\}\end{aligned} \tag{5}$$

其中，$x(x,y,z)$是图片 x 在坐标（x,y）处通道 z 的值，ε 为模型的扰动上界。基础迭代方法通过构造的裁剪函数迭代构造对抗样本，设初始输入为 x，BIM 的迭代过程如下。

$$\begin{aligned}&x_0'=x,x_{N+1}'=\\&clip_{x,\varepsilon}\{x_N'+\alpha\text{sign}(\nabla xJ(x_N',y_{\text{true}}))\}\end{aligned} \tag{6}$$

其中，α 为步长。BIM 克服了 FGSM 对输入改动较大的缺点，同时攻击效果较好，在诸多对抗样本攻防比赛中获得了广泛的应用。

(6) 迭代最小可能类（LLC，iterative least-likely class）方法

FGSM 和 BIM 关注与生成随机的错误分类。这些方法应用于诸如 MNIST 和 CIFAR-10 之类的数据集，其中，类的数量很小并且所有类的高度不同。在 ImageNet[35]上，由于类别数量更多，类别之间的差异程度不同，将一种雪橇犬识别为另一种雪橇犬和将火车识别为飞机的错误程度显然不同。为了构造更有实用意义的错误，文献引入了迭代最小可能类方法。LLC 方法并不手动选择目标类，而是根据模型对输入的预测，选择最不可能的类 $y_{LL}=argmin\{P(y\mid X)\}$。对于一个已经训练好的分类器，最不可能的类通常与真正的类

区分度更大，因此这种攻击方法更为有效。对于LLC 攻击，其迭代过程为

$$x'_0 = x,\ x'_{N+1} = clip_{x,\varepsilon}\{x'_N - \alpha \mathrm{sign}\,(\nabla x J(x'_N, y_{\mathrm{LL}}))\} \quad (7)$$

(7) JSMA（Jacobian-based saliency map attack）方法

Papernot 等[36]提出 JSMA 方法使用雅可比矩阵来评估模型对每个输入特征的敏感度，找到整幅图片中有利于攻击目标实现的显著像素点。这种方法通过计算正向导数（雅可比矩阵）来寻找显著点，以求找到导致 DNN 输出发生重大变化的输入特征。与 FGSM 计算反向梯度形成对比，该算法每次修改一个原始的图像像素，并监控变化对分类结果的影响。通过使用网络层输出的梯度来计算显著性列表。一旦计算出显著性列表，该算法就会选择最有效的像素来欺骗网络。JSMA 方法对原始输入修改较少，同时由于 JSMA 方法采用正向传播计算显著点，计算过程相对较为简单。

(8) 实用的黑盒攻击（PBA，practical black-box attacks）

Papernot 等[37]首次在黑盒的情况下，攻击了远程托管的深度神经网络模型。敌手对神经网络模型的知识限定为深度学习模型的输入与输出。敌手将目标神经网络结构视作谕言机，为了训练替代模型，首先构造了一批随机数据，通过询问谕言机构造出合成样本。利用合成样本，敌手训练替代模型 F 模拟原始神经网络的输入与输出，模仿原始模型的决策边界。利用合成数据训练替代模型后，基于已有的替代模型，使用 FGSM 方法，产生对抗样本，攻击了由 MetaMind、亚马逊、谷歌托管的深度学习模型，并产生了超过 80%的错误分类。PBA 方法的构造相对复杂，但是相对其他构造方法更加难以防御。2018 年，Ilyas 等[38]提出，现实世界中的对抗攻击与黑盒攻击相比存在更多限制，定义了 3 种更为贴近现实场景的威胁模型，并克服了查询次数的现状，成功地攻击了谷歌托管的深度学习 API。

(9) ATN（adversarial transformation network）

Baluja 等[39]提出以自监督学习方式训练对抗性转换网络的前馈神经网络，将输入转换为对抗样本。ATN 在给定原始输入的情况下最小化地修改分类器的输出，同时约束新的分类以匹配对抗目标类。文献[39]展示了 ATN 在白盒、黑盒场景下的应用，并分析了其针对各种分类器的有效性。

(10) 对抗样本生成网络（AdvGAN，generating adversarial examples with adversarial network）

Xiao 等[40]提出对抗样本生成网络，使用生成对抗网络构造对抗样本，他们将提出的生成对抗网络称为 AdvGAN。一旦训练了生成器，它就可以为任何实例有效地产生对抗性扰动，从而对潜在的防御性对抗训练进行加速。文献[40]在半白盒和黑盒攻击设置中应用 AdvGAN。与传统的白盒攻击相比，半白盒攻击在生成器训练之后不需要访问原始目标模型。在黑盒攻击中，AdvGAN 动态训练黑盒模型相应的蒸馏模型并优化其生成器。与其他攻击相比，AdvGAN 在有防御情况下具有较高的攻击成功率。

5　防御方法

现有的对抗样本防御方法主要有两大类，分别为数据检测及预处理、增强模型稳健性，下面详细介绍防御者抵御对抗攻击的一些常见防御方法。

5.1　基于数据检测及预处理的防御方法

这类防御方法主要通过技术手段对输入数据进行检测或清洗，预先发现对抗样本或破坏某些构成对抗样本的关键结构。

(1) 基于密钥的模型检测

2018 年，Zhao 等[41]提出基于密钥的模型检测机制，用于隐藏输入对应的标签。他们利用随机选择的标签集上多个二进制分类器产生的二进制代码向量作为签名，匹配正常图像，拒绝对抗样本；基于纠错输出码，实现对抗性样本与正常样本的区分。

为了检测对抗样本，可以验证输入计算的代码向量是否满足以特定精度和某类的签名匹配。如果输出为否，则将输入视为对抗性样本。对于实际的黑盒和灰盒场景，攻击者不知道编码方案，很难设计出满足相应标签的对抗样本。实验中，该方案对于迭代和自适应攻击具有良好的抵抗效果，但是该方法在实验中使用的数据集规模较小。

(2) MagNet

Meng 等[42]于 2017 年提出名为 MagNet 的框架，通过逼近正常样本的流形训练模型以检测对抗样本。MagNet 由一个或多个独立的检测器网络和一个重组器网络组成。检测器网络测量给定测试样本与正常流形之间的距离，如果距离超过阈值则拒绝该样本。重组器网络使用自动编码器将对抗样本偏向流形，使之成为相似的合法样本。MagNet 不改变受保护的分类器，无须了解构造对抗性样本过程的相关知识，因此具有相当强的泛化能力。MagNet 对于黑盒及灰盒攻击具有较好的防御效果，在白盒攻击的情况下，其性能会显著下降。

(3) 特征挤压

Xu 等[43]提出采用特征挤压，减少敌手可用搜索空间。这种防御背后的主要思想是降低表示数据的复杂程度，使对抗样本由于较低的灵敏度消失。对于图像数据集，文献主要采用了两种方法：降低像素级别的颜色深度，即使用较少的值对颜色进行编码；在图像上使用平滑滤波器。因此，原始空间特征向量相对减少，这使模型获得了抵抗噪声和对抗样本的能力，但是降低了模型的准确率。

(4) 迁移性抑制

对抗样本攻破诸多防御策略的重要原因是对抗样本具有很强的迁移性，即在一个模型上生成的对抗样本被另一个模型检测也会生成同样的错误，攻击者并不需要了解过多的先验知识，就可以利用对抗样本的迁移性攻击可能的模型。Hossein 等[44]于 2017 年提出了 NULL 标记方法，阻止对抗样本从一个网络转移到另一个网络。该方法在输出类集合中添加一个新的 NULL 标签，并训练模型将对抗样本分类为 NULL 标签。迁移性抑制的优点是将扰动输入标记为 NULL 标签，而不是将它们分类为原始标签。这种方法可以准确地区分出对抗性样本，同时不会损害原始数据的准确性。

(5) 胶囊神经网络

胶囊网络[45]是 Hinton 于 2017 年提出的一种新的网络构架，胶囊网络将卷积神经网络的标量输出替换为向量输出。同时文献[45]的研究表明，添加一个利用姿态参数和顶层胶囊特征重建输入图像的网络，可以提升胶囊网络的判别性能。2018 年，Nicholas 等[46]训练胶囊网络根据顶层胶囊的姿态参数和身份来重构图像。由于对抗样本与目标类典型成员具有相当大的差异，因此从该类的顶层胶囊生成重构图像时，它们会有更大的重构误差。通过设立合理的重构误差，文献[46]提出名为 DARCCC（detecting adversaries by reconstruction from class conditional capsules）的技术用于对抗样本检测。

5.2 基于模型稳健性增强的防御

大量文献侧重增强模型面对小扰动的稳健性，尽管观察者无法察觉这些扰动，这种扰动却容易误导深度学习网络。在这种情况下，研究者通过改变或隐藏模型的某些结构，提高深度学习模型对小扰动的稳健性。

(1) 梯度隐藏

针对利用梯度信息构造对抗样本攻击方法（FGSM、BIM 等）的有效防御手段是梯度隐藏。如果模型是不可微分的（如决策树、最近邻分类器或随机森林），那么基于梯度的攻击变得无效。然而，文献[47]中指出相当一部分依赖于梯度隐藏机制的防御手段并不完善，通过文献[35]中学习具有类似梯度的替代黑盒模型并对其构造对抗样本可以攻破这类防御。

(2) 梯度对抗训练

Sinha 等[48]在 2018 年提出了一种新的训练框架，该框架提出梯度更新信息在统计上难以区分。因此，通过梯度正则化，可以去除可能导致对抗样本的显著信息。文献[48]引入了辅助网络处理梯度张量，同时主网络作为辅助网络的对手进行对抗训练实验表明，其框架在训练过程中具有较好的稳健性。

(3) 对抗训练

对抗训练是一种暴力训练方法，该方法通过对训练集添加预先构造的对抗样本，提升模型针对对抗样本的稳健性。文献[15]通过学习添加扰动的损失函数 $\tilde{J}(\boldsymbol{\theta},\boldsymbol{x},y)=\alpha J(\boldsymbol{\theta},\boldsymbol{x},y)+(1-\alpha)J(\boldsymbol{\theta},\boldsymbol{x}+\varepsilon\text{sign}(\nabla xJ(\theta,x,y)),y)$，确保原始样本以及通过 FGSM 方法生成的对抗样本产生相同的预测分类。2016 年，Kurakin 等[49]对 ImageNet 数据集添

加大规模对抗样本，对模型进行对抗训练，并讨论了模型针对黑盒攻击的稳健性。对抗训练只对基于已知方法制作的对抗样本有效，对于本地生成替代模型产生对抗样本的黑盒攻击，对抗性训练则很难防御。

(4) 防御性蒸馏

2016 年，Papernot 等[50]在蒸馏（distillation）技术的基础上提出了防御性蒸馏技术。蒸馏是一种模型训练方法，该方法用小型模型模仿大规模的、计算量大的模型，得到与原有模型相似的结果并尽量保留原有模型的准确性[51]。防御性蒸馏并不改变模型的规模，它只是为了让模型的输出更平滑，稳健性更高，对基于快速梯度标志攻击方法与基于显著图攻击方法的抵抗力较强，但是针对黑盒攻击的抵抗能力较弱。

(5) 生成对抗网络

2017 年，Lee 等[52]提出一种既可以检测对抗样本，也可以增强稳健性的模型训练方法。利用生成对抗网络，他们采用两个网络交替进行训练，一个网络生成对抗样本，另一个网络尝试进行分类。通过两个网络之间的博弈，他们构造的分类网络具有更好的稳健性。

现有的大部分对抗样本防御方法缺乏统一的评估机制，相当一部分方法仅通过小型数据集（如 MNIST）测试其效果，防御强度无法得到严谨的量化评估。同时，现有的对抗样本防御措施往往聚焦于深度学习算法的某一环节，缺乏系统性的考量。实际上，由于基础理论模型的限制，对抗样本领域的防御研究滞后于攻击，现有技术难以建立完备、可靠、安全的深度学习模型。

6　结束语

随着深度学习在各种领域的进一步推广，深度学习技术逐渐成为驱动经济社会各个领域从数字化、网络化向智能化加速发展的重要引擎。面对对抗样本带来的潜在安全威胁，本文首先介绍了一系列对抗样本相关概念及性质；然后，总结了现有对抗样本存在性解释，并归纳了存在的问题；接着，列出了经典的对抗样本构造方法及其对抗样本在不同领域的研究现状；最后，梳理了一系列对抗样本防御方法后给出总结并进行展望。对抗样本在深度学习实践中是一种极大的威胁，现有的深度学习技术普遍面临对抗样本带来的安全挑战。进一步深化研究对抗样本机理，开拓更多领域应用的对抗样本研究，建设安全可信的深度学习模型，仍然存在大量亟须解决的问题。最大限度降低风险，确保人工智能安全、可靠、可控发展具有重要的科学意义和应用价值。

即将到来的 5G 网络所产生的海量数据将进一步促进深度学习技术的发展。同时，对抗样本随之带来的安全问题受到了学术界和工业界的极大关注。结合目前对抗样本研究领域亟待解决的问题，本文归纳并总结了以下几个对抗样本领域的研究方向。

1) 增强模型可解释性。现有深度学习的不可解释性带来了更多的业务风险，而增强深度学习系统的可解释性有助于更好地分析深度学习系统的逻辑漏洞。因此，引入有效的数学工具对深度学习模型进行分析，针对对抗样本的成因构造完备的理论模型是对抗样本研究领域的重点。

2) 健全深度学习稳健性评估体系。当前对抗样本的攻击与防御缺乏统一完备的评价标准，建立统一的测试数据集将是对抗样本研究的有益补充。因此，围绕模型的完整性和可用性构建普适、健全的深度学习评估防御体系是现有对抗样本研究亟待解决的问题。

3) 引入密码学手段。数据安全与隐私保护是深度学习系统的重要组成部分，相应地，密码学技术可以为深度学习系统提供有力的补充与保障。因此，在兼顾效率的情况下，使用差分隐私保护、同态加密等密码学技术，保护用户数据隐私，保障数据的完整与可用是对抗样本研究中值得探索的方向。

4) 在实际应用场景中开展对抗样本研究。自动驾驶等深度学习应用往往具有更高的安全需求，在实际场景中开展对抗样本研究具有更大的现实意义。同时，不同领域的对抗样本研究可以加深人们对深度学习技术的认知，从而进一步推动深度学习发展，构建可靠可信的深度学习模型。

参考文献：

[1] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]//The IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2818-2826.

[2] TANG T A, MHAMDI L, MCLERNON D, et al. Deep learning approach for network intrusion detection in software defined networking[C]//2016 International Conference on Wireless Networks and Mobile Communications (WINCOM). 2016: 258-263.

[3] COLLOBERT R, WESTON J. A unified architecture for natural language processing: deep neural networks with multitask learning[C]//The 25th International Conference on Machine Learning. 2008: 160-167.

[4] CHEN C, SEFF A, KORNHAUSER A, et al. Deepdriving: learning affordance for direct perception in autonomous driving[C]//The IEEE International Conference on Computer Vision. 2015: 2722-2730.

[5] CHING T, HIMMELSTEIN D S, BEAULIEU-JONES B K, et al. Opportunities and obstacles for deep learning in biology and medicine[J]. Journal of The Royal Society Interface, 2018, 15(141).

[6] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[J]. arXiv preprint arXiv:1312.6199, 2013.

[7] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial examples in the physical world[J]. arXiv preprint arXiv:1607.02533, 2016.

[8] ALZANTOT M, SHARMA Y, ELGOHARY A, et al. Generating natural language adversarial examples[J]. arXiv preprint arXiv:1804.07998, 2018.

[9] QIN Y, CARLINI N, GOODFELLOW I, et al. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition[J]. arXiv preprint arXiv:1903.10346, 2019.

[10] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436.

[11] PAPERNOT N, MCDANIEL P, GOODFELLOW I. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples[J]. arXiv preprint arXiv:1605.07277, 2016.

[12] PAPERNOT N, MCDANIEL P, JHA S, et al. The limitations of deep learning in adversarial settings[C]//The 1st IEEE European Symposium on Security and Privacy. 2016.

[13] 宋蕾, 马春光, 段广晗. 机器学习安全及隐私保护研究进展[J]. 网络与信息安全学报, 2018, 4(8): 1-11.
SONG L, MA C G, DUAN G H. Machine learning security and privacy: a survey[J]. Chinese Journal of Network and Information Security, 2018, 4(8): 1-11.

[14] GU S, RIGAZIO L. Towards deep neural network architectures robust to adversarial examples[J]. arXiv preprint arXiv:1412.5068, 2014.

[15] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[C]//2015 International Conference on Learning Representations. 2015: 1-10.

[16] TABACOF P, VALLE E. Exploring the space of adversarial images[J]. arXiv preprint arXiv:1510.05328, 2015.

[17] TRAM`ER F, PAPERNOT N, GOODFELLOW I, et al. The space of transferable adversarial examples[J]. arXiv preprint arXiv:1704.03453, 2017.

[18] KROTOV D, HOPFIELD J J. Dense associative memory is robust to adversarial inputs[J]. arXiv preprint arXiv:1701.00939, 2017.

[19] LUO Y, BOIX X, ROIG G, et al. Foveation-based mechanisms alleviate adversarial examples[J]. arXiv preprint arXiv:1511.06292, 2015.

[20] TANAY T, GRIFFIN L. A boundary tilting perspective on the phenomenon of adversarial examples[J]. arXiv preprint arXiv:1608.07690, 2016.

[21] MOOSAVI-DEZFOOLI S M, FAWZI A, FAWZI O, et al. Universal adversarial perturbations[C]//The IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1765-1773.

[22] MOOSAVI-DEZFOOLI S M, FAWZI A, FAWZI O, et al. Analysis of universal adversarial perturbations[J]. arXiv preprint arXiv:1705.09554, 2017.

[23] TRAM`ER F, KURAKIN A, PAPERNOT N, et al. Ensemble adversarial training: attacks and defenses[J]. arXiv preprint arXiv:1705.07204, 2017.

[24] MOOSAVI-DEZFOOLI S M, FAWZI A, FAWZI O, et al. Robustness of classifiers to universal perturbations: a geometric perspective[C]//International Conference on Learning Representations. 2018.

[25] SONG Y, KIM T, NOWOZIN S, et al. Pixeldefend: leveraging generative models to understand and defend against adversarial examples[J]. arXiv preprint arXiv:1710.10766, 2017.

[26] MENG D, CHEN H. Magnet: a two-pronged defense against adversarial examples[C]//The 2017 ACM SIGSAC Conference on Computer and Communications Security. 2017: 135-147.

[27] GHOSH P, LOSALKA A, BLACK M J. Resisting adversarial attacks using gaussian mixture variational autoencoders[J]. arXiv preprint arXiv:1806.00081, 2018.

[28] LEE H, HAN S, LEE J. Generative adversarial trainer: defense to adversarial perturbations with gan[J]. arXiv preprint arXiv:1705.03387, 2017.

[29] GILMER J, METZ L, FAGHRI F, et al. Adversarial spheres[J]. arXiv preprint arXiv:1801.02774, 2018.

[30] GILMER J, METZ L, FAGHRI F, et al. The relationship between high-dimensional geometry and adversarial examples[J]. arXiv:1801.02774v3, 2018.

[31] EYKHOLT K, EVTIMOV I, FERNANDES E, et al. Robust physical-world attacks on deep learning visual classification[C]//The IEEE Conference on Computer Vision and Pattern Recognition. 2018: 1625-1634.

[32] MOOSAVI-DEZFOOLI S M, FAWZI A, FROSSARD P. Deepfool: a simple and accurate method to fool deep neural networks[C]//The 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016.

[33] PAPERNOT N, MCDANIEL P, SWAMI A, et al. Crafting adversarial input sequences for recurrent neural networks[C]//MILCOM 2016-2016 IEEE Military Communications Conference. 2016: 49-54.

[34] GROSSE K, PAPERNOT N, MANOHARAN P, et al. Adversarial examples for malware detection[C]//European Symposium on Research in Computer Security. 2017: 62-79.
[35] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.
[36] PAPERNOT N, MCDANIEL P, JHA S, et al. The limitations of deep learning in adversarial settings[C]//2016 IEEE European Symposium on Security and Privacy. 2016: 372-387.
[37] PAPERNOT N, MCDANIEL P, GOODFELLOW I, et al. Practical black-box attacks against machine learning[C]//The 2017 ACM on Asia Conference on Computer and Communications Security. 2017: 506-519.
[38] ILYAS A, ENGSTROM L, ATHALYE A, et al. Black-box adversarial attacks with limited queries and information[J]. arXiv preprint arXiv:1804.08598, 2018.
[39] BALUJA S, FISCHER I. Adversarial transformation networks: Learning to generate adversarial examples[J]. arXiv preprint arXiv:1703.09387, 2017.
[40] XIAO C, LI B, ZHU J Y, et al. Generating adversarial examples with adversarial networks[C]//The 27th International Joint on Artificial Intelligence Main track.2019: 3805-3911.
[41] ZHAO P, FU Z, HU Q, et al. Detecting adversarial examples via key-based network[J]. arXiv preprint arXiv:1806.00580, 2018.
[42] MENG D, CHEN H. Magnet: a two-pronged defense against adversarial examples[C]//The 2017 ACM SIGSAC Conference on Computer and Communications Security. 2017: 135-147.
[43] XU W, EVANS D, QI Y. Feature squeezing: detecting adversarial examples in deep neural networks[J]. arXiv preprint arXiv:1704.01155, 2017.
[44] HOSSEIN H, CHEN Y, KANNAN S, et al. Blocking transferability of adversarial examples in black-box learning systems[J]. arXiv:1703.04318, 2017.
[45] SABOUR S, NICHOLAS F, HINTON G E. Dynamic routing between capsules[C]//Neural Information Processing Systems. 2017.
[46] NICHOLAS F, SABOUR S, HINTON G. DARCCC: detecting adversaries by reconstruction from class conditional capsules[J]. arXiv preprint arXiv:1811.06969, 2018.
[47] TRAMÈR F, KURAKIN A, PAPERNOT N, et al. Ensemble adversarial raining: attacks and defenses[J]. arXiv:1705.07204, 2017.
[48] SINHA A, CHEN Z, BADRINARAYANAN V, et al. Gradient adversarial training of neural networks[J]. arXiv preprint arXiv:1806.08028, 2018.
[49] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial machine learning at scale[J]. arXiv preprint arXiv:1611.01236, 2016.
[50] PAPERNOT N, MCDANIEL P, WU X, et al. Distillation as a defense to adversarial perturbations against deep neural networks[C]//2016 IEEE Symposium on Security and Privacy. 2016: 582-597.
[51] HINTON G E, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J]. arXiv:1503.02531.
[52] LEE H, HAN S, LEE J. Generative adversarial trainer: defense to adversarial perturbations with GAN[J]. arXiv preprint arXiv:1705.03387, 2017.

[作者简介]

段广晗（1994– ），男，黑龙江海伦人，哈尔滨工程大学博士生，主要研究方向为深度学习、对抗样本、机器学习。

马春光（1974– ），男，黑龙江双城人，山东科技大学教授、博士生导师，主要研究方向为密码学、数据安全与隐私、人工智能安全与隐私、区块链技术与应用。

宋蕾（1989– ），女，黑龙江牡丹江人，哈尔滨工程大学博士生，主要研究方向为机器学习安全与隐私保护、云计算、网络安全。

武朋（1974– ），女，黑龙江齐齐哈尔人，山东科技大学讲师，主要研究方向为网络安全、隐私保护。