深度学习中的对抗攻击与防御

刘西蒙[1,2]，谢乐辉[1]，王耀鹏[1]，李旭如[3]
（1. 福州大学数学与计算机科学学院，福建 福州 350108；
2. 广东省数据安全与隐私保护重点实验室，广东 广州 510632；
3. 华东师范大学计算机与科学学院，上海 200241）

摘　要：对抗样本是被添加微小扰动的原始样本，用于误导深度学习模型的输出决策，严重威胁到系统的可用性，给系统带来极大的安全隐患。为此，详细分析了当前经典的对抗攻击手段，主要包括白盒攻击和黑盒攻击。根据对抗攻击和防御的发展现状，阐述了近年来国内外的相关防御策略，包括输入预处理、提高模型鲁棒性、恶意检测。最后，给出了未来对抗攻击与防御领域的研究方向。

关键词：对抗样本；对抗攻击；对抗防御；深度学习安全
中图分类号：TP18
文献标识码：A
doi: 10.11959/j.issn.2096−109x.2020071

Adversarial attacks and defenses in deep learning

LIU Ximeng[1,2], XIE Lehui[1], WANG Yaopeng[1], LI Xuru[3]
1. College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350108, China
2. Guangdong Provincial Key Laboratory of Data Security and Privacy Protection, Guangzhou 510632, China
3. School of Computer Science and Technology, East China Normal University, Shanghai 200241, China

Abstract: The adversarial example is a modified image that is added imperceptible perturbations, which can make deep neural networks decide wrongly. The adversarial examples seriously threaten the availability of the system and bring great security risks to the system. Therefore, the representative adversarial attack methods were analyzed, including white-box attacks and black-box attacks. According to the development status of adversarial attacks and defenses, the relevant domestic and foreign defense strategies in recent years were described, including pre-processing, improving model robustness, malicious detection. Finally, future research directions in the field of adversarial attacks and adversarial defenses were given.

Key words: adversarial examples, adversarial attacks, adversarial defenses, deep learning security

收稿日期：2020−03−31；**修回日期**：2020−05−12

通信作者：刘西蒙，snbnix@gmail.com

基金项目：国家自然科学基金（U1804263，61702105）；广东省数据安全与隐私保护重点实验室开放项目（2017B030301004-12）；陕西省重点研发项目（2019KW-053）

Foundation Items: The National Natural Science Foundation of China (U1804263, 61702105), Opening Project of Guangdong Provincial Key Laboratory of Data Security and Privacy Protection (2017B030301004-12), The Key Research and Development Program of Shaanxi Province, China (2019KW-053)

论文引用格式：刘西蒙，谢乐辉，王耀鹏，等. 深度学习中的对抗攻击与防御[J]. 网络与信息安全学报, 2020, 6(5): 36-53.
LIU X M, XIE L H, WANG Y P, et al. Adversarial attacks and defenses in deep learning[J]. Chinese Journal of Network and Information Security, 2020, 6(5): 36-53.

1　引言

随着计算机运算能力的提升和社会数据量爆发式增长，深度学习在数据特征提取上表现出独特的优势。如今，深度学习已经被广泛应用于各个领域，在诸如计算机视觉[1]、语音识别[2-3]、文字处理[4]、恶意软件检测[5]等场景下均有不俗的表现。其在围棋[6]、游戏[7]等领域已经达到人类顶尖的水平。以计算机视觉领域为例，在 2012 年的一个大规模图像[8]识别任务中，Krizhevsky 等[1]利用卷积神经网络[9]将识别率提高至 84.7%，达到前所未有的高度。

尽管深度学习在许多领域中取得了令人瞩目的成绩，但 Szegedy 等[10]在其针对图像分类的研究成果中指出，尽管深度神经网络有着极高的准确率，但却非常脆弱，容易受到一种往图像中添加人眼无法察觉的微小扰动攻击。这类攻击不仅能够以极高的置信度误导深度神经网络输出错误的分类结果，并具有可迁移性，即同一张扰动图像可能攻击多个网络模型。这类攻击被称为对抗攻击，这个扰动图像即为对抗样本。如图 1 所示，假设图像分类模型能够以 57.7%的置信度正确识别出图片中的熊猫，但在图片中有针对性地添加一些人眼不可察觉的微小扰动后，虽然肉眼看上去图片并没有变化，但该图像分类模型却以 99.3%的置信度错误地把将图片分类成了长臂猿[11]。

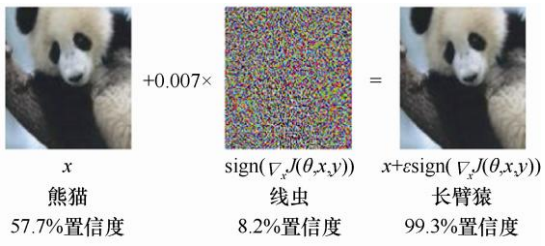


图 1　对抗样本实例
Figure 1　Adversarial example

自 Szegedy 等[10]提出针对深度神经网络的对抗攻击之后，针对计算机视觉中深度学习的对抗攻击及其防御逐渐成为研究热点，由此产生了许多对抗攻击和防御方法。既有白盒攻击中 Moosavi-Dezfooli 等[12]提出对于多个图像适用的通用扰动，也有黑盒攻击中 Su 等[13]提出只针对图像中的一个像素点进行扰动并误导分类器的极端攻击方法。而防御方法主要分为 3 类，即输入预处理、提高模型鲁棒性以及恶意检测。涉及的具体方法多种多样，包括预处理修正网络、像素偏移、图像变换、对抗生成网络防御、特征去噪、防御蒸馏、对抗性训练和随机化等。虽然目前提出了大量防御方法，但却无法完全防御对抗攻击，许多防御方法并不完善、存在片面性，其有效性也遭到了质疑[14-16]。

本文总结了目前主流的对抗攻击和防御方法，并从这两个角度出发，进行相关的介绍和总结。

2　相关概念

本节主要介绍一些相关的基本概念、对抗攻击威胁模型和攻击类型。

2.1　基本概念

Szegedy 等[10]首次发现在原始图像上添加人眼无法察觉的噪声，添加噪声后的图像能够误导神经网络模型以高置信度错误分类该图像，这类通过添加噪声导致网络分类错误的图像被称为对抗样本，其所添加的噪声称为对抗扰动。Moosavi-Dezfooli 等[12]发现一种通用型对抗扰动，这种扰动添加到特定数据集的所有图像构成对抗样本，具体细节在第 3 节中介绍。此外，对抗样本具有迁移性，即对抗样本能够攻击特定目标模型以外的神经网络模型。攻击者为了使生成的对抗样本与原始图像肉眼无法区分，对抗样本 x' 与原始样本 x 的相似程度使用 ℓ_p 范数衡量，即 $\|x'-x\|_p$，$\|\cdot\|_p$ 的定义如式(1)所示。

$$\|\boldsymbol{v}\|_p=\sqrt[p]{\sum_{i=0}^{m}|v_i|^p} \tag{1}$$

其中，$\boldsymbol{v}$ 是 m 维向量；p 是实数。

针对对抗样本问题，Goodfellow 等[11]首次提出通过对抗训练来提高模型鲁棒性，即在原有的模型训练过程中加入对抗样本的训练方法，旨在提升模型对于微小扰动的鲁棒性，具体的方法将在第 5 节中介绍。

2.2　威胁模型

根据攻击者对攻击目标模型所掌握信息的程度，可以把威胁模型分为黑盒模型、灰盒模型、白盒模型。① 黑盒攻击：攻击者只能够通过神经

网络输入，获得相应的输出，根据输出的反馈构造对抗样本。② 灰盒攻击：相比黑盒攻击，灰盒攻击的攻击者所拥有的信息更加丰富。攻击者能够获得攻击目标模型结构、参数，但不了解目标模型防御机制相关的任何信息，灰盒模型通常用于评测防御方法。③ 白盒模型：攻击者掌握目标模型结构、参数、防御机制等全部信息，有时候还包括训练集，在这种情况下，攻击者的能力最强。

2.3　攻击类型

自 Szegedy 等[10]提出对抗样本的概念之后，便产生了许多对抗样本生成算法。攻击算法可根据有无针对性目标分为定向攻击和非定向攻击。定向攻击指攻击者产生的对抗样本能够让神经网络把对抗样本分类成攻击者指定类别。非定向攻击指神经网络将对抗样本分类为除正确类别以外的任意类别。非定向攻击相比定向攻击更加容易实现。而根据攻击威胁模型，攻击算法可分为白盒攻击和黑盒攻击。白盒攻击者能够获得网络模型结构、参数、防御机制，有时候甚至还包括训练集。黑盒攻击者只能够根据神经网络输入，获得相应输出，并不知道网络模型的结构和参数。在攻击威胁模型的基础上，可根据攻击者所能掌握目标模型信息程度细分成 4 个类别：梯度攻击、置信度攻击、决策攻击、迁移攻击，如图 2 所示。梯度攻击建立在攻击者能够获取目标模型结构和参数的基础上，利用反向传播的梯度构造对抗样本，属于白盒攻击。置信度攻击、决策攻击、迁移攻击是黑盒攻击模型中的子类别。置信度攻击通过网络模型输出的置信度构造对抗样本。相比置信度攻击，决策攻击只需根据最终的输出类别，不需要各类别的置信度。迁移攻击则在代理模型上生成对抗样本，利用对抗样本的迁移性对目标模型发起攻击。迁移攻击所需要的信息量最少，不需要目标模型的任何信息，就可以实现对目标模型的攻击。

3　对抗攻击

自对抗样本被发现以来，出现了许多不同的对抗样本生成算法及其相应的改进版本。本文根据威胁模型的不同，将对抗攻击算法分为白盒攻击与黑盒攻击两类。

3.1　白盒攻击

3.1.1　L-BFGS 攻击

Szegedy 等[10]发现了两点深度神经网络的反直觉特性。① 神经网络中深层次的语义信息由整

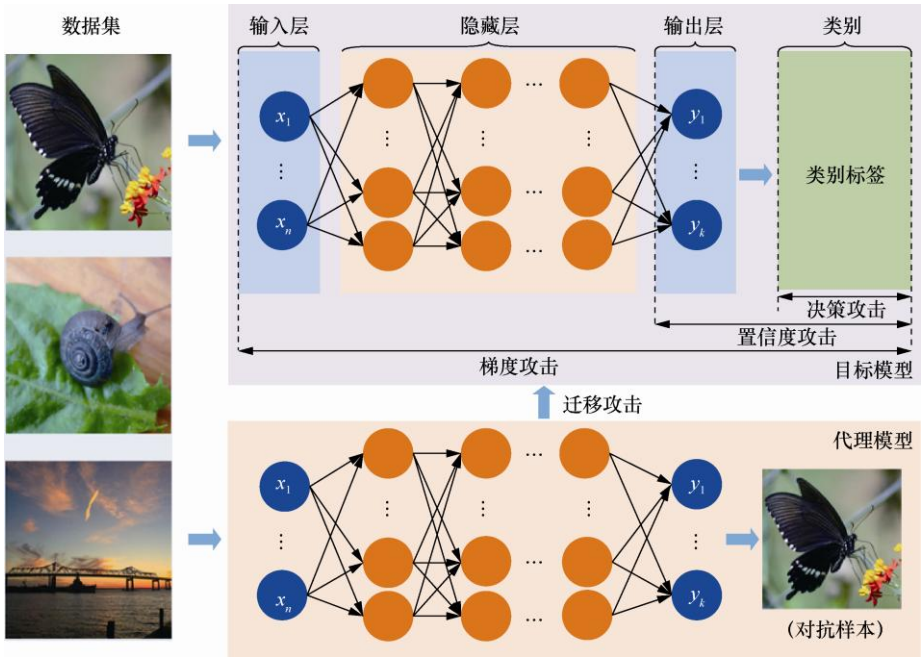


图 2　攻击类型
Figure 2　Attack type

个单元空间构成，而不是单个神经元。② 深度神经网络学习到的输入输出之间的映射关系具有不连续的特性，并由此提出一种在原始图像上添加肉眼难以观察到的微小扰动就能够误导分类器的图片，即对抗样本。文献[10]首次提出对抗样本概念，并将对抗样本生成方法建立成模型，采用 L-BFGS 算法[17]简化成如式(2)、式(3)所示。

$$\min_r c\cdot|\delta|+J(x+\delta,t) \tag{2}$$

$$x+\delta\in[0,1]^m \tag{3}$$

其中，x 表示原始图像，这里将 RGB 图像转换成 m 维向量表示；c 是超参数，该算法利用线性搜索找到一个可以产生最小距离对抗样本的常数；式(2)中最小化扰动 δ 和攻击目标类别 t 的损失函数 J（通常是交叉熵）来误导深度学习网络模型；式(3)保证添加扰动后的图像取值 $x+\delta\in[0,1]^m$ 稳定在正常范围。

3.1.2　快速梯度符号算法攻击

Szegedy 等[10]认为对抗样本的存在是由神经网络模型的非线性和过拟合造成的。而 Goodfellow 等[11]认为，即使是简单的线性模型也会存在对抗样本，并且将线性特征应用于非线性模型，提出了第一个基于梯度的快速梯度符号算法（FGSM，fast gradient sign method），该算法与 L-BFGS 有两点不同：① 采用 L_∞ 优化方式；② 重点在于计算效率，对抗性能不足。核心公式如式(4)所示。

$$\delta=\epsilon\cdot\mathrm{sign}\left(\nabla_x J(x,y_{\mathrm{true}})\right) \tag{4}$$

其中，$\nabla_x J(x,y_{\mathrm{true}})$ 即损失函数的梯度，y_{true} 是真实类别标签；sign(·) 为符号函数，决定像素点改变的方向；ϵ 是控制扰动大小。

3.1.3　基础迭代法攻击

Kurakin 等[18]提出基础迭代法（BIM，basic iterative method）。它是对 FGSM[11]的扩展：不再利用单一步长 ϵ，而是分成多次小步 α 进行迭代，采用 ϵ 裁剪对应结果，如式(5)、式(6)所示。

$$x'_0=x \tag{5}$$

$$x'_i=\mathrm{clip}_{X,\epsilon}\left\{x'_{i-1}+\alpha\mathrm{sign}(\nabla_x J(x'_{i-1},y_{\mathrm{true}}))\right\} \tag{6}$$

其中，x'_i 表示对抗样本的第 i 次迭代；clip 为裁剪函数，定义如式(7)所示。

$$\mathrm{clip}_{x,\epsilon}(x'_i)=\min\{1,x+\epsilon,\max(0,x-\epsilon,x'_i)\} \tag{7}$$

裁剪函数对图像 x' 的每个像素进行裁剪，保证 x'_i 维持在原始图像 x 的 ϵ 邻域。实验表明，BIM 比 FGSM 更有效，但计算效率有所降低。

3.1.4　基于雅可比矩阵的显著图攻击

Papernot 等[19]提出攻击算法：基于雅克比的显著图（JSMA，Jacobian based saliency map attack）算法。该文献利用 L_0 范数将扰动控制在图像中的几个像素点上，保证对抗样本的真实性。该方法首先利用雅克比矩阵计算深度神经网络的前向导数，如式(8)所示。

$$\nabla F(x)=\frac{\partial F(x)}{\partial x}=\left[\frac{\partial F_j(x)}{\partial x_i}\right]_{i\in1,\cdots,M_{\mathrm{in}},j\in1,\cdots,M_{\mathrm{out}}} \tag{8}$$

其中，F 表示 Softmax 层输出函数，M_{in} 表示输入层数量，M_{out} 表示输出层数量。然后通过前向导数计算对应的对抗性显著图（adversarial saliency map），并利用贪心算法选择对抗性显著图中数值最大的一个像素点进行扰动。算法迭代以上步骤直到达到最大扰动像素数量或者成功误导模型，最终得到对抗样本。

3.1.5　Carlini&Wagner（C&W）攻击

防御蒸馏可以为目标网络模型提供很强的鲁棒性，能够将当时已经出现的攻击算法成功率由 95%锐减到 0.5%。Carlini 和 Wagner[20]提出 ℓ_0、ℓ_2、ℓ_∞ 3 种优化方式的 C&W 攻击算法，能够对蒸馏或未蒸馏的神经网络达到 100%的攻击效果，计算如式(9)所示。

$$\min_\delta\|\delta\|_p+c\cdot f(x+\delta) \tag{9}$$

其中，δ 即对抗性扰动，对应原始图像和对抗样本的差值，该部分越小，意味着越不容易被察觉。而 $f(\cdot)$ 表示目标函数。文献[20]提供了 7 种候选目标函数，该文献在实验中实际应用的函数之一如式(10)所示。

$$f(x')=\max\left(\{\max Z(x')_i:i\neq t\}-Z(x')_t,-k\right) \tag{10}$$

其中，$Z(x')_i$ 表示类别 i 的逻辑值，将最大逻辑值（对应类别不同于 t）记为 $\max\{Z(x')_i:i\neq t\}$，并通过优化降低 $\max\{Z(x')_i$ 来提高攻击有效性。利用 k 控制错误分类的置信度，参数 k 与对抗样本 x' 攻击的成功率呈正相关，k 越大，x' 分类为 t 的

可能性越大。超参数 c 用来权衡两个部分之间关系，C&W 算法中使用二分查找来确定 c 值。

3.1.6 Deepfool 攻击

Moosavi-Dezfooli 等[21]提出了基于分类问题的 Deepfool 对抗样本生成方法。在二分类问题中，原始图像与决策边界的最短距离为垂直距离，为了生成最小扰动，使原始图像朝着垂直于决策边界的方向前进最短距离，到达决策边界，生成对抗样本使其误导分类模型。但大多数神经网络是高度非线性的，问题由二分类延伸至多分类。多分类问题可以看作多个二分类问题的集合，即寻找原始样本与其所在凸区域的边界之间的最小距离，并通过多次迭代到达分类边界，使攻击成功。

3.1.7 通用型扰动攻击

Moosavi-Dezfooli 等[12]首次证明了存在一种非定向的通用型扰动攻击，将其添加到给定图像集中的大部分图片上能够实现以高置信度误导网络模型，并给出了通用型扰动生成算法。算法将目标数据集 X 记为数据集 $X=\{x_1,x_2,\cdots,x_m\}$，并对数据集 X 中的所有图像添加通用型扰动 δ，使数据集 X 中大部分图像能够误导分类模型。算法寻找的扰动 δ 满足式(11)。

$$\mathbb{P}_{x\sim\mu}\left(\hat{k}(x+\delta)\neq\hat{k}(x)\right)\geqslant 1-\eta,\ \delta_p\leqslant\epsilon \quad (11)$$

假设原始样本从图像分布 μ 中采样得到，$\hat{k}(x)$ 表示对图像 x 的预测类别，η 表示欺骗率。算法利用 DeepFool 攻击方法[21]依次将数据集 X 中的图像逐渐推到各自的决策边界，并依次将图像投影在半径为 ϵ 的 ℓ_p 球面上，迭代生成通用型扰动 δ。

3.2 黑盒攻击

3.2.1 单像素攻击

Su 等[13]提出一种基于差分进化算法的极少像素攻击。这是一种极端的对抗攻击方法，仅改变一个像素就能够使网络模型分类错误。该算法通过迭代地修改单个像素并生成子图像，将其与母图像对比，根据选择标准保留攻击效果最好的子图像以实现对抗攻击。单像素攻击可以通过修改少数不同像素来达到攻击的目的，如修改 1、3 或 5 个像素点，成功率分别为 73.8%、82.0%、87.3%。与以往的攻击方法不同，该算法仅需要黑盒反馈（概率标签），不需要目标网络的内部信息，如梯度和网络结构。

3.2.2 期望变换攻击

噪声、扭曲、仿射变换等图像变换操作会导致对抗样本失效。针对这个问题，Athalye 等[22]提出变换期望（EOT，Expection Over Transformation）攻击算法。其核心如式(12)所示。

$$\arg\min_{x'} E_{t\sim T}\left[\log P\left(y_t\left|t\left(x'\right)\right.\right)-\lambda\mathrm{LAB}\left(t\left(x'\right)\right)-\mathrm{LAB}\left(t\left(x\right)\right)\right\|_2\Big] \quad (12)$$

其中，x' 是对抗样本；x 为原始图像；LAB 为图像颜色对立空间[23]；T 是图像变换分布。

算法基本思想是变换分布 T 可以对感知扭曲进行建模，如随机旋转、平移或添加噪声。该算法不仅能够模拟简单变换，也可执行诸如纹理的三维渲染之类的操作。

3.2.3 零阶优化攻击

受到 C&W 算法[20]的启发，Chen 等[24]提出了基于置信度的零阶优化（ZOO，zeroth order optimization）方法。其通过输入样本和输出类别置信度，针对深度神经网络模型进行黑盒攻击。ZOO 不需要训练替代模型，它利用零阶优化近似估计网络梯度，同时使用降维、分层攻击、重要抽样技术提高计算效率。该算法的优化方案与 C&W 算法[20]一致，但其区别在于该算法为黑盒攻击，无法获取模型梯度，ZOO 使用近似梯度替代模型梯度，利用对称差商计算近似梯度和黑塞矩阵。在得到梯度和黑塞矩阵的前提下，通过随机坐标下降方法生成最优扰动，并利用 ADAM 方法[25]提高收敛效率。

3.2.4 边界攻击

Brendel 等[26]指出当前的主流攻击方式为基于梯度的攻击和基于置信度的攻击。但在现实场景中无法获取攻击所需的网络模型信息，这两种方式都不适用于现实场景。因此，文献[26]提出了仅需输出类型的决策边界攻击，其基本思想是在保证得到对抗样本的情况下，不断地迭代靠近原始图像，最终得到与原始图像相近的对抗样本。这种攻击能够有效地应用于现实场景，并且与迁移攻击相比，边界攻击所需的模型信息量更少、

鲁棒性更强、更容易应用于现实场景。

3.2.5　有偏边界攻击

边界攻击是从多维正态分布中提取扰动候选项，这意味着算法使用的是无偏采样的攻击方式。虽然这种方法灵活性较高，但对于鲁棒性较强的模型效率不高。Brunner 等[27]则提出有偏的边界攻击，大大提高了攻击的效率，主要从 3 方面进行改进。

1) 低频扰动：由于经典的对抗样本生成算法产生的对抗扰动是高频噪声，所以 Brunner 等[27]采用了低频的 Perlin 噪声[28]绕过检测机制。

2) 区域掩码：利用区域掩码，在对抗样本和原始图像差异较大的区域进行更新，在极度相似部分则不进行更新，有效减少了搜索空间。

3) 替代模型梯度：对抗样本具有迁移性，即替代模型的梯度对于攻击目标模型是有帮助的。因此，Brunner 等[27]利用替代模型的梯度指引边界攻击的更新方向，使其提高攻击效率。

虽然上述改进提高了算法效率，但替代模型梯度依赖于模型的可移植性。为了解决这个问题，Chen 等[29]提出不依赖可移植性假设，利用蒙特卡洛估计评估梯度的方向，进一步改善了攻击效率。

3.3　攻击方法总结

攻击的主要目标是以人眼无法察觉的扰动，使深度学习模型对输入图像分类错误。因此，本文根据攻击目标有无、扰动范数、攻击类型 3 个特点对这些经典的方法加以对比，如表 1 所示。从表 1 可以看出，攻击算法非定向的居多，扰动范数主要是 ℓ_0、ℓ_2、ℓ_∞ 3 类。早期提出的攻击算法类型中，梯度攻击最为常见，但现实应用中迁移攻击、置信度攻击、决策攻击更符合实际。

4　对抗攻击的实例

第 3 节介绍了图像分类领域内的对抗样本生成算法，但存在一定的条件限制，如标准数据集、白盒模型等约束。这些条件的约束可能不足以让人为深度学习模型安全担忧。然而，已有针对网络空间攻击、目标检测、停车牌识别、语义分割、人脸识别等现实应用场景的攻击实例，本节对这些领域内的对抗攻击实例进行介绍。

4.1　图像分类

Kurakin 等[30]首次证明了对抗攻击的威胁存在于物理世界。他们使用 Inception-v3 模型[31]生成对抗样本，然后打印出对抗样本的图像，并用手机摄像头拍摄，把拍摄的图像输入 TensorFlow Camera Demo 应用程序。结果表明，即使是用相机拍摄，这些图像仍能够导致模型分类错误。简单利用对抗样本的可移植性的攻击方式容易受到图像变换的影响，效果不够稳定。因此，Athalye 等[22]提出一种构造 3D 对抗样本的方法，通过

表 1　对抗攻击方法总结

Table 1　Summary of adversarial attacks

威胁模型	攻击算法	有无目标	扰动范数	攻击类型
白盒	L-BFGS 攻击[10]	定向	ℓ_0	梯度攻击
	FGSM 攻击[11]	非定向	ℓ_∞	梯度攻击
	BIM 攻击[18]	非定向	ℓ_∞	梯度攻击
	JSMA 攻击[19]	定向	ℓ_0	梯度攻击
	C&W 攻击[20]	定向	ℓ_0、ℓ_2、ℓ_∞	梯度攻击
	通用对抗扰动攻击[12]	非定向	ℓ_2、ℓ_∞	迁移攻击
	Deepfool 攻击[21]	非定向	ℓ_2、ℓ_∞	梯度攻击
黑盒	单像素攻击[13]	非定向	ℓ_0	置信度攻击
	EOT 攻击[22]	定向	ℓ_2	迁移攻击
	边界攻击[26]	定向、非定向	ℓ_2、ℓ_∞	决策攻击
	有偏边界攻击[27]	定向、非定向	ℓ_2、ℓ_∞	决策攻击
	零阶优化攻击[24]	定向、非定向	ℓ_2	置信度攻击

EOT 技术首次合成了现实 3D 对抗样本，该方法模拟多种图像变换的影响，即使在图像变换的干扰下，仍然保持对抗样本的特性，使对抗样本在现实场景中更加鲁棒。

4.2 网络空间攻击

Papernot 等[32]提出了网络空间上的现实对抗样本攻击，在合成数据集上训练了一个代理模型，用于产生对抗样本，并对 MetaMind、亚马逊和谷歌的远程托管神经网络发起攻击。结果表明，模型错误分类率分别是 84.24%、96.19%和 88.94%。同样地，Liu 等[33]利用对抗样本的迁移性实施攻击，其基本思想是生成一个能够同时让多个模型分类出错的对抗样本，用来实施迁移攻击。这一方法实现了在大数据集 ImageNet[8]上的黑盒攻击，并且成功攻击了当时提供最先进图像分类服务的商业公司 Clarifai。

与迁移攻击不同，Li 等[34]分别对单像素攻击和边界攻击进行改进。在单像素攻击[13]基础上，通过逐步增加像素修改的数量，并融入语义分割的思想提高效率。而在边界攻击[26]中则引入语义分割和贪心的思想提高效率。Li 等[34]还对亚马逊、微软、谷歌、百度和阿里巴巴五大云服务提供商提供的与计算机视觉相关的服务（如图像分类、对象识别、非法图像检测）分别进行了黑盒攻击，成功率几乎达 100%。

4.3 目标检测

Wei 等[35]提出了一种基于生成对抗网络[36]框架的对抗样本生成方法，为了增加对抗样本的可移植性，引入网络的特征损失作为损失函数的一部分。通过结合高级别的类别损失和低级别的特征损失训练的生成器，生成的对抗样本具备很好的可迁移性，可以同时攻击两个具有代表性的目标检测器快速区域卷积神经网络[37]（Faster-RCNN，faster regions with convolutional neural networks）和单发多框检测器[38]（SSD，single shot multibox detector）。

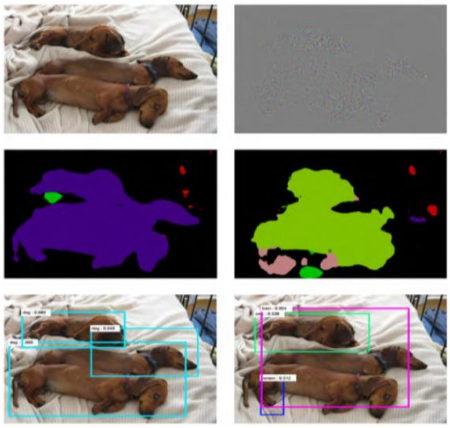
Thys 等[39]提出了基于“你只需看一次”（YOLO，you only look once）[40]模型的动态人物目标检测攻击方法。他们通过优化图像的方式生成一个对抗补丁，将其放置在人体中心，成功绕过检测模型检测。他们把优化目标损失函数分为 3 个部分，即 L_{nps}、L_{tv} 和 L_{obj}。L_{nps} 表示当前补丁的颜色能否应用于现实生活；L_{tv} 体现了图像的平滑度；L_{obj} 为图像中最大的目标检测置信度。在优化过程中，神经网络模型参数不变，仅改变对抗性补丁，并将每次修改过的补丁进行旋转、缩放等基本变换之后，再次应用到数据集图像中，以提高对抗性补丁鲁棒性，使其能够成功误导检测模型。

4.4 停车牌识别

Evtimov 等[41]基于先前攻击算法[20,33]，提出了一种通用的攻击算法（鲁棒物理扰动），用于在不同的物理条件下（如距离、角度、扭曲）产生具有鲁棒性的视觉对抗性扰动。在实际驾驶环境中，鲁棒性物理扰动成功欺骗路标识别系统。为了证实鲁棒性物理扰动具有通用性，他们将用鲁棒性物理扰动生成的涂鸦贴在微波炉上，成功误导 Inception-v3 分类器[31]将微波炉识别成手机。Lu 等[42]对道路标志图像与检测器的物理对抗样本进行了实验，实验表明 YOLO[40]和 Faster-RCNN[37]等检测器目前没有被 Evtimov 等[41]提出的攻击所欺骗。然而 Eykholt 等[43]声称能够产生一个小的贴纸来欺骗 YOLO 检测器[40]，也可以欺骗 Faster-RCNN[37]。Chen 等[44]进一步使用了 EOT 技术[22, 45]技术，使攻击更具有鲁棒性，成功误导 Faster-RCNN 检测器[37]。

4.5 语义分割

Hendrik 等[46]给出了针对语义分割和目标检测的通用对抗样本。随后，Xie 等[47]首先提出了一种系统的算法（稠密对抗生成器），生成用于对象检测和分割任务的对抗样本。如图 3 所示，在添加扰动后，语义分割和目标检测同时预测出错。稠密对抗性生成器的基本思想是同时考虑检测和分割任务中的所有目标，优化总体损失。此外，为了解决目标检测任务中建议数量较多的问题，引入交并比来保持增加但合理的建议数量。文献[48]发现，在分割任务中，广泛使用的对抗损失与准确性之间的关系没有在分类任务中那么明确，因此，提出了使用 Houdini 损失来近似真实的对抗损失，使对抗扰动更不易被人眼察觉。



(a) 语义分割正确　　(b) 语义分割错误

图 3　语义分割对抗样本实例
Figure 3　Adversarial example of semantic segmentation

4.6　人脸识别

Sharif 等[49]开发了一种针对人脸识别系统进行攻击的系统性方法，只需要通过添加一副眼镜框就能够让人脸识别系统识别错误。Zhou 等[50]研究了一个现实中对抗攻击的有趣例子，发现红外光也可以用来干扰人脸识别系统。攻击者可以在一顶帽子的帽檐安装 LED 灯，利用 LED 灯照射到脸部产生人眼无法察觉但能被相机传感器捕捉到的紫色光线，躲避人脸识别系统检测。

5　防御策略

近年来，研究人员提出了许多针对对抗样本问题的防御策略。本文将防御策略分为预处理、提高模型鲁棒性、恶意检测 3 类，如图 4 所示。

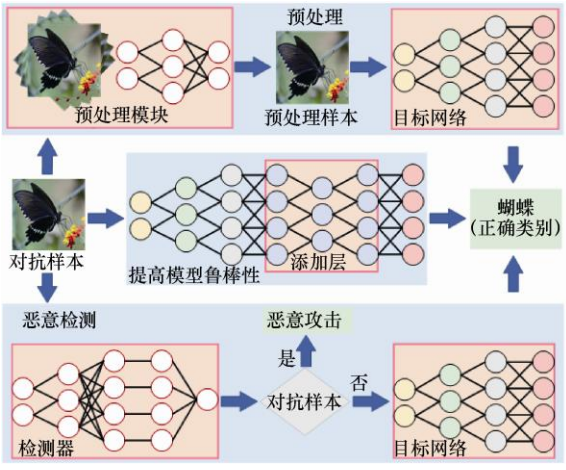


图 4　对抗防御框架
Figure 4　A framework of adversarial defense

预处理是在图像输入网络之前对图像进行去噪、随机化、重构、缩放、变换、增强等操作，减轻对抗扰动对模型分类的影响，通常无须对模型进行任何的修改，可以直接应用于已经训练好的模型，计算开销较低。提高模型鲁棒性则通过修改模型架构、训练方式、正则化、特征去噪等方式实现，增强模型抵抗对抗样本的能力，但需对模型进行重新训练，计算开销较大。恶意检测防御方法检测用户输入的图像是否是恶意图像，从而阻止对抗攻击。预处理与恶意检测方法无须修改预训练模型，但恶意检测的防御方法不能够恢复对抗样本的正确类别。提高模型鲁棒则着重对模型自身入手，不借助附加机制进行预处理和检测。

5.1　预处理

5.1.1　随机化

Wang 等[51]通过在样本中随机消除特征（类似 dropout[52]的方式）来阻止攻击者构建有效的对抗样本，虽然 dropout[52]训练时随机让神经元失效，但测试阶段的防御效果不显著。而随机消除特征在测试阶段能够随机让神经元失效。Prakash 等[53]提出一种像素偏移的方法，包括重新分配像素值和小波去噪两个部分。重新分配像素值利用了卷积神经网络对自然图像中噪声的鲁棒性，随机将一些像素替换为一个小邻域中随机选择的像素。小波去噪修复重新分配像素值和对抗性扰动带来的图像破坏，使图像变得更加自然。然而，Athalye 等[15]利用反向近似梯度技术绕过像素偏移过程的不可微问题，成功绕过防御方法。相似地，Ho 等[54]提出用像素重绘来防御对抗样本。首先训练一个预测模型用于生成预测图像，并对图像像素值的取值范围进行区间划分。原始图像经预测模型生成预测图像，获得预测图像的每个像素值所在区间，再用区间内的随机值替换原始图像的像素值。

5.1.2　图像变换

Dziugaite 等[55]研究发现对于 FGSM 算法[11]产生微小扰动的对抗样本，JPEG 压缩能够减轻对抗样本导致的分类错误，但随着扰动程度的增大，JPEG 压缩的防御效果会降低。Das 等[56]进一步研究发现 JPEG 压缩的一个重要能力就是它能够去

除图像内部的高频信号分量，相当于有选择性地模糊图像，可以消除图像上的对抗性扰动。由此 Das 等[56]提出一种能够快速搭建在已训练好的网络模型上的 JPEG 压缩预处理模块。然而，Guo 等[57]发现总方差最小化[58]和图像缝合[59]比固定性的去噪过程（如 JPEG 压缩[55]、位深度缩减[60]、非局部均值滤波[61]等）具有更强的防御能力。在众多简单防御的基础上，Raff 等[62]提出集合一系列简单的防御（如位深度缩减[60]、JPEG 压缩[55]、小波去噪[63]、均值滤波[64]、非局部均值滤波[61]等）来构建一个强的防御机制来抵抗对抗样本，并且深入考虑了模糊梯度的问题[16]。其基本思想是从大量的随机变换中随机选择几个变换，并在图像输入网络之前按随机顺序应用每个变换。该方法在大规模数据集[8]中也具有鲁棒性。

然而，这些基本的变换大多数已经被证明不能够有效防御对抗样本[53, 55, 60]，而且 He 等[65]曾声称简单防御方法的组合并不能有效防御对抗样本。

5.1.3 去噪网络

Akhtar 等[66]针对通用型扰动攻击算法[12]，提出防御框架。该框架包含扰动校正网络和扰动检验器两部分。该框架将扰动校正网络作为额外的预输入层附加到目标网络中，在不更新目标模型参数的情况下训练它们对被扰动后的图像进行校正，使分类器对对抗样本和其原始图像预测结果相一致。扰动检验器是将扰动修正网络输入输出的特征差异作为输入，通过支持向量机[67]学习得到二元分类器。输入图像首先经过扰动修正网络，然后使用扰动检验器进行检测是否存在扰动。如果检测到扰动，就用经过扰动修正网络修正后的图像代替原始图像作为分类器输入。

传统的去噪自编码器[68]是流行的去噪模型。但传统的去噪自编码器在编码器和解码器之间节点数太少会限制重建高分辨率图像所需的精细尺度信息。为了解决这个问题，Liao 等[69]使用 U-net[70]作为去噪网络模型。U-net[70]网络模型与传统的自编码器有两点区别：第一，去噪网络使用的不是像素层面的损失函数，而是使用特征图作为损失函数；第二，U-net 去噪网络学习的是对抗扰动，而不是构造整个图像。利用去噪网络得到对抗性扰动，并结合原图像得到去噪后的图像。但 Athalye 等[15]指出这个方法不能有效防御白盒攻击。

5.1.4 对抗生成网络

Samangouei 等[71]提出一个基于对抗生成网络[36]的防御框架，主要思想是采用原始数据集训练一个对抗生成网络，利用生成器的表达能力重构一个与原始图像相似但不含对抗扰动的重构图像。整个防御框架如图 5 所示。输入图像经过对抗性生成网络进行重构后，得到一个与原始图像相似的重构图像，再将重构的图像输入目标网络模型进行分类。其中引入了随机种子，使整个网络模型难以攻击。但这个方法在 CIFAR-10 数据集[72]上不能够有效防御白盒攻击。并且 Athalye 等[16]在 MNIST 数据集[73]上采用反向传播近似梯度技术对这种防御机制进行攻击，但成功率只有 48%。

相似地，Bao 等[74]提出基于双向生成对抗网络[75-76]的特征双向对抗生成网络，它描述了高维数据空间和低维语义潜在空间之间的双向映射。输入图像经过特征双向生成网络映射，提取语义特征，这些特征不随扰动而变化，并根据语义特征把输入图像重构成无扰动图像。输入图像经过双向对抗生成网络的重构后输入目标模型进行分类。实验表明，在白盒和灰盒攻击下，这种防御方法对于预先训练好的任意分类器都有效。

5.1.5 超分辨率

Mustafa 等[77]提出一种基于超分辨率[78]和小波去噪[79]的防御方法。其基本思想是使用超分辨

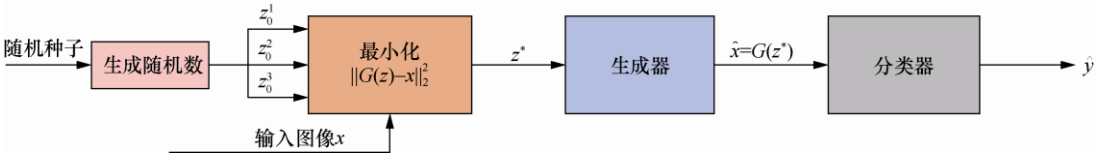


图 5 对抗生成网络防御框架

Figure 5 Defense framework of generative adversarial network

率网络将流形外的对抗样本引入自然图像流形中，从而将分类恢复到正确的类别。算法采用小波去噪减少恶意对抗扰动，然后使用超分辨率网络加强图像视觉质量。该方法在无须重新训练模型的前提下，可以补充到现有的防御机制，同时提升分类准确率。

5.2　提高模型鲁棒性

5.2.1　对抗性训练

对抗性训练作为目前能够有效提高模型鲁棒性的防御方式，其缺点是训练模型的开销太大且无法对所有攻击类型的对抗样本进行对抗训练。

Goodfellow 等首次[11]提出对抗训练来提高模型鲁棒性，Kurakin 等[18]利用批量归一化[80]方法，成功将其扩展到 Inception-v3 模型[31]和 ImageNet 数据集[8]。但其缺点是只能防御单步的攻击[11]，不能防御迭代的攻击[81]。而 Chang 等[82]提出了一种基于双对抗样本的训练方式，既能够抵抗单步的对抗样本，也可以防御迭代的对抗样本。许多对抗训练只能防御特定的对抗攻击[11, 83-84]，在 Athalye 等[16]的评测中，Madry 等[81]利用梯度投影下降攻击算法进行对抗训练的防御方法是唯一没有被完全攻破的防御方法。但 Madry 等[81]只在 MNIST[73]和 CIFAR-10[72]数据集上进行对抗训练。随后，Kannan 等[85]成功将其扩展到 ImageNet 数据集[8]，将相似的样本构成一对，配对样本的模型输出相似程度作为损失函数的一部分。该方法在 ImageNet 数据集[8]上具有鲁棒性，而且超过了当时表现最好的集成对抗训练方法[86]。

此外，Li 等[87]利用距离度量学习中常见的三重损失函数来构建对抗训练模型。三重损失函数可以优化嵌入空间，即具有相同标签的样本在空间中尽可能地接近，不同标签的样本尽可能地远离。由于很难找到一个能够代表对抗样本域的样本，因此对抗训练泛化能力较弱。Song 等[88]提出了一种具有领域适应能力的对抗训练方法，旨在学习样本有意义的表示，这种表示在对抗样本和原始数据集上是不变的。

5.2.2　批量调整网络梯度

Rozsa 等[89]认为在一批训练样本中，正确分类的样本往往对模型权重更新贡献较小，在正常样本周围难以形成更加平坦、不变的区域，这导致很小的扰动就能够使分类器分类错误。因此，他们提出了一种简单、高效、可以提高模型鲁棒性的训练方法（批量调整网络梯度）对梯度进行调整，用于提高分类正确样本对模型参数更新的贡献值。该方法的优点在于不依赖任何形式的数据扩充和使用对抗样本进行对抗性训练，同时能够保持甚至增强整体分类性能并抵抗对抗攻击。

5.2.3　正则化

Tomar 等[90]提出了一种深度学习和流形学习相结合的方法（流形正则化网络）。在模型损失上附加一个盲点特性的流形损失项。结果表明，该方法能够抵抗对抗样本，并且能够使模型在流形上泛化。此外在 MNIST[73]、CIFAR-10[72]、SVHN[91]数据集上结合 dropout[52]能够获得较好的表现。而缺点在于梯度正则化是二阶的，这导致模型的训练时间翻了一倍。Sankaranarayanan 等[92]观察到隐藏层激活的对抗性扰动在不同的样本中普遍存在，提出通过中间激活函数层的对抗扰动来提供更强的正则化。该方法不仅比 dropout[52]有更强的正则化，而且比传统对抗性训练方法的鲁棒性更强。

神经网络模型不能够学习到关键特征，图像轻微的变化就导致分类器分类错误。针对这个问题，Liu 等[93]提出通过基于非线性注意力模块和 L_2 特征正则化的特征优先模型，使模型分类依赖于关键特征。其中，注意力模块通过给关键特征分配更大的权重，促使模型学习到关键特征。对模型进行 L_2 正则化促使提取原始图像和对抗样本相似的本质特征，有效忽略了添加的扰动。

5.2.4　防御蒸馏

Papernot 等[94]基于网络蒸馏[95]思想提出一种防御方法（防御蒸馏）。网络蒸馏技术原本是为了把网络模型部署到移动端而设计的一种网络压缩技术，它能够把大的网络压缩成小的网络，而且能够保持网络的性能。在蒸馏的过程中通过调整温度（softmax 函数中的参数 T），产生更加平滑、对扰动更加不敏感的模型，从而提升模型对对抗性性样本的鲁棒性。文献[94]声称防御蒸馏能够抵抗 90%的对抗样本。随后，Carlini 等[96]指出蒸馏并不能防御对抗攻击，只要对 C&W[20]进行微调便能攻击成功。

5.2.5 特征去噪

微小的对抗性扰动在网络中被逐层放大，导致网络的特征图中出现大量的噪声。原始图像的特征主要集中在图像中的语义特征，而对抗样本的特征在语义不相关的区域被激活。因此，Xie 等[97]开发了新的卷积网络架构，这个架构包含了用于特征图去噪的去噪模块。虽然去噪模块并不能提高在原始数据集下的分类准确率，但去噪模块和对抗性训练结合在一起，在白盒攻击和黑盒攻击中能够显著提高模型的鲁棒性。在 2018 年 CAAD（对抗样本攻防赛）中，该方法在 48 种未知攻击下获得 50.6%的分类准确率。

5.2.6 随机消除特征

由于深度学习模型不能够很好的学习到关键特征，攻击者通过向不相关的非关键特征维度添加微小的扰动，就可以导致模型分类错误。因此，Gao 等[98]提出了一种 DeepCloak 机制，具体过程如图 6 所示。利用掩码的方式删除网络模型中不必要的特征，限制了攻击者生成对抗样本的能力，从而提升模型的鲁棒性。与其他的防御机制相比，DeepCloak 机制更易于实现且计算效率较高。

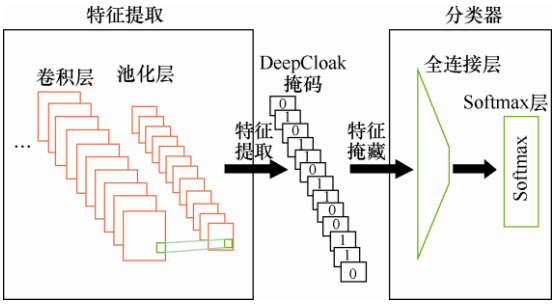


图 6 随机消除特征框架

Figure 6 A framework of random feature nullification

5.2.7 卷积稀疏编码

Sun 等[99]基于卷积稀疏编码[100]，构造了一个分层的低维准自然图像空间。文献[99]通过将原始图像和对抗样本映射到一个低维准自然图像空间来实现高水平的鲁棒性。这个准自然图像空间近似于自然图像空间。同时该方法消除了对抗性扰动，使对抗样本接近其在特征空间中的原始输入。在训练阶段，作者在输入图像和神经网络的第一层之间引入一个稀疏变换层来有效地将图像映射到准自然图像空间中，用映射到准自然空间的图像对分类模型进行训练。在测试阶段，把原始输入映射到准自然空间的图像作为分类模型的输入。与其他不可知攻击的对抗性防御方法相比，该方法在对抗性扰动的大小、各种不同的图像分辨率和数据集规模方面，都有具有更强的鲁棒性。

5.2.8 深度收缩网络

研究发现去噪自编码器[101]可以消除大量的对抗性噪声，但当去噪自编码器与原始神经网络叠加时，会再次受到失真程度更小的对抗样本攻击。Rifai 等[102]认为这是自动编码机训练上的问题，并提出了深度收缩网络模型。模型采用了一种全新的端到端的训练过程，该过程采用伸缩自编码器相似的平滑损失、逐层惩罚，近似地使网络输出的方差相对于输入的扰动最小，使模型在训练数据点周围更加平滑。这增强了模型对对抗样本的鲁棒性，并且没有明显的性能损失。

5.2.9 阻止可移植性

对抗样本具有可迁移性的特点，由特定网络模型生成的对抗样本可能误导其他网络模型结构不同或者不同训练集上的分类器。针对这个问题，Hosseini 等[103]提出了空标签的方法来防御黑盒下的对抗样本可移植性攻击。其主要的思想是通过对数据集的标签增加一个空类别，并进行对抗性训练。该方法的优势在于能够将对抗样本的错误标签分类到空类别中，而不是其他的错误类别，有效阻止对抗样本的可移植性，同时图像能够保持模型的精度。

5.3 恶意检测

5.3.1 图像变换

Tian 等[104]认为对抗性的样本通常对某些图像变换操作敏感，如旋转和移位，但原始图像通常是免疫这种操作的，因此，提出一种基于图像变换的对抗样本检测方法。首先，对一幅图像进行一组的变换操作，生成多幅变换后的图像。然后，使用这些变换后的图像的分类结果作为特征训练一个神经网络来预测原始图像是否受到了攻击者的干扰。为了防御更复杂的白盒攻击，在转换过程中引入随机性。对多个图像数据集的实验结果表明，C&W[20]算法产生的对抗样本的检测率达到 99%。其中，对于白盒攻击，该方法的检测率达到 70%以上。

5.3.2　有状态检测

针对白盒对抗样本的防御已被证明是难以实现的。白盒攻击在实际场景中不太现实，云平台提供的服务一般是基于询问的方式。因此，Chen 等[105]首次提出基于有状态检测的黑盒防御方法，相比目前研究的无状态防御，该方法增强了防御方的能力。有状态检测方法思想是记录一定量的用户询问记录，并且在用户下次询问时，将用户询问与以往的记录进行对比，如果相似程度在规定的一定阈值内则认为是恶意攻击，具体流程如图 7 所示。首先，为了能够压缩用户询问记录的存储，使用了相似编码进行压缩编码。然后，将用户询问输入和以往记录比较，利用 k 近邻算法计算距离 d，如果 d 小于阈值 δ 则认为用户是在进行恶意攻击。利用黑盒攻击 NES[106]和边界攻击[26]进行评估，实验分析表明，基于询问的黑盒攻击通常需要几十万到上百万的询问，这非常容易触发用户的防御机制。即使不触发防御机制，攻击所需的存储服务也需消耗大量资源。有状态的检测方法缺点是无法防御不需要询问的迁移攻击。但是，该方法能够与防御迁移攻击的集成对抗性训练方法进行结合，可以弥补有状态检测的不足，使其在黑盒攻击情况下有较好的表现。

5.3.3　隐藏层特征

Metzen 等[107]提出增加一个小的检测子网络来扩充深度神经网络，这个子网络是用于检测对抗样本的二分类器。文献[101]将子网络嵌入网络内部，结合对抗性训练，成功检测由 FGSM[11]、BIM[18]、Deepfool[21]攻击算法产生的对抗样本。Li 等[108]则通过分析对抗样本与正常的样本是否来自相同分布的方式来进行对抗样本的检测。该检测方法使用中间层的特征，即基于卷积层的输出，而不是直接使用原始图像和对抗样本进行统计分布。文献[108]用这些中间特征训练了一个级联（层叠）分类器，有效地检测对抗样本。此外，从一个特定的对抗算法[10]生成数据并进行训练，得到的分类器可以成功地检测从完全不同机制[109]生成的对抗样本。Feinman 等[110]认为对抗样本与原始图像属于不同的分布，因此使用高斯混合模型对神经网络最后一层的输出进行建模，用来检测对抗样本。由于神经网络尾部的隐藏层可以捕获到输入的高级语义信息，在最后一层上使用一个简单的分类器将比其应用于原始输入图像更加准确可靠。相似地，Zheng 等[111]通过分析深度神经网络中隐藏神经元的输出分布，使用高斯混合模型来近似深度神经网络分类器的隐藏状态分布，然后通过判断输入样本状态分布是否异常来检测对抗样本。该方法能够应用于任何深度神经网络结构上，并可以与其他防御策略相结合，以此提高模型的鲁棒性。实验表明，该方法能够防御黑盒和灰盒攻击。

Carlini 等[14]通过构造新的损失函数，成功绕过了 10 种检测方法，其中包括上面所提到的前 3 种方法[107-108, 110]。

5.3.4　流形学习

Meng 等[112]提出一种防御对抗样本的方法，由若干个检测器和一个修正器组成。检测器在训

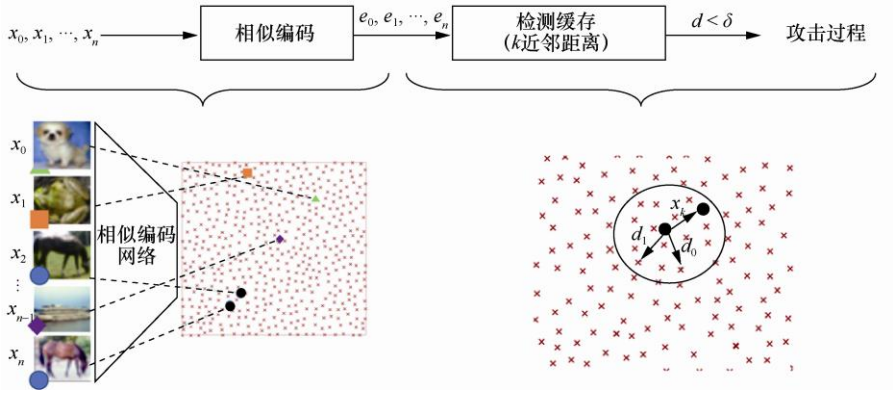


图 7　有状态检测流程

Figure 7　The pipeline of stateful detection

练期间学习对抗样本的流行分布，对于远离流形边界的样本判定为对抗样本。而修正器在流形区域中寻找一个与输入 x 相近的 x' 替换 x 并输入分类器进行分类。然而，Carlini 等[113]指出这个方法只能抵抗轻微对抗扰动。Huang 等[114]观察到对抗样本子空间在方向上与原始数据子流形非常接近，因此通过添加随机扰动来扩展对抗子空间能够使对抗样本被重新分类正确。而且由于神经网络模型的鲁棒性，添加微弱的随机扰动并不会影响分类结果。Huang 等[114]提出了一种模型无关方法来解决检测对抗样本。该方法通过分析模型对于随机扰动输入的反应，以相对变化的置信度作为检测对抗样本的鉴别标准，在理论框架中学习范数有界对抗性扰动的鲁棒性，可以轻易地部署到现成的深度学习模型中。

5.3.5 自适应去噪

传统的去噪对扰动较大的情况效果显著，但在扰动较小的情况，去噪会使图像模糊，导致分类准确率低。针对这个问题，Liang 等[115]利用交叉熵的大小作为标准实现自适应去噪。其中包含两种图像降噪方式：标量量化和空间平滑滤波。先使用交叉熵自适应量化区间大小，然后判断是否需要进行空间平滑滤波。降噪后的图像和原始输入分别用分类器进行分类，如果分类结果一致，则认为原始输入是正常样本，否则认为是对抗样本。

5.4 防御方法总结

各类防御方法都有其相应的优点和缺点，本文将具有代表性的防御方法加以比较，如表 2 所示。预处理和恶意检测类的方法着重对输入数据进行去噪，然而这些方法不能够有效过滤对抗扰动，存在一些问题。例如，只能针对特定数据集或者算法有效，不能够很好泛化，但其优点是开销较小，往往无须对模型进行修改和重新训练原有模型，容易补充到已有的防御系统。提高模型鲁棒性的方法针对模型自身进行改进，往往需要对模型架构和训练方式进行修改，并进行重新训练。目前公认的有效防御对抗样本的方法是对抗训练，其主要的缺点是往往需要改变原有模型的架构、训练策略进行对抗训练，这就导致模型训

表 2 防御方法总结

Table 2 A summary of adversarial defenses

类别	方法	优点	缺点
预处理	文献[54]	可以预防未知攻击，防止模型过拟合	对扰动大小敏感
	文献[62]	不依赖混淆梯度[16]，无须重新训练模型，可以防御白盒攻击	需大量的简单防御作为基础
	文献[69]	具有较强的泛化能力	不能防御白盒攻击，去噪能力依赖于模型的表达能力
	文献[71]	无须重新训练模型和改变模型架构	对抗生成网络的重构能力受限，在数据集 CIFAR-10[72]和 ImageNet[8]上无效
	文献[77]	简单，容易补充到已有的防御系统中，可以防御未知攻击	依赖超分辨率网络的表达能力
提高模型鲁棒性	文献[85]	能够在大数据集 ImageNet[8]上进行对抗训练	需要重新进行模型架构修改，并进行对抗训练，计算开销较大
	文献[89]	简单，不依赖对抗样本和数据增强	不同数据集不能够自适应，需要调整参数
	文献[93]	引入注意力机提促使模型学习鲁棒性特征	需要重新进行模型架构修改，并进行对抗训练，计算开销较大
	文献[97]	特征去噪模块化，添加去噪模块简单	修改现有模型，并进行对抗训练，开销较大
	文献[99]	防御未知攻击、不受数据集规模影响、无须修改原有模型架构	需要对原有数据集进行映射，并进行对抗训练
恶意检测	文献[104]	操作简单、开销较低、模型无关	图像旋转角度不能够自适应
	文献[105]	阻止对手生成对抗样本	不能够防御迁移攻击，需要存储用户查询记录
	文献[111]	简单、可集成到现有防御系统中	依赖模型的内部信息
	文献[115]	根据扰动大小自适应去噪、模型无关	不适用只改变小部分像素值的攻击算法

练开销较大。

6　未来展望

6.1　对抗样本

对抗样本最早只针对图像分类任务[10]提出，然而根据本文第 4 节中的描述，针对图像分类[30]、网络空间攻击[32,34]、停车牌识别[41]、目标检测[35,39]、语义分割[46-47]、人脸识别[49]等任务的对抗样本被发现。尽管已经有针对不同应用场景的对抗样本，但与现实世界中的应用场景还有较大差距。未来针对现实物理环境具有鲁棒性的对抗样本将是研究的热点。

6.2　对抗攻击

对抗攻击从梯度攻击[11, 18, 81]发展到置信度攻击[24]，再到决策攻击[106, 116]的攻击，其攻击所需的信息量逐渐减少，而且已经成功应用于现实场景。尽管近年来对抗样本的生成算法取得了较大的进步，但仍然有许多不足和限制，如询问的次数过多、攻击不够稳定等。未来针对更高效、更微小扰动、稳定的黑盒攻击算法仍是研究的热点。

6.3　对抗性防御

正如本文中所提到的大部分防御方法只针对特定的已知攻击算法有效，在未知攻击算法中并不能够很好地泛化。迄今为止，仍然没有一个对抗样本的防御方法能够完全抵抗白盒攻击。因此，研究人员开始着重于针对灰盒、黑盒攻击[77]以及对抗样本检测机制[105]方向的研究。相比于强大的白盒攻击，灰盒和黑盒攻击者所拥有的信息更少，相对更容易防御，而且在实现环境中更加符合实际情况。如何训练一个鲁棒性模型以及有恶意检测机制的模型是未来的研究热点。

7　结束语

本文针对深度学习领域存在的对抗攻击问题，首先给出了对抗样本和对抗攻击的基本概念，构建了威胁模型；介绍了近年来深度学习领域具有代表性的对抗攻击手段及现实中的对抗攻击实例；研究了典型的防御方法，并根据最新研究进展分析了方法的有效性；针对深度学习的对抗攻击能够直接影响到系统的可用性，诱导系统输出错误的结果。目前该领域内仍然没有一个能够有效防御对抗样本的防御机制，但训练一个鲁棒性模型是可能的。最后，根据对研究现状的分析，本文讨论了这一领域未来研究方向。

参考文献：

[1] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[C]//Advances in Neural Information Processing Systems. 2012: 1097-1105.

[2] DAHL G E, YU D, DENG L, et al. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 20: 30-42.

[3] HINTON G, DENG L, YU D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups[J]. IEEE Signal Processing Magazine, 2012, 29: 82-97.

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

[5] DAHL G E, STOKES J W, DENG L, et al. Large-scale malware classification using random projections and neural networks[C]//2013 IEEE International Conference on Acoustics, Speech and Signal Processing. 2013: 3422-3426.

[6] SILVER D, HUANG A, MADDISON C J, et al. Mastering the game of go with deep neural networks and tree search[J]. Nature, 2016, 529: 484.

[7] VINYALS O, BABUSCHKIN I, CZARNECKI W M, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning[J]. Nature, 2019, 575: 350-354.

[8] DENG J, DONG W, SOCHER R, et al. Imagenet: a large-scale hierarchical image database[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. 2009: 248-255.

[9] LECUN Y, BOSER B, DENKER J S, et al. Backpropagation applied to handwritten zip code recognition[J]. Neural Computation, 1989, 1: 541-551.

[10] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[J]. arXiv preprint arXiv: 1312. 6199, 2013.

[11] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv preprint arXiv: 1412. 6572, 2014.

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

[13] SU J, VARGAS D V, SAKURAI K. One pixel attack for fooling deep neural networks[J]. IEEE Transactions on Evolutionary Computation, 2019, 23: 828-841.

[14] CARLINI N, WAGNER D. Adversarial examples are not easily detected: Bypassing ten detection methods[C]//Proceedings of the

10th ACM Workshop on Artificial Intelligence and Security. 2017: 3-14.

[15] ATHALYE A, CARLINI N. On the robustness of the CVPR 2018 white-box adversarial example defenses[J]. arXiv preprint arXiv: 1804. 03286, 2018.

[16] ATHALYE A, CARLINI N, WAGNER D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples[J]. arXiv preprint arXiv: 1802. 00420, 2018.

[17] FLETCHER R. Practical methods of optimization[M]. John Wiley & Sons, 2013.

[18] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial machine learning at scale[J]. arXiv preprint arXiv: 1611. 01236, 2016.

[19] 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 (EuroS&P). 2016: 372-387.

[20] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//2017 IEEE Symposium on Security and Privacy (SP). 2017: 39-57.

[21] MOOSAVI-DEZFOOLI S-M, FAWZI A, FROSSARD P. Deepfool: a simple and accurate method to fool deep neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2574-2582.

[22] ATHALYE A, ENGSTROM L, ILYAS A, et al. Synthesizing robust adversarial examples[J]. arXiv preprint arXiv: 1707. 07397, 2017.

[23] MCLAREN K. XIII—The development of the CIE 1976 (L* a* b*) uniform colour space and colour-difference formula[J]. Journal of the Society of Dyers and Colourists, 1976, 92: 338-341.

[24] CHEN P-Y, ZHANG H, SHARMA Y, et al. Zoo: zeroth order optimization based black-box attacks to deep neural networks without training substitute models[C]//Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. 2017: 15-26.

[25] KINGMA D P, BA J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv: 1412. 6980, 2014.

[26] BRENDEL W, RAUBER J, BETHGE M. Decision-based adversarial attacks: reliable attacks against black-box machine learning models[J]. arXiv preprint arXiv: 1712. 04248, 2017.

[27] BRUNNER T, DIEHL F, LE M T, et al. Guessing smart: biased sampling for efficient black-box adversarial attacks[C]//Proceedings of the IEEE International Conference on Computer Vision.2019: 4958-4966.

[28] PERLIN K. An image synthesizer[J]. ACM Siggraph Computer Graphics, 1985, 19: 287-296.

[29] CHEN J, JORDAN M I, WAINWRIGHT M J. HopSkipJumpAttack: a query-efficient decision-based attack[J]. arXiv preprint arXiv: 1904. 02144, 3: 2019.

[30] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial examples in the physical world[J]. arXiv preprint arXiv: 1607. 02533, 2016.

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

[32] PAPERNOT N, MCDANIEL P, GOODFELLOW I, et al. Practical black-box attacks against machine learning[C]//Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security. 2017: 506-519.

[33] LIU Y, CHEN X, LIU C, et al. Delving into transferable adversarial examples and black-box attacks[J]. arXiv preprint arXiv: 1611. 02770, 2016.

[34] LI X, JI S, HAN M, et al. Adversarial examples versus cloud-based detectors: a black-box empirical study[J]. IEEE Transactions on Dependable and Secure Computing, 2019.

[35] WEI X, LIANG S, CHEN N, et al. Transferable adversarial attacks for image and video object detection[J]. arXiv preprint arXiv: 1811. 12641, 2018.

[36] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Advances in Neural Information Processing Systems. 2014: 2672-2680.

[37] REN S, HE K, GIRSHICK R, et al. Faster r-CNN: Towards real-time object detection with region proposal networks[C]// Advances in Neural Information Processing Systems. 2015: 91-99.

[38] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector[C]//European Conference on Computer Vision. 2016: 21-37.

[39] THYS S, VAN RANST W, GOEDEMÉ T. Fooling automated surveillance cameras: adversarial patches to attack person detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2019.

[40] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 7263-7271.

[41] EVTIMOV I, EYKHOLT K, FERNANDES E, et al. Robust physical-world attacks on deep learning models[J]. arXiv preprint arXiv: 1707. 08945, 2017.

[42] LU J, SIBAI H, FABRY E, Et al. Standard detectors aren't (currently) fooled by physical adversarial stop signs[J]. arXiv preprint arXiv: 1710. 03337, 2017.

[43] EYKHOLT K, EVTIMOV I, FERNANDES E, et al. Note on attacking object detectors with adversarial stickers[J]. arXiv preprint arXiv: 1712. 08062, 2017.

[44] CHEN S T, CORNELIUS C, MARTIN J, et al. Shapeshifter: robust physical adversarial attack on faster r-CNN object detector[C]//Joint European Conference on Machine Learning and Knowledge Discovery in Databases. 2018: 52-68.

[45] BROWN T B, MANÉ D, ROY A, et al. Adversarial Patch[J]. CoRR, abs/1712. 09665: 2017.

[46] HENDRIK METZEN J, CHAITHANYA KUMAR M, BROX T, et al. Universal adversarial perturbations against semantic image segmentation[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 2755-2764.

[47] XIE C, WANG J, ZHANG Z, et al. Adversarial examples for se-

mantic segmentation and object detection[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 1369-1378.

[48] CISSE M, ADI Y, NEVEROVA N, et al. Houdini: Fooling deep structured prediction models[J]. arXiv preprint arXiv: 1707. 05373, 2017.

[49] SHARIF M, BHAGAVATULA S, BAUER L, et al. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition[C]//Proceedings of the 2016 ACM Sigsac Conference on Computer and Communications Security. 2016: 1528-1540.

[50] ZHOU Z, TANG D, WANG X, et al. Invisible mask: Practical attacks on face recognition with infrared[J]. arXiv preprint arXiv: 1803. 04683, 2018.

[51] WANG Q, GUO W, ZHANG K, et al. Adversary resistant deep neural networks with an application to malware detection[C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2017: 1145-1153.

[52] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. The Journal of Machine Learning Research. 2014, 15: 1929-1958.

[53] PRAKASH A, MORAN N, GARBER S, et al. Deflecting adversarial attacks with pixel deflection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 8571-8580.

[54] HO J, KANG D-K. Pixel Redrawn For A Robust Adversarial Defense[EB].

[55] DZIUGAITE G K, GHAHRAMANI Z, ROY D M. A study of the effect of jpg compression on adversarial images[J]. arXiv preprint arXiv: 1608. 00853, 2016.

[56] DAS N, SHANBHOGUE M, CHEN S T, et al. Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression[J]. arXiv preprint arXiv: 1705. 02900, 2017.

[57] GUO C, RANA M, CISSE M, et al. Countering adversarial images using input transformations[J]. arXiv preprint arXiv: 1711. 00117, 2017.

[58] RUDIN L I, OSHER S, FATEMI E. Nonlinear total variation based noise removal algorithms[J]. Physica D: Nonlinear Phenomena, 1992, 60: 259-268.

[59] EFROS A A, FREEMAN W T. Image quilting for texture synthesis and transfer[C]//Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques. 2001: 341-346.

[60] XU W, EVANS D, QI Y. Feature squeezing: Detecting adversarial examples in deep neural networks[J]. arXiv preprint arXiv: 1704. 01155, 2017.

[61] BUADES A, COLL B, MOREL J M. A non-local algorithm for image denoising[C]//2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). 2005: 60-65.

[62] RAFF E, SYLVESTER J, FORSYTH S, et al. Barrage of random transforms for adversarially robust defense[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 6528-6537.

[63] ANTONINI M, BARLAUD M, MATHIEU P, et al. Image coding using wavelet transform[J]. IEEE Transactions on Image Processing. 1992, 1: 205-220.

[64] HUANG T, YANG G, TANG G. A fast two-dimensional median filtering algorithm[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing. 1979, 27: 13-18.

[65] HE W, WEI J, CHEN X, et al. Adversarial example defense: Ensembles of weak defenses are not strong[C]//11th $$USENIX$$ Workshop on Offensive Technologies ($$WOOT$$ 17), 2017.

[66] AKHTAR N, LIU J, MIAN A. Defense against universal adversarial perturbations[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 3389-3398.

[67] CORTES C, VAPNIK V. Support-vector networks[J]. Machine Learning, 1995, 20: 273-297.

[68] VINCENT P, LAROCHELLE H, BENGIO Y, et al. Extracting and composing robust features with denoising autoencoders[C]// Proceedings of the 25th International Conference on Machine Learning. 2008: 1096-1103.

[69] LIAO F, LIANG M, DONG Y, et al. Defense against adversarial attacks using high-level representation guided denoiser[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 1778-1787.

[70] RONNEBERGER O, FISCHER P, BROX T. U-net: convolutional networks for biomedical image segmentation[C]//International Conference on Medical Image Computing and Computer-assisted Intervention. 2015: 234-241.

[71] SAMANGOUEI P, KABKAB M, CHELLAPPA R. Defense-gan: protecting classifiers against adversarial attacks using generative models[J]. arXiv preprint arXiv: 1805. 06605, 2018.

[72] KRIZHEVSKY A, NAIR V, HINTON G. CIFAR-10 (Canadian Institute for Advanced Research)[D]. Toronto: University of Toronto, 2009.

[73] YANN L, BERNHARD B, JOHN S D, et al. Backpropagation applied to handwritten zip code recognition[J]. Neural Computation, 1989, 1(4): 541-551.

[74] BAO R, LIANG S, WANG Q. Featurized Bidirectional GAN: adversarial defense via adversarially learned semantic inference[J]. arXiv preprint arXiv: 1805. 07862, 2018.

[75] DUMOULIN V, BELGHAZI I, POOLE B, et al. Adversarially learned inference[J]. arXiv preprint arXiv: 1606. 00704, 2016.

[76] DONAHUE J, KRÄ HENBÜ HL P, DARRELL T. Adversarial feature learning[J]. arXiv preprint arXiv: 1605. 09782, 2016.

[77] MUSTAFA A, KHAN S H, HAYAT M, et al. Image super-resolution as a defense against adversarial attacks[J]. IEEE Transactions on Image Processing. 2019, 29: 1711-1724.

[78] LIM B, SON S, KIM H, et al. Enhanced deep residual networks for single image super-resolution[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2017: 136-144.

[79] CHANG S G, YU B, VETTERLI M. Adaptive wavelet thresholding for image denoising and compression[J]. IEEE Transactions on Image Processing, 2000, 9: 1532-1546.

[80] IOFFE S, SZEGEDY C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[J]. arXiv preprint arXiv: 1502. 03167, 2015.

[81] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[J]. arXiv preprint arXiv: 1706. 06083, 2017.

[82] CHANG T J, HE Y, LI P. Efficient Two-Step Adversarial Defense for Deep Neural Networks[J]. arXiv preprint arXiv: 1810. 03739, 2018.

[83] HUANG R, XU B, SCHUURMANS D, et al. Learning with a strong adversary[J]. CoRR, abs/1511. 03034: 2015.

[84] SHAHAM U, YAMADA Y, NEGAHBAN S. Understanding adversarial training: Increasing local stability of supervised models through robust optimization[J]. Neurocomputing, 2018, 307: 195-204,.

[85] KANNAN H, KURAKIN A, GOODFELLOW I. Adversarial logit pairing[J]. arXiv preprint arXiv: 1803. 06373, 2018.

[86] TRAMÈR F, KURAKIN A, PAPERNOT N, et al. Ensemble adversarial training: Attacks and defenses[J]. arXiv preprint arXiv: 1705. 07204, 2017.

[87] LI P, YI J, ZHOU B, et al. Improving the Robustness of Deep Neural Networks via Adversarial Training with Triplet Loss[J]. arXiv preprint arXiv: 1905. 11713, 2019.

[88] SONG C, HE K, WANG L, et al. Improving the generalization of adversarial training with domain adaptation[J]. arXiv preprint arXiv: 1810. 00740, 2018.

[89] ROZSA A, GUNTHER M, Boult T E. Towards robust deep neural networks with BANG[J]. arXiv preprint arXiv: 1612. 00138, 2016.

[90] TOMAR V S, ROSE R C. Manifold regularized deep neural networks[C]//Fifteenth Annual Conference of the International Speech Communication Association. 2014.

[91] NETZER Y, WANG T, COATES A, et al. Reading digits in natural images with unsupervised feature learning[R]. 2011.

[92] SANKARANARAYANAN S, JAIN A, CHELLAPPA R, et al. Regularizing deep networks using efficient layerwise adversarial training[C]//Thirty-Second AAAI Conference on Artificial Intelligence. 2018:

[93] LIU C, JAJA J. Feature prioritization and regularization improve standard accuracy and adversarial robustness[J]. arXiv preprint arXiv: 1810. 02424, 2018.

[94] 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 (SP). 2016: 582-597.

[95] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J]. arXiv preprint arXiv: 1503. 02531, 2015.

[96] CARLINI N, WAGNER D. Defensive distillation is not robust to adversarial examples[J]. arXiv preprint arXiv: 1607. 04311, 2016.

[97] XIE C, WU Y, MAATEN L V D, et al. Feature denoising for improving adversarial robustness[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 501-509.

[98] GAO J, WANG B, LIN Z, et al. Masking deep neural network models for robustness against adversarial samples[J]. arXiv preprint arXiv: 1702. 06763,

[99] SUN B, TSAI N-H, LIU F, et al. Adversarial Defense by Stratified Convolutional Sparse Coding[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 11447-11456.

[100] CHOUDHURY B, SWANSON R, HEIDE F, et al. Consensus convolutional sparse coding[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 4280-4288.

[101] GU S, RIGAZIO L. Towards deep neural network architectures robust to adversarial examples[J]. arXiv preprint arXiv: 1412. 5068, 2014.

[102] RIFAI S, VINCENT P, MULLER X, et al. Contractive auto-encoders: explicit invariance during feature extraction[C]// Proceedings of the 28th International Conference on International Conference on Machine Learning. 2011: 833-840.

[103] HOSSEINI H, CHEN Y, KANNAN S, et al. Blocking transferability of adversarial examples in black-box learning systems[J]. arXiv preprint arXiv: 1703. 04318, 2017.

[104] TIAN S, YANG G, CAI Y. Detecting adversarial examples through image transformation[C]//Thirty-Second AAAI Conference on Artificial Intelligence, 2018.

[105] CHEN S, CARLINI N, WAGNER D. Stateful detection of black-box adversarial attacks[J]. arXiv preprint arXiv: 1907. 05587, 2019.

[106] ILYAS A, ENGSTROM L, ATHALYE A, et al. Black-box adversarial attacks with limited queries and information[J]. arXiv preprint arXiv: 1804. 08598, 2018.

[107] METZEN J H, GENEWEIN T, FISCHER V, et al. On detecting adversarial perturbations[J]. arXiv preprint arXiv: 1702. 04267, 2017.

[108] LI X, LI F. Adversarial examples detection in deep networks with convolutional filter statistics[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 5764-5772.

[109] NGUYEN A, YOSINSKI J, CLUNE J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 427-436.

[110] FEINMAN R, CURTIN R R, SHINTRE S, et al. Detecting adversarial samples from artifacts[J]. arXiv preprint arXiv: 1703. 00410, 2017.

[111] ZHENG Z, HONG P. Robust detection of adversarial attacks by modeling the intrinsic properties of deep neural networks[C]//Advances in Neural Information Processing Systems, 2018: 7913-7922.

[112] MENG D, CHEN H. Magnet: a two-pronged defense against adversarial examples[C]//Proceedings of the 2017 ACM SIGSAC

Conference on Computer and Communications Security. 2017: 135-147.

[113] CARLINI N, WAGNER D. Magnet and" efficient defenses against adversarial attacks" are not robust to adversarial examples[J]. arXiv preprint arXiv: 1711. 08478, 2017.

[114] HUANG B, WANG Y, WANG W. Model-agnostic adversarial detection by random perturbations[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence, 2019: 4689-4696.

[115] LIANG B, LI H, SU M, et al. Detecting adversarial image examples in deep networks with adaptive noise reduction[J]. arXiv preprint arXiv: 1705. 08378, 2017.

[116] CHENG M, LE T, CHEN P Y, et al. Query-efficient hard-label black-box attack: An optimization-based approach[J]. arXiv preprint arXiv: 1807. 04457, 2018.

[作者简介]

刘西蒙（1988− ），男，陕西西安人，博士，福州大学研究员、博士生导师，主要研究方向为隐私计算、密文数据挖掘、大数据隐私保护、可搜索加密。

谢乐辉（1997− ），男，福建建瓯人，福州大学硕士生，主要研究方向为人工智能安全。

王耀鹏（1995− ），男，福建泉州人，福州大学硕士生，主要研究方向为人工智能安全。

李旭如（1995− ），女，安徽宣城人，华东师范大学博士生，主要研究方向为无线通信网络、网络空间安全。