Contents lists available at ScienceDirect

Engineering

journal homepage: www.elsevier.com/locate/eng

Research
Artificial Intelligence—Article

深度学习中的对抗性攻击和防御

任奎 [a,b,*], Tianhang Zheng [c], 秦湛 [a,b], Xue Liu [d]

[a] Institute of Cyberspace Research, Zhejiang University, Hangzhou 310027, China
[b] College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China
[c] Department of Electrical and Computer Engineering, University of Toronto, Toronto M5S 2E8, Canada
[d] School of Computer Science, McGill University, Montreal H3A 0E9, Canada

ARTICLE INFO

Article history:
Received 3 May 2019
Revised 6 September 2019
Accepted 26 December 2019
Available online 3 January 2020

关键词
机器学习
深度神经网络
对抗实例
对抗攻击
对抗防御

摘要

在深度学习（deep learning, DL）算法驱动的数据计算时代，确保算法的安全性和鲁棒性至关重要。最近，研究者发现深度学习算法无法有效地处理对抗样本。这些伪造的样本对人类的判断没有太大影响，但会使深度学习模型输出意想不到的结果。最近，在物理世界中成功实施的一系列对抗性攻击证明了此问题是所有基于深度学习系统的安全隐患。因此有关对抗性攻击和防御技术的研究引起了机器学习和安全领域研究者越来越多的关注。本文将介绍深度学习对抗攻击技术的理论基础、算法和应用。然后，讨论了防御方法中的一些代表性研究成果。这些攻击和防御机制可以为该领域的前沿研究提供参考。此外，文章进一步提出了一些开放性的技术挑战，并希望读者能够从所提出的评述和讨论中受益。

1. 引言

计算能力的万亿倍增长使得深度学习（deep learning, DL）在处理各种机器学习（machine learning, ML）任务中得到广泛应用，如图像分类[1]、自然语言处理[2]和博弈论[3]。然而研究者发现现有DL算法存在着严重的安全隐患：攻击者可以通过给良性样本添加特定噪声而轻易地欺骗DL模型，并且通常不会被人发现[4]。攻击者利用人的视觉/听觉无法感知的扰动，足以使正常训练的模型输出置信度很高的错误预测，研究者将这种现象叫做对抗攻击，它被认为是在生产中部署DL模型之前的巨大障碍，因此激发了人们对对抗攻击和防御研究的广泛兴趣。

根据威胁模型可以将现有的对抗性攻击分为白盒、灰盒和黑盒攻击。这3种模型之间的差异在于攻击者了解的信息。在白盒攻击的威胁模型中，假定攻击者具有关于其目标模型的完整知识，包括模型体系结构和参数。因此攻击者可以通过任何方式直接在目标模型上制作对抗性样本。在灰盒威胁模型中，攻击者了解的信息仅限于目标模型的结构和查询访问的权限。在黑盒威胁模型中，攻击者只能依赖查询访问的返回结果来生成对抗样本。在这些威胁模型的框架中，研究者开发了许多用于对抗样本生成的攻击算法，比如基于有限内存的BFGS（limited-memory Broyden-Fletcher-Goldfarb-Shan-

* Corresponding author.
E-mail address: kuiren@zju.edu.cn (K. Ren).

2095-8099/
英文原文: Engineering 2020, 6(3): 346–360
引用本文: Kui Ren, Tianhang Zheng, Zhan Qin, Xue Liu. Adversarial Attacks and Defenses in Deep Learning. *Engineering*, https://doi.org/10.1016/j.eng.2019.12.012

no, L-BFGS）[4]、快速梯度符号法（fast gradient sign method, FGSM）[5]、基本迭代攻击/投影梯度下降（basic iterative attack/projected gradient descent, BIA/PGD）[6]、分布式对抗攻击（distributionally adversarial attack, DAA）[7]、Carlini和Wagner（C&W）攻击[8]、基于雅可比的显著图攻击（Jacobian-based saliency map attack, JSMA）[9]以及DeepFool [10]。尽管这些攻击算法最初是在白盒威胁模型下设计的，但是由对抗样本在模型之间的可传递性可知：它们同样适用于灰盒威胁模型和黑盒威胁模型[11,12]。

同时，近年来我们还发现了多种对抗性样本分类/检测的防御技术，包括启发式和可证明式防御。启发式防御是指对某些特定攻击可能具有良好性能，但没有给出防御性能的理论性保障。当前最成功的启发式防御是对抗训练，它试图通过将对抗样本纳入训练阶段来提高模型的鲁棒性。根据经验结果，PGD对抗训练可在MNIST、CIFAR10和ImageNet [13,14]等多个基准数据集上抵御各种L_∞攻击，从而得到当前最好的防御效果。其他启发式防御可能依赖于输入/特征转换和降噪来减轻数据/特征域中对抗样本的影响。相反，在明确知道对抗性攻击类别的情况下，可证明式防御能够计算模型输出的最低精度。最近流行的可证明式防御是制定对抗性多面体，并通过凸松弛来限制它的上界。宽松弛过后的上界可以作为已训练模型的一个保障，它可以证明在限定的限制条件下，没有任何攻击可以超过该上界对应的攻击成功率。但是这些可证明式防御措施的实际性能仍然比对抗训练的性能差很多。

本文将调查并总结对抗性攻击和防御研究领域中最前沿的研究成果。此外，我们将根据目前最新的研究进展对这些攻击和防御方式的有效性进行评述。本文的其余部分安排如下：第2节首先介绍背景；第3节详细介绍具有代表性的攻击方法；第4节介绍对抗性攻击在工业某些潜在领域中的应用；第5节介绍最新的防御方法；第6节讨论我们对该研究领域的见解，包括白盒和黑盒攻击技术之间的差异、对抗攻击和防御发展趋势的差异、模型鲁棒性的最新理论结果、面临的主要挑战；第7节总结全文。

2. 预备知识

2.1. 定义和符号

本节将阐明本文中使用的定义和符号。本文中数据集定义为$\{x_i, y_i\}_{i=1}^N$，其中，x_i是带有标签y_i的数据样本；N是数据集的大小。我们将神经网络记为$f(\cdot)$，其中输入为x，预测结果为$f(x)$。相应的优化损失函数（也称为对抗损失函数）用$J(\theta, x, y)$表示，其中，θ表示模型权重。对于分类任务，将$f(x)$与标签（独热编码）y之间的交叉熵用作优化损失函数，用$J(f(x); y)$表示。当样本x'在特定距离度量函数下接近样本x，但$f(x') \neq y$时，称样本x'是样本x的对抗样本。将x的对抗样本定义为：

$$x' : D(x, x') < \eta, f(x') \neq y \tag{1}$$

式中，$D(\cdot,\cdot)$是距离度量函数；η是预定义的距离约束，也称为允许扰动。根据经验可以利用较小的η来保证x和x'之间的相似性，从而使x'与x不可区分。

2.2. 距离度量

由上述定义知，对抗样本x'和良性样本x在特定的距离量度方式下应该很接近。最常用的距离度量是L_p距离度量[8]。x和x'之间的L_p距离用$||x–x'||_p$表示，其中，$||\cdot||_p$定义为：

$$||v||_p = (|v_1|^p + |v_2|^p + \cdots + |v_r|^p)^{1/p} \tag{2}$$

式中，p是实数。

具体来说，L_0距离表示因为对抗攻击而发生修改的良性样本x中的元素的数量；L_2距离测量x和x'之间的标准欧式距离。最受欢迎的距离度量方式是L_∞距离，该距离测量良性样本和对抗样本之间的对应元素值最大的差异。对于离散数据，也有几种对抗攻击方式，这些攻击应用了其他距离度量，如文献[15]中的删除点数和文献[16]中的语义相似度。

2.3. 威胁模型

对抗攻击和防御有3种主流的威胁模型，即黑盒模型、灰盒模型和白盒模型。这3个模型是根据攻击者所知道的待攻击模型信息量的多少定义的。在黑盒模型中，攻击者不知道其目标模型的结构和参数，但是它们可以与模型进行交互，以查询某些特定数据的预测结果。攻击者将查询得到的成对的数据、预测结果和其他的良性对抗样本用于替代的分类器，并在替代的分类器上生成对抗样本。由于对抗性样本的可传递性，黑盒攻击会损害正常训练的非防御性模型。在灰盒攻击中，假

定攻击者没有模型权重，但知道其目标模型的体系结构，并且还可以在发起攻击之前与模型进行交互。在这种威胁模型中，攻击者会在相同网络体系结构的替代分类器上制作对抗样本。由于存在额外的网络结构信息，因此在攻击性能方面，灰盒攻击总是比黑盒攻击的效果更好。最强大的攻击方式是白盒攻击，这种攻击可以完全访问其目标网络的模型（包括所有参数），这意味着攻击者可以直接在目标网络模型上制作对抗样本。目前许多防御措施可以有效防御黑盒/灰盒攻击，但却对白盒无能为力。例如，ICLR2018中9种启发式防御里有7种会受到文献[17]中提出的自适应白盒攻击的破坏。

3. 对抗攻击

本节将介绍一些具有代表性的对抗攻击算法。这些算法主要用于攻击图像分类的模型，但也可以应用于其他DL模型。第4节详细介绍针对其他通用DL模型的特定对抗攻击。

3.1. L-BFGS

文献[4]首先发现了深度神经网络（deep neural network, DNN）无法有效处理对抗样本的情况，作者发现某些难以察觉的对抗扰动会引起模型对图片的分类错误。文献[4]提出了一种称为L-BFGS的方法，通过最小化L_p范数可以找到欺骗DNN的对抗性扰动，其公式为：

$$\min_{x'} ||x - x'||_p \text{ subject to } f(x') \neq y' \tag{3}$$

式中，$||x–x'||_p$是对抗性扰动的L_p范数；y'是对抗攻击的目标标签（$y' \neq y$）。由于此优化问题不易求解，因此文献[4]提出最小化混合损失，即用$c||x–x'||_p + J(\theta, x', y')$近似代替该优化问题的目标函数，并通过线性搜索/网格搜索找到最优解c。

3.2. 快速梯度符号法

Goodfellow等[5]首先提出了一种有效的无目标攻击方法，称为快速梯度符号法（FGSM），该方法通过在良性样本的L_∞范数限制下生成对抗样本，如图1所示。FGSM是典型的一步攻击算法，它沿着对抗性损失函数$J(\theta, x, y)$的梯度方向（即符号）执行一步更新，以增加最陡峭方向上的损失。FGSM生成的对抗性样本表示如下：

$$x' = x + \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y)) \tag{4}$$

式中，ϵ是扰动大小。通过降低$J(\theta, x, y')$的梯度（其中y'表示目标类别）可以将FGSM轻松地扩展为目标攻击算法（targeted FGSM）。如果将交叉熵作为对抗损失，则此更新过程可以减少预测概率向量和目标概率向量之间的交叉熵。目标攻击算法的梯度更新可以表示为：

$$x' = x - \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y')) \tag{5}$$

此外，在良性样本上先添加随机扰动再执行FGSM可以提高FGSM生成对抗样本的性能和多样性。

3.3. 基本迭代攻击和投影梯度下降

Kurakin等[6]提出了BIA方法，该方法通过将一个迭代优化器迭代优化多次来提高FGSM的性能。BIA以较小的步长执行FGSM，并将更新后的对抗样本裁剪到有效范围内，通过这样的方式总共迭代T次，在第k次迭代中的梯度更新方式如下：

$$x'_{t+1} = \text{Clip}\{x'_t + \alpha \cdot \text{sign}(\nabla_x J(\theta, x'_t, y))\} \tag{6}$$

式中，$\alpha T = \epsilon$。投影梯度下降（PGD）可以看作是BIA的广义形式，这种方法没有约束$\alpha T = \epsilon$。为了约束对抗

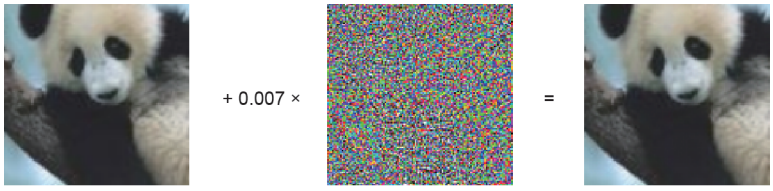

图1. 通过将FGSM应用于GoogleNet产生的对抗性样本示例[5]。FGSM生成的难以察觉的干扰使GoogleNet将该图像识别为长臂猿。

性扰动，PGD将每次迭代学习的对抗性样本投影到良性样本的ϵ- L_∞邻域中，从而使对抗性扰动值小于ϵ。其更新方式如下：

$$x'_{t+1} = \text{Proj}\{x'_t + \alpha \cdot \text{sign}(\nabla_x J(\theta, x'_t, y))\} \quad (7)$$

式中，Proj会将更新后的对抗样本投影到ϵ- L_∞邻域和有效范围内。

3.4. 动量迭代攻击

受动量优化器的启发，Dong等[18]提出将动量记忆集成到BIM的迭代过程中，并推导了一种新的迭代算法Momentum Iterative FGSM (MI-FGSM)。该方法通过以下方式迭代更新其对抗样本：

$$x'_{t+1} = \text{Clip}\{x'_t + \alpha \cdot \text{sign}(g_{t+1})\} \quad (8)$$

式中，梯度g_{t+1}通过$g_{t+1} = \xi \cdot g_t + \nabla_x J(\theta, x_t', y)/||\nabla_x J(\theta, x_t', y)||_1$更新。文献[18]提出的方案是以一组集成模型为目标，在黑盒/灰盒设置下攻击一个不可见的模型。其基本思想是考虑多个模型相对于输入的梯度，并综合确定一个梯度方向，这种攻击方法生成的对抗样本更可能转移攻击其他黑盒/灰盒模型。MI-FGSM与集成攻击方案的结合在NIPS 2017无目标攻击和度量攻击竞赛（黑盒设置）中获得了第一名。

3.5. 分布式对抗攻击

Zheng等[7]提出了一种新的对抗攻击方法分布式对抗攻击（DAA），该方法在概率度量的空间上运行。其与PGD不同的是：PGD会针对每个良性样本独立生成对抗性样本，而DAA对潜在的对抗性分布执行优化。此外该方法提出的目标函数首次将对抗样本盒良性样本数据分布之间的KL散度包含在了对抗损失函数中，从而在优化过程中增加了对抗性攻击泛化的强度。此分布优化问题可以表示如下：

$$\max_{\mu} \int_{\mu} J(\theta, x', y) \mathrm{d}\mu + \text{KL}(\mu(x') || \pi(x)) \quad (9)$$

式中，μ 表示对抗性数据分布；$\pi(x)$表示良性数据分布。

由于对分布进行直接优化比较困难，因此作者利用两种粒子优化方法进行近似。与PGD相比，DAA探索了新的生成对抗样本模式，如图2所示。值得注意的是，DAA在MIT MadryLab的白盒排行榜上排名第二[13]，是对当时几种防御措施最有效的L_∞攻击之一。

3.6. Carlini 和 Wagner 攻击

Carlini和Wagner [8]提出了一组基于优化的对抗攻击C&W，它们可以生成L_0、L_2和L_∞范数限制下的对抗样本CW_0、CW_2和CW_∞。与L-BFGS类似，将优化目标函数[8]表示为：

$$\min_{\delta} D(x, x+\delta) + c \cdot f(x+\delta) \ \textit{subject to}\ x+\delta \in [0,1] \quad (10)$$

式中，δ 是对抗扰动；$D(\cdot,\cdot)$表示L_0、L_2或L_∞距离度量；$f(x+\delta)$是自定义的对抗损失，当且仅当DNN的预测为攻击目标时才满足$f(x+\delta) \leqslant 0$。为了确保$x+\delta$产生有效的图像（即$x+\delta \in [0,1]$），引入了一个新变量来代替δ [8]，如式（11）所示：

$$\delta = \frac{1}{2}(\tanh(\kappa) + 1) - x \quad (11)$$

这样，$x+\delta = 1/2(\tanh(\kappa) + 1)$在优化过程中始终位于[0, 1]中。除了在MNIST、CIFAR10和ImageNet的正

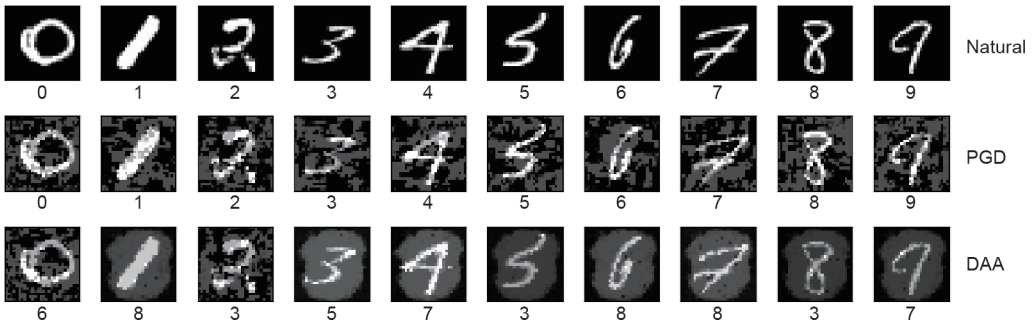

图2. PGD和DAA之间的比较。DAA倾向于产生更多的结构化扰动[7]。

常训练DNN模型上获得100%的攻击成功率外，C＆W攻击还可以破坏防御性蒸馏模型，而这些模型可以使L-BFGS和Deepfool无法找到对抗性样本。

3.7. 基于雅可比的显著性图方法

Papernot等[9]提出了一种称为基于雅可比的显著性图方法（JSMA）的有效攻击方式，它可以利用较小的L_0扰动来欺骗DNN。该方法首先在softmax层之前计算logit层输出的$l(x)$的雅可比矩阵：

$$\nabla l(x) = \frac{\partial l(x)}{\partial x} = \left[\frac{\partial l_j(x)}{\partial x_\gamma}\right]_{\gamma\in 1,\cdots,M_{\text{in}},j\in 1,\cdots,M_{\text{out}}} \quad (12)$$

雅可比矩阵表示输入x的各个分量如何影响不同类别的logit层输出。然后根据雅可比矩阵，攻击者定义了一个对抗性显著图$S(x, y')$用以选择应该受到干扰的特征/像素，以便在logit层的输出中获得所需的变化。他们选择扰动具有最高$S(x, y')[\gamma]$的元素x_i，从而增加目标类别的logit层输出或显著减少其他类别的logit层输出，这样对一小部分元素的扰动已经可以影响$l(x)$并欺骗神经网络。

3.8. Deepfool

Moosavi-Dezfooli等[10]提出了一种新的称为Deepfool的算法，该算法可以在仿射二进制分类器和通用二进制可微分类器上找到最小化L_2范数的对抗性扰动。对于仿射分类器$f(x) = w^{\mathrm{T}}x+b$，更改样本x的分类结果的最小扰动就是移动x到决策边界超平面$\mathcal{F} = \{x: w^{\mathrm{T}}x+b = 0\}$，该距离为$-\frac{f(x_0)}{||w||_2^2}w$。对于一般的可微分类器，Deepfool假设$f$在$x_t'$的领域是线性的，并且迭代计算扰动$\delta_t$：

$$\underset{\delta_t}{\text{argmin}}\,||\delta_t||_2 \text{ subject to } f(x_t') + \nabla f(x_t')^{\mathrm{T}}\delta_t = 0 \quad (13)$$

该过程将持续到$f(x_t') \neq f(x)$为止，并最终用δ_t的总和来近似最小扰动。该方法也可以扩展为攻击通用的多分类器，只需要将问题改变为计算从x到所有类之间的决策边界形成的凸多面体P表面的距离即可，如图3 [10]所示。实验表明，在一些基准数据集上，Deepfool产生的扰动小于FGSM。

3.9. 对 DNN 的弹性网攻击

Chen等[19]提出的对抗攻击将对抗样本的生成过程视为一个弹性网正则化优化的问题，即对DNN的弹性网攻击（elastic-net attack to DNN, EAD）。总的来说，EAD希望在同时最大化减少L_1和L_2距离度量下扰动的前提下，能找到可以欺骗神经网络的对抗性样本。该优化问题表述为：

$$\min_{x'}\; cJ(\theta, x', y') + \beta||x'-x||_1 + ||x'-x||_2^2 \quad \text{subject to } x' \in [0,1] \quad (14)$$

式中，$J(\theta, x', y')$是目标对抗损失函数，$\beta||x'-x||_1+||x'-x||_2^2$用于惩罚对抗性样本$x'$与良性样本$x$之间的$L_1$和$L_2$距离。EAD首次在对抗攻击中引入了$L_1$范数约束，并产生了一组能与其他最新方法一较高下的对抗样本。

3.10. 通用对抗攻击

上述所有攻击都是针对良性样本来精心设计对抗性扰动的。换句话说，对抗性扰动不会在良性样本之间传递。因此一个自然的问题是：是否存在一种普遍的扰动会欺骗大多数良性样本的网络？文献[20]首先尝试通过使用所有目标良性样本迭代更新扰动来寻求这种扰动向量。在每次迭代中，对于当前扰动无法欺骗的良性样本，将求解一个类似于L-BFGS [4]的优化问题，以找到危害这些样本所需的最小附加扰动。附加扰动将添加到当前

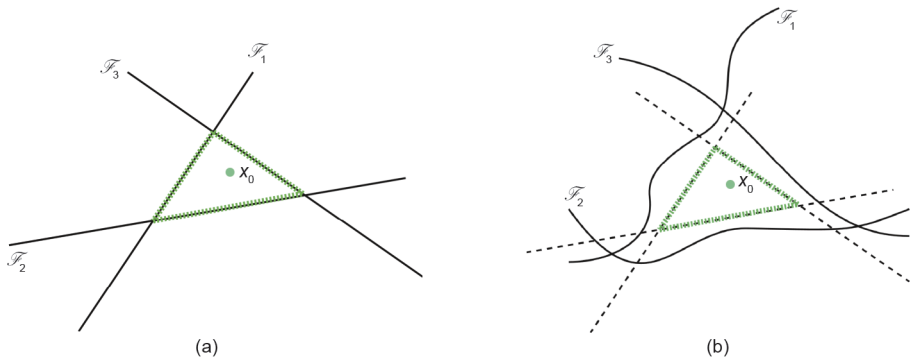

图3. 所有类别之间的决策边界形成的凸多面体。（a）线性模型；（b）非线性模型[10]。

扰动中。最终，扰动使大多数良性样本欺骗了网络。实验表明，这种简单的迭代算法可以有效地攻击深度神经网络，如CaffeNet [21]、GoogleNet [22]、VGG [23]和ResNet [24]。出乎意料的是，这种可在不同样本中传递的扰动同时可以应用到其他不同的模型中，例如，在VGG上制作的通用扰动在其他模型上也可以达到53%以上的欺骗率。

3.11. 对抗补丁

直接使用上述攻击算法将会对良性样本的所有分量（如良性图像中的所有像素）进行干扰。最近的研究[25,27,28]显示良性样本的局部区域/片段中的扰动也可能使DL模型失控，我们称这种扰动为对抗补丁（adversarial patch）。Sharif等[25]提出了只在面部图片里附着的眼镜架上制作对抗性扰动的方法，如图4所示。该方法通过优化常用的对抗性损失（即交叉熵）从而使局部产生的扰动也能轻易欺骗VGG-Face卷积神经网络（convolutional neural network, CNN）[26]。作者还3D打印了这种带有干扰的眼镜，并在物理世界中实现了对抗攻击。他们还提供了视频演示，其中戴有对抗眼镜的人们被真实的VGG-Face CNN系统识别为攻击目标。文献[27]提出了一种生成通用鲁棒对抗补丁的方法。这种方法基于良性图像、补丁变换和补丁位置定义了优化补丁的对抗损失函数，然后通过将补丁放在所有良性图像上进行优化来实现通用性，此外作者通过使用EoT方法[28]计算不敏感于噪声盒三维变换的梯度用于优化，从而实现了对噪声和三维变换的鲁棒性。Liu等[29]提出在良性样本上添加特洛伊木马补丁，以生成对抗性样本。该攻击首先选择几个能显著影响网络输出的神经元，然后初始化对抗补丁区域中的像素值，从而使选定的神经元达到最大值，最后利用良性图像和带有特洛伊木马补丁的图像对模型进行重新训练，以调整与所选神经元相关的权重。尽管在良性图像上该模型执行的操作与原始模型类似，但重新训练的模型会在带有对抗补丁的图像上表示出恶意行为。

3.12. 基于生成对抗网络的攻击

Xiao等[30]首先提出利用生成对抗网络（generative adversarial network, GAN）生成对抗样本的方法。该方法通过最大化目标对抗损失$J(\theta, x', y')$和GAN损失L_{GAN}来训练生成器，以使其学习对抗样本的分布。软铰链损失被作为惩罚来约束生成的对抗样本x'和良性样本x之间的L_p距离。值得注意的是，在动态蒸馏的设置下，代理分类器（蒸馏模型）也会通过目标分类器在生成的对抗样本上的输出与生成器一起训练，该攻击将Madry-Lab's MNIST未知模型的准确性降低到92.74% [13]。这是目前最好的黑盒攻击结果。Yang等[31]训练辅助分类器生成对抗网络（auxiliary classifier GAN, AC-GAN）[32]，从而对每个类别的数据分布建模。他们提出的攻击是通过优化明确定义的目标函数来实现的，这样可以找到特定类别的潜在代码从而生成样本，这些样本会被目标分类器分类为另一个类别。由于生成的对抗性样本与任何现有的良性样本都不接近，因此将它们称为非限制性对抗样本。这种攻击不遵循通常为对抗性样本定义的常规约束，因此它能更有效地攻击满足常规约束的对抗性训练模型。

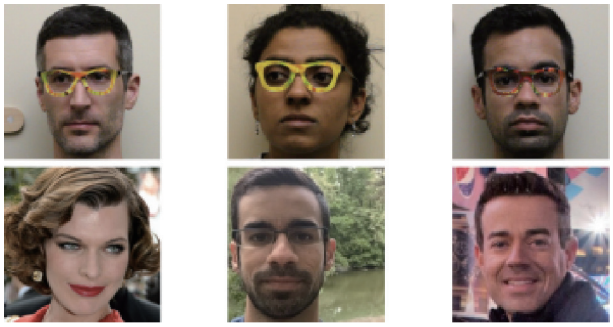

图4. 具有对抗扰动的眼镜欺骗了一种面部识别系统，可以将第一行中的人脸错误的识别为第二行中的人脸[25]。

3.13. 实践性攻击

尽管PGD和C＆W等对抗性攻击算法在数字领域非常有效，但将其扩展到物理世界仍然需要克服两个关键问题。第一个问题是环境噪声和自然变化将破坏数字空间中计算出的对抗性扰动。例如，模糊、噪声和联合图像专家小组（joint photographic experts group, JPEG）编码等会对对抗性攻击的破坏率超过80%[6]。第二个问题是攻击仅限于使用图像/视频的ML任务，其中只有与某些对象相对应的像素才能在物理世界中被干扰，也可以说攻击者不可能干扰背景。Athalye等[28]提出了一种称为转换期望（expectation over transformation, EoT）的方法来解决第一个问题。EoT不是使用理想数字域中计算出的梯度，而是在输入上添加/应用了一组随机噪声/自然变换，然后取这些噪声/自然变换的输入计算得到梯度的平均值用于优化。在基于梯度的攻击算法（如FGSM和PGD）中采用这种平均梯度，可以提高生成的

对抗样本的鲁棒性。对抗补丁的思想则可以简单地解决第二个问题，即空间约束。Eykholt等[29]提出了一种掩模/补丁变换来分离背景和目标，从而可以将对抗性扰动限制在目标区域内。此外，文献[33]还考虑了因为打印和受扰动RGB值之间的差异引起的制造误差，如图5所示。这种方法的特点是在优化损失中包含一个附加惩罚项，称为不可打印分数（non-printable score, NPS）。最终文献[29]成功地在现实世界的交通标志上生成了可打印的对抗干扰，总体攻击成功率达到80%以上。

3.14. 混淆梯度规避攻击

Athalye等[17]说明了大多数启发式防御方法（包括ICLR2018中公布的9种防御中的8种）所共有的一个常见问题。这一问题是这些防御模型的梯度要么是不存在的，要么是由于采用诸如量化和随机化之类的附加操作而不确定。对于这些防御，文献[17]提出了3种方法可以绕过附加组件操作，从而获得用于生成对抗样本的有效梯度。第一种：对于依赖于不可微的附加操作（如量化）的防御，通过使用可微函数来逼近它们；第二种：对于设置有随机变换等不确定性操作的防御系统，只需使用EoT [28]来确定所有可能变换中的一般梯度方向的期望，并沿着该一般梯度方向更新对抗样本；第三种：对于由优化循环引起的梯度爆炸或消失的防御方法，提出进行变量更改以便将优化循环转换为可微函数。利用这3种方法近似的梯度，打破了ICLR2018中9种启发式防御中的7种[17]。

4. 对工业界广泛使用的应用进行对抗性攻击

第3节主要介绍了一些典型的攻击算法，其中大多数最初是为图像分类而设计的。但是这些方法也可以应用于其他领域，如图像/视频分割[34,35]、3D识别[36,37]、音频识别[38]和强化学习[39]，这引起了学术界和工业界越来越多的关注。因为特定的数据和应用程序可能导致独特的对抗攻击，所以在本节中，我们还将概述那些针对其他普及应用独特的对抗攻击。

图5.（a）显示了由原始Inception v3模型识别为微波炉的原始图像；（b）显示了被识别为电话的对抗样本[33]。

4.1. 语义分割模型中的对抗性攻击

Xie等[40]首先提出了一种密集对抗生成算法（dense adversarial generation, DAG），该算法可以用于生成目标检测和语义分割任务的对抗样本。DAG的基本思想是同时考虑检测/分割任务中的所有目标并优化总体损失，如图6所示。此外，为了解决像素级对象检测任务中的大量候选单元[即以$O(K^2)$缩放，其中K是像素数]，DAG通过在优化阶段修改交并比（intersection-over-union, IoU）来保证候选单元数量增加并维持在合理的范围。在文献[41]中，作者发现在分割任务中没有很好地建立对抗性损失与分类任务中的准确度之间的关系。因此他们提出了一种新的替代损失Houdini，并用这个损失来逼近实际对抗损失，它是随机边际和任务损失的乘积。随机边际表示真值与预测目标的预测概率之间的差值，并且任务损失与模型无关，它对应于最大化目标函数。文献[41]进一步推导出在给定输入的情况下新替代损失梯度的近似值，从而可以对输入进行基于梯度的优化。实验表明，Houdini算法在语义分割方面达到了最先进的攻击性能，人眼很难识别出这种对抗性扰动。

4.2. 3D 识别中的对抗性攻击

点云是3D物体识别的重要数据表示形式。PointNet [35]、PointNet ++ [42]和动态图CNN（dynamic graph CNN, DGCNN）[43]是基于点云的分类/分割的3种最受欢迎的DL模型。但是，最近发现这3种模型也容易受到对抗攻击[15,44,45]。在文献[44]中，作者首先将C＆W攻击扩展到这些3D点云DL模型。在文献[44]中，空间点位置与像素值相对应，通过移动空间点（即扰动空间点的位置）来优化C＆W损失。同样，文献[45]将BIA/

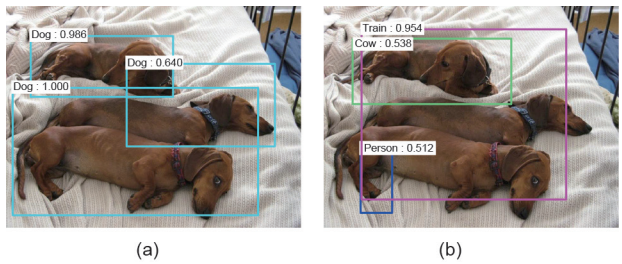

图6. 在左侧的良性图像中，Faster CNN正确地检测了三只狗并识别了它们的区域，而在DAG生成的右侧对抗图像中，检测结果是完全错误的[40]。

PGD应用于点云分类并实现了很高的攻击成功率。在文献[15]中，作者基于丢弃点云中现有的空间点的思想提出了新的攻击方法。该方法通过将点移动到点云的质心来近似每个点对分类结果的贡献，然后丢弃具有较大正贡献的点。随着一定数量的这些点被丢弃，PointNet、PointNet ++和DGCNN的分类精度显著降低。对于点云模型，文献[46]建议在三维网格上添加对抗性扰动，以便三维网格的二维投影可以误导二维图像分类模型。这种三维攻击是通过对目标二维模型带有对抗损失的混合损失和保证三维网格对抗样本看起来像实物的损失函数进行优化来实现的。

4.3. 音频和文本识别中的对抗性攻击

Carlini和Wagner [47]通过对C&W损耗函数的优化，成功地构建了高质量的音频对抗性样本。对于任何音频信号，只要在DeepSpeech上对音频信号的1%进行对抗性干扰，即可在文本翻译中最多修改50个单词[48]。他们还发现，构建的对抗性音频信号对点噪声和MP3压缩具有鲁棒性。但是由于麦克风和录音机的非线性影响，被扰动的音频信号在空气中播放后不会保持对抗性。文献[49]中的作者提出一种模拟非线性效应和噪声，并在攻击过程中将它们考虑在内的攻击方法。具体而言，作者将接收信号建模为发射信号的函数，该建模包括模型受到带通滤波器、脉冲响应和白高斯噪声的影响的变换。对抗损失函数是在接收信号而不是发射信号上定义的。这种方法成功地在物理世界中产生了对抗性的音频样本，即使在空气中播放，也能攻击音频识别模型。在文本识别领域，Liang等[50]提出了3种针对文本数据的词级干扰策略，包括插入、修改和删除。攻击者首先确定影响分类结果的最重要的文本项，然后对这些文本项采用其中一种扰动方法。实验表明，这种攻击可以成功地欺骗一些基于DNN的最新文本分类器。TextBugger [16]对文本数据采用了5种干扰操作，包括插入、删除、交换、字符替换和单词替换，如图7所示。在白盒设置中，先使用雅可比矩阵[9]识别重要单词，再对这些单词执行上述5种干扰操作。在黑盒威胁模型中，矩阵是不可计算的，但假设攻击者有权获得句子和文档预测结果的置信值，在这种情况下，每个句子的重要性被定义为它对预测类的置信值。在对分类结果最重要的句子中，每个词的重要性是由含有该词和不含有该词的句子的置信值之间的差来定义的。

4.4. 深度强化学习中的对抗性攻击

Huang等[51]发现通过在策略的原始输入上添加对抗性扰动，现有的攻击方法也可用于在深度强化学习中降低已训练策略的性能。比如构建一个替代损失函数$J(\theta, x, y)$，这个函数包含参数θ、策略的输入x以及所有可能操作y的加权得分[51]。FGSM [5]对3种不同前馈训练策略的算法进行了攻击，包括深度Q-networks（deep Q-network, DQN）[52]、异步优势动作评价（asynchronous advantage actor-critic, A3C）[53]和信任区域策略优化（trust region policy optimization, TRPO）[54]。在大多数情况下，白盒攻击可以将智能体的准确性降低50%，另外，在黑盒攻击同样也有效。尽管攻击性能可能会下降，但这3种算法之间的对抗效果是能够相互转移的。文献[55]建议扰动Q函数$Q(s_{t+1}, a, \theta_t)$中的输入状态s_t，从而在学习过程中产生对抗性动作a'。文章中推荐使用FGSM和JSMA算法生成对抗性扰动。Lin等[56]提出了深度强化学习的两种攻击策略，包括策略定时攻击和附魔攻击。策略定时攻击是指仅在几个特定的时间步长内扰动图像输入从而使奖励最小化。这种攻击是通过优化奖励的扰动来进行的。附魔攻击可以对抗性地干扰图像帧从而将智能体引诱到目标状态。这种攻击需要一个生成模型来预测未来的状态和动作，然后产生一个误导性的动作序列指导生成图像帧上的扰动。

5. 对抗防御

本节将总结近年来具有代表性的对抗防御方法，主要包括对抗训练、基于随机化的方法、降噪方法、可证

Task: sentiment analysis. **Classifier:** CNN. **Original label:** 99.8% negative. **Adversarial label:** 81.0% positive.

Text: I love these ~~awful~~ **awf ul** 80's summer camp movies. The best part about "Party Camp" is the fact that it ~~litorally~~ **literaly** has ~~no~~ **No** plot. The ~~eliches~~ **clichs** here are limitless: the nerds vs. the jocks, the secret camera in the girls locker room, the hikers happening upon a nudist colony, the contest at the conclusion, the secretly horny camp administrators, and the ~~embarrassingly~~ **embarrasing1y** ~~foolish~~ **fo0lish** sexual innuendo littered throughout. This movie will make you laugh, but never intentionally. I repeat, never.

图7. TextBugger生成的对抗文本[16]。负面评论被误分类为正面评论。

明的防御以及其他一些新防御。我们还将简要讨论它们在不同环境下对不同攻击的有效性。

5.1. 对抗训练

对抗训练是一种针对对抗样本的直观防御方法，该方法试图通过利用对抗样本进行训练来提高神经网络的鲁棒性。从形式上讲，这是一个Min-Max的游戏，可以表述为：

$$\min_{\theta} \max_{D(x,x') < \eta} J(\theta, x', y) \tag{15}$$

式中，$J(\theta, x', y)$是对抗损失函数；θ是网络权重；x'是对抗输入；y是标签真值。$D(x, x')$表示x和x'之间的某种距离度量。内部的最大化优化问题是找到最有效的对抗样本，这可以通过精心设计的对抗攻击实现，如FGSM [5]和PGD [6]。外部的最小化优化问题是损失函数最小化的标准训练流程。最终的网络应该能够抵抗训练阶段用的生成对抗性样本的对抗性攻击。最近的研究[13,14,57,58]表明：对抗性训练是对抗性攻击最有效的防御手段之一。主要是因为这种方法在几个基准数据集上达到了最高的精度。因此在本节中，我们将详细介绍过去几年里表现最好的对抗训练技术。

5.1.1. FGSM 对抗训练

Goodfellow等[5]首先提出用良性和FGSM生成的对抗样本训练神经网络以增强其网络鲁棒性的方法。他们提出的对抗目标函数可以表达为：

$$J(\theta, x, y) = cJ(\theta, x, y) + (1 - c)J(\theta, x + \epsilon \cdot \mathrm{sign}(\nabla_x J(\theta, x, y)), y) \tag{16}$$

式中，$x + \epsilon \cdot \mathrm{sign}(\nabla_x J(\theta, x, y))$是良性样本$x$根据FGSM方法生成的对抗样本；$c$是用于平衡良性和对抗性样本的准确性。文献[5]中的实验表明，该网络对于FGSM方法生成的对抗性样本变得有些稳健。具体来说，在对抗训练中对抗样本的错误率从89.4%急剧下降至17.9%。尽管该方法对FGSM的攻击有效，但是训练后的模型仍然容易受到基于迭代/优化方式的对抗攻击。因此许多研究进一步挖掘了具有更强对抗性攻击（如BIA/PGD攻击）的对抗性训练。

5.1.2. PGD 对抗训练

大量评估表明，PGD攻击可能是通用的一阶L_∞攻击[13]。如果是这样，那么针对PGD的模型鲁棒性意味着可以抵抗各种一阶L_∞攻击。基于这种推测，Madry等[13]提出使用PGD对抗训练一个稳健的网络，出人意料的是PGD对抗训练确实提高了CNN和ResNets的鲁棒性[24]，从而能够抵抗几种具有代表性的一阶L_∞攻击，如在黑盒和白盒设置下的FGSM、PGD和CW_∞攻击。即使是当前最强的L_∞攻击（即DAA），也只能将PGD对抗训练的MNIST模型和CIFAR-10模型的准确性分别降低到88.56%和44.71%。在最近的对抗攻击与防御竞赛（competition on adversarial attack and defense, CAAD）中，针对ImageNet对抗样本的一流防御实际上也依赖于PGD对抗训练[14]。通过PGD对抗训练，原始的ResNets [23]在20步PGD下已经达到了50%以上的精度，而文献[14]中提出的降噪架构实际上仅进一步将精度提高了3%。以上所有研究和结果表明，PGD对抗训练总体上是针对L_∞攻击的最有效对策。但是由于生成PGD对抗样本需要大量计算成本，因此PGD对抗训练不是一种有效率的防御措施。例如，在TITAN-V图形处理器（graphics processing unit, GPU）上使用针对CIFAR-10数据集的简化ResNet进行PGD对抗训练大约需要3 d，而在CAAD中排名第一的模型在128个Nvidia V100 GPU上花费了52 h。此外，PGD对抗训练模型仅对L_∞攻击具有鲁棒性，这样的模型依然容易受其他L_p-norm攻击者的攻击，如EAD [19,59]和CW_2 [8]。

5.1.3. 集成对抗训练

为了避免PGD对抗训练带来的大量计算成本，文献[60]提出通过FGSM对抗训练和随机启动（RAND + FGSM）结合来生成鲁棒的ImageNet模型。但是经过对抗训练的模型甚至很容易受到黑盒攻击。为了解决这个问题，文献[61]提出了一种训练方法，该方法利用多个预先训练的模型中转移的对抗样本进行对抗训练，即集成对抗训练（ensemble adversarial training, EAT）[61]。直观上，EAT增加了用于对抗训练的对抗样本的多样性，从而增强了针对从其他模型转移过来的对抗样本的神经网络鲁棒性。实验结果表明，EAT模型对其他模型遭受的各种单步和多步攻击产生的对抗样本都具有较强的鲁棒性。在某些情况下，EAT对黑盒和灰盒攻击的抵抗性能甚至优于PGD对抗训练。

5.1.4. 对抗 Logit 配对

Kannan等[62]提出了一种新的称为对抗Logit配对

（adversarial Logit-pairing, ALP）的对抗训练方法。与文献[63]中提出的稳定性训练策略相似。该方法通过将良性样本x的logits层和相应的扰动样本x'之间的交叉熵包括在训练损失函数中，来鼓励提高成对样本在学习到的Logits层的相似性。唯一的区别是文献[62]中使用的x'是PGD对抗样本。该方法训练的损失函数为：

$$J\left(\theta,x,x',y\right)=J(\theta,x,y)+cJ(\theta,x,x') \tag{17}$$

式中，$J(\theta, x, y)$是原始损失；而$J(\theta, x, x')$是x和x'的Logits层的交叉熵。文献[62]中的实验表明，这种配对损失有助于在多个基准数据集上提高PGD对抗训练的性能，如SVHN、CIFAR-10和小型ImageNet。文献[62]声称ALP在白盒PGD攻击下可以将Inception V3模型的准确性从1.5%提高到27.9%，并且在抵御黑盒攻击方面也表现得与EAT差不多。然而文献[64]评估了经过ALP训练的ResNet的鲁棒性，发现在文献[62]考虑的目标攻击下，ResNet仅能实现0.6%的正确分类率。文献[64]还指出ALP不太适合梯度下降，因为ALP有时会引起“凹凸不平”，例如，在损失函数空间中，输入样本点附近可能被凹陷的损失情况包围。因此ALP可能不如文献[62]中所描述的那样稳健。

5.1.5. 生成对抗训练

以上所有对抗训练策略均采用确定性攻击算法来生成训练样本。文献[65]首先提出在对抗性训练的过程中利用非确定性生成器来生成对抗性样本。作者设置了一个生成器，该生成器的输入是训练好的模型在良性样本点上的梯度，并产生了类似FGSM的对抗性扰动。通过在良性样本和生成样本上训练分类器，与FGSM对抗训练模型相比，文献[65]获得了一个对FGSM具有更强鲁棒性的模型。Liu等[66]首先提出使用AC-GAN架构[32]进行数据扩充，从而提高PGD对抗训练的通用性。通过将PGD对抗样本作为真实样本输入到鉴别器中，AC-GAN学会了生成与PGD对抗性样本相似的伪造样本。类似于PGD的假样本将被用来训练辅助分类器和预训练的鉴别器。根据文献[66]，生成器、鉴别器、辅助分类器和PGD攻击的这种组合在单个网络中不仅会形成更强大的分类器，而且可以形成更好的生成器。

5.2. 随机化

最近的许多防御措施都采用随机化来减轻输入/特征域中对抗性扰动的影响，因为从直觉上看，DNN总是对随机扰动具有鲁棒性。基于随机化的防御方法试图将对抗性效应随机化为随机性效应，当然这对大多数DNN而言都不是问题。在黑盒攻击和灰盒攻击的设置下，基于随机化的防御获得了不错的性能，但是在白盒攻击下，EoT方法[28]能够通过在攻击过程中考虑随机过程来破坏大多数防御方法。本节将详细介绍几种基于随机化的代表性防御方式，并介绍其针对不同环境中各种防御的性能。

5.2.1. 随机输入变换

Xie等[67]利用随机调整大小和填充这两种随机变换来减轻推理时的对抗效果。随机调整大小是指在将输入图像输入DNN之前将其调整为随机大小。随机填充是指以随机方式在输入图像周围填充零。这种快速而敏锐的机制如图8所示。该防御方法在黑盒攻击下取得了卓越的性能，在NIPS 2017对抗样本防御挑战中排名第二，然而在白盒攻击下这种防御会被EoT方法破坏[28]。当使用30个随机调整大小和填充的图像集合来逼近梯度时，EoT能够通过8/255 L_∞扰动将模型精度降低到0。Guo等[68]的防御方法是将图像送入到CNN之前使用具有随机性的图像变换，如位深度减小、JPEG压缩、总方差最小化和图像缝合。这种方法可以抵抗由多种主流攻击方法生成的60%的强灰盒对抗样本和90%的强黑盒对抗样本。但是它也会受到EoT方法的损害[28]。

5.2.2. 随机噪声

Liu等[69]提出名为RSE (random self-ensemble)的随机噪声机制来防御对抗性干扰。在训练和测试阶段，RSE在每个卷积层之前添加一个噪声层，并集成随机噪声的预测结果以确保DNN有稳定的输出，如图9所示[69]。Lecuyer等[70]从差分隐私（differential privacy, DP）的角度看待随机噪声的防御方式[71]，并提出了一种基于DP的防御PixelDIP。PixelDP在DNN内集成了DP噪声层，基于范数的微小扰动会引起预测结果概率分布的变化，通过这种变化可以给出DP边界。PixelDP可在使用Laplacian/Gaussian DP机制防御的前提下抵抗L_1/L_2攻击。受PixelDP的启发，文献[72]中的作者进一步提出在分类之前将随机噪声直接添加到对抗性样本的像素中，从而消除对抗性扰动的影响。基于Renyi散度理论，文献[72]证明了利用输出概率分布

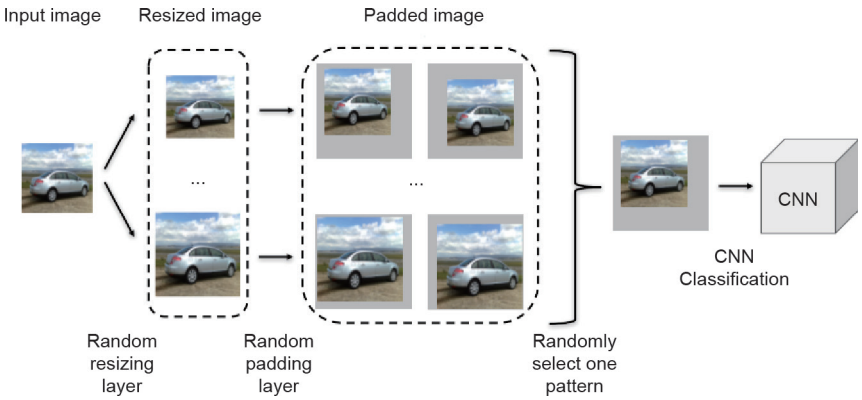

图8. Xie等[67]提出的基于随机化的防御机制流程图。输入图像首先被随机调整大小，然后被随机填充。

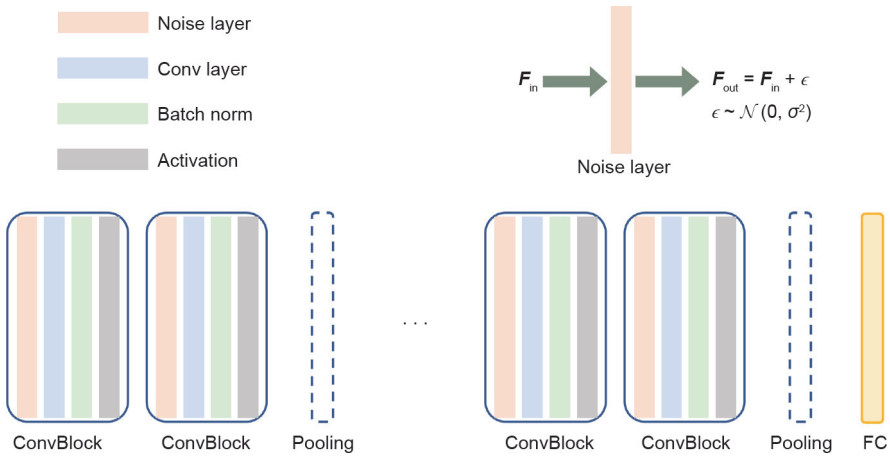

图9. RSE的结构[69]。

（向量）的第一和第二大概率可以确定其对抗扰动的鲁棒上限。

5.2.3. 随机特征修剪

Dhillon等[73]提出了一种称为随机激活修剪（stochastic activation pruning, SAP）的方法进行对抗防御，该方法通过随机修剪每一层中的一部分激活子集，并优先保留更大幅度的激活项来保护预训练网络免受对抗样本的攻击。在激活修剪之后，SAP会扩展尚存的激活从而标准化每个层的输入。但是在CIFAR-10上，EoT [28]还是可以通过8/255 L_∞对抗扰动将SAP的精度降低为0。Luo等[74]通过随机掩盖卷积层输出的特征图设计了一种新的CNN结构。通过随机掩盖输出特征图使每个过滤器仅从局部位置提取要素。Luo等[74]声称这将有助于过滤器学习与掩模图案一致分布的特征，因此CNN可以捕获有关局部特征空间结构的更多信息。

5.3. 去噪

就减轻对抗性扰动/效果而言，降噪是一种非常简单的方法。之前的工作指出了设计这种防御的两个方向，包括输入降噪和特征图降噪。其中第一个方向试图从输入中部分或完全消除对抗性扰动，第二个方向是减轻对抗性扰动对DNN学习高级功能的影响。本节将详细介绍这两个方向上的几种著名防御方法。

5.3.1. 常规输入整流

为了减轻对抗效果，Xu等[75]首先利用两种压缩（去噪）方法：位减少和图像模糊，以减少自由度并消除对抗性扰动，如图10所示。通过比较原始图像和压缩图像上的模型预测值来实现对抗样本检测。如果原始输入和压缩输入产生的输出与模型有很大差异，则原始输入可能是对抗样本。Xu等[76]进一步声称在文献[75]中提出的特征压缩方法可以减轻C＆W攻击，但是He等[77]证明了特征压缩仍然容易受到知识适应性的攻击者

的攻击，在实验过程中采用CW_2损失作为对抗损失。在每步优化之后，作者可从优化器里获得中间图像。Xu等的检测系统会检查通过减小色深获得的中间图像，这种优化将运行多次，所有可以绕过Xu等系统的中间对抗性样本都将被统计，整个自适应攻击可以在比文献[75]中小得多的扰动下破坏输入压缩系统。Sharma和Chen [78]也表明EAD和CW_2可以通过增强攻击者的实力绕过输入压缩系统。

5.3.2. 基于 GAN 的输入清理

生成对抗网络（GAN）是一种功能强大的工具，其可用于学习数据分布并形成生成器。大量的工作试图利用GAN来学习良性数据分布，从而在对抗性输入的前提下生成良性预测。防御-GAN（Defense-GAN）和对抗干扰消除-GAN（adversarial perturbation elimination-GAN, APE-GAN）是这类工作的两个代表。Defence-GAN [79]训练生成器来对良性图像的分布进行建模，如图11 [79]所示。在测试阶段，Defense-GAN通过在其学习的分布中搜索接近于对抗样本的图像来清除对抗样本，然后将良性图像输入分类器。这种策略可以用来防御各种对抗攻击，目前针对Defence-GAN最有效的攻击方案是基于BPDA [17]的攻击方法，它可以通过$0.005L_2$的对抗扰动将其准确性降低到55%。APE-GAN [80]直接学习生成器，它将对抗样本作为输入，输出其对应的良性样本，从而清晰对抗样本。尽管APE-GAN在文献[80]的实验平台上取得了很好的性能，但在文献[81]中指出自适应白盒攻击CW_2可以很容易击败APE-GAN。

5.3.3. 基于自动编码器的输入去噪

在文献[82]中作者介绍了一种称为MagNet的两节防御系统，其中包括一个探测器和一个重整器。在MagNet中使用自动编码器来学习良性样本的多种形式。检测器根据样本与学习到的良性样本的多种形式之间的关系来区分良性样本和对抗样本。重整器用于将对抗样本纠正为良性样本。作者通过实验证明了MagNet可以有效抵抗灰盒和黑盒设置（如FGSM、BIA和C＆W）下的各种对抗攻击。然而文献[81]证明了MagNet容易受到CW_2攻击产生的可转移对抗样本的攻击。

5.3.4. 特征去噪

Liao等[83]提出了一种基于高级表示法指导的去噪器（high-level representation guided denoiser, HGD），这种去噪器可以改善受对抗性扰动影响的特征。HGD不是使用像素级去噪，而是使用特征级损失函数训练降噪的U-NET [34]，这样可以最大限度地减少良性和对抗性样本之间的特征级差异。在NIPS2017比赛中，HGD获得了防御赛道的第一名（黑盒攻击）。尽管这种方法在

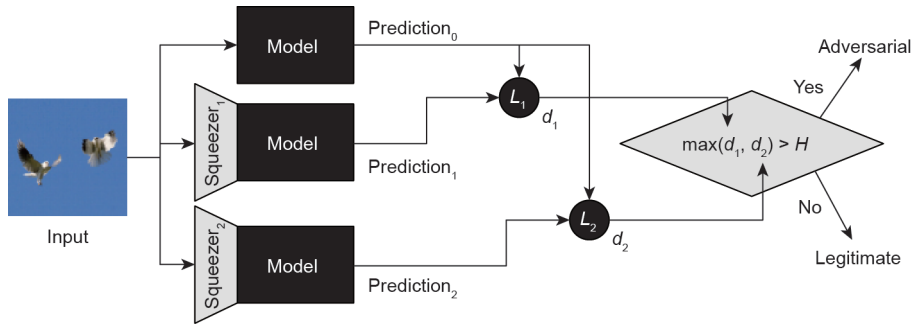

图10. Xu等[75]提出的特征压缩框架。d_1和d_2：模型对压缩输入的预测与原始输入的预测之间的差异；H：用于检测对抗示例的阈值。

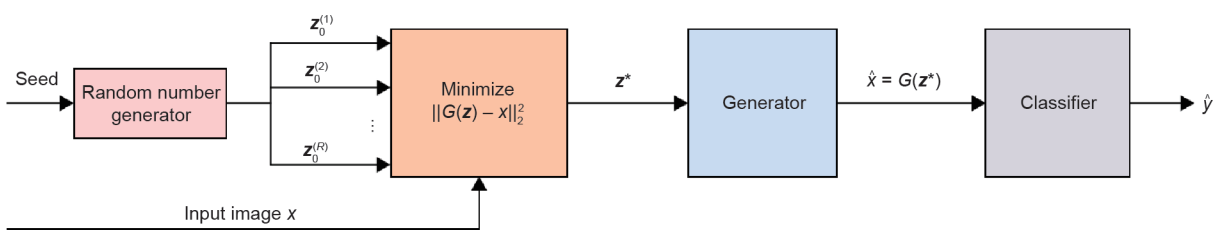

图11. Defense-GAN的流程图[79]。G：可以从低维向量z生成高维输入样本的生成模型；R：随机数生成器生成的随机向量的数量。

黑盒设置下有效，但在白盒设置下HGD会受到PGD攻击者的损害[84]。文献[84]中的实验表明，4/255 L_∞扰动的PGD攻击已经可以将HGD的精度降低到0。Xie等[14]设计了一个块来学习几种去噪操作，从而纠正DNN中间层学习到的特征。改进后的PGD对抗训练网络在CAAD 2018对抗防御赛道中排名第一。尽管Xie等[14]取得了显著的成功，但与PGD对抗训练相比，特征去噪块对网络鲁棒性的贡献并不显著。因为在白盒PGD攻击下，PGD对抗训练的基线ResNet也能达到近50%的准确率，而文献提出的去噪块仅使该基线的准确率提高了3%。

5.4. 可证明式防御

以上所有介绍的防御都是启发式防御，这意味着这些防御的有效性只在实验上得到验证，而没有在理论上得到证明。如果无法计算理论上的错误率，这些启发式防御可能会被未来的新攻击所打破。因此许多研究者致力于探索可证明的防御方法，在一类定义明确的攻击下，这些方法始终能保持一定的准确性。本节将介绍几种具有代表性的可证明的防御。

5.4.1. 基于半定规划的可证明式防御

Raghunathan等[85]首先提出了一种针对两层网络生成的对抗样本的可证明的防御方法。作者导出了攻击者损失函数上界的半定松弛，并将其作为正则化项加入训练损失函数中。

这种训练方法产生的一个网络被证明在不超过0.1/1.0 L_∞扰动的情况下，不会对MNIST造成超过35%的测试误差。在文献[86]中，Raghunathan等进一步提出了一种新的半定性松弛来证明任意ReLU网络的防御性能。新提出的松弛比以前的松弛更严格[85]，并且可以在3个不同的网络上保证其鲁棒性。

5.4.2. 基于对偶方法的可证明式防御

借鉴文献[85]的思路，Wong等[87]提出了一个对抗问题来界定对抗样本的输出多维区域。他们表明可以通过在另一个深度神经网络上进行优化来解决对偶问题。与仅适用于两层全连接网络[85]有所不同，该方法可应用于具有任意线性算子层（如卷积层）的深层网络。文献[88]将文献[87]中的技术扩展到了具有跳过连接和任意非线性激活的更通用的网络。文献[88]还提出了一种非线性随机投影技术，这种方法使得估计边界问题的规模与隐藏元的规模是线性相关的，这使得该方法适用于较大的网络。在MNIST和CIFAR数据集上，利用文献[88]提出的技术训练分类器可以大大改善先前可证明式防御方法的鲁棒性：在$\epsilon = 0.1$的L_∞扰动下，在MNIST上的错误率从5.8%下降到3.1%；在$\epsilon = 2/255$的L_∞扰动的情况下，在CIFAR上的错误率从80%下降到36.4%。

5.4.3. 分布稳健性证明

从分布优化的角度来看，文献[89]将对抗性分布的优化问题表述为：

$$\min_{\theta} \sup_{\varphi \in \phi} E_{\varphi}[J(\theta, x, y)] \tag{18}$$

式中，ϕ是围绕良性数据所有分布的候选集，可以由散度球[90]或Wasserstein球[91]构造。在此分布目标上进行的优化等效于将对良性数据相邻的所有样本（即对抗样本的所有候选项）的经验风险最小化。由于P会影响计算的复杂性，并且难以在任意P上进行直接优化，因此文献[80]使用Wasserstein距离度量和计算效率高的松弛，可求出能处理的集合P，即使$J(\theta, x, y)$不为凸函数也可计算。文献[89]还提供了对抗训练方法，其计算性能和统计性能可以利用数学进行证明。在所提出的训练过程中，文献[89]引入了一个惩罚项来描述对抗稳健性区域。由于很难对该惩罚项进行优化，因此提出了针对该惩罚项的拉格朗日松弛法，从而实现了对分布损失函数的鲁棒优化。作者还能够确保鲁棒鞍点问题的经验性最小值，并给出了域适应问题的专门界限，这也为分布鲁棒性证明提供了启发。

5.5. 稀疏权重 DNN

Guo等[92]首先证明了针对FGSM和DeepFool生成的对抗样本的权重稀疏性和网络稳健性之间的内在联系。对于线性模型，文献[92]证明了对抗样本的优化会导致网络权重的稀疏化。对于非线性神经网络，文献[92]应用了文献[93,94]中确保鲁棒性的方法，并证明当权重矩阵较稀疏时，网络Lipchitz常数倾向于更小。由于研究发现最小化Lipchitz常数有助于提高网络的鲁棒性[93]，因此可以得出结论，稀疏权重也将导致神经网络更鲁棒。文献[95]还显示了权重稀疏性对网络稳健性验证是有益的。作者证明稀疏度可以将计算上难以解决的验证问题转变为易于解决的问题。文献[95]通过用L_1正则化训练神经网络来改善权重稀疏性，并发现权值稀

疏性显著加快了用于网络鲁棒性验证的线性规划（linear programming, LP）求解器[96]。

5.6. 基于 KNN 的防御

Wang等[97]首先开发了一个框架用于分析k最近邻（KNN）算法的对抗鲁棒性。该框架确定了具有不同鲁棒性的k的两个不同状态。假设常数为k的KNN在条件概率$p(y = 1|x)$位于(0, 1)的区域中，这样的模型在大样本限制条件下不具有鲁棒性。对于$k = \Omega(\sqrt{dn\lg n})$，其中，$d$是数据维度；$n$是样本大小，在大样本限制条件下，基于KNN的分类的鲁棒性区域接近贝叶斯最佳分类器的鲁棒性区域。由于$k = \Omega(\sqrt{dn\lg n})$对于具有高数据维度和大量样本的真实数据集而言太大，文献[97]提出了一种有效的1-最近邻算法。该算法中当反向标记的点相距较远时，最接近的1个邻居是鲁棒的。基于上述现象，该算法会删除附近的反向标记点，并保留与其邻居共享相同标记的点。在MNIST上，对于较小的对抗性扰动（低攻击半径），此算法与基于1-最近邻的分类相比其性能稍逊于其他防御（如对抗训练的分类器），而在较大攻击半径的情况下，其性能优于那些防御。Papernot等[98]提出了一种称为DkNN的基于KNN的防御方法，该方法是对DNN每一层学习的数据表示执行KNN算法。KNN算法主要用于估计测试输入的异常预测。当DNN学习的中间表示与那些和预测共享同一标签的训练样本的表示不接近时，则认为该预测是异常的。实验表明，在多次对抗攻击下，尤其是在C&W攻击下，DkNN可以显著提高DNN的准确性。

5.7. 基于贝叶斯模型的防御

Liu等[99]将贝叶斯神经网络（Bayesian neural network, BNN）[100]与对抗训练相结合，从而学习在对抗攻击下的最优模型的权重分布。具体来说，作者假设网络中的所有权重都是随机的，并使用BNN理论中常用的技术训练网络[100]。通过对抗性训练，这种随机的BNN [99]，与RSE [69]和CIFAR10以及STL10和ImageNet143的常见的对抗性训练相比，显著提高了对抗鲁棒性。Schott等[101]建议基于贝叶斯模型对输入数据的分类条件分布进行建模，并将新样本分类为相应类别条件模型产生最高似然性的类别。他们将模型命名为Analysis by Synthesis model (ABS)。ABS被称为MNIST数据集上针对L_0、L_2和L_∞攻击的第一个稳健模型。ABS在抵抗L_0和L_2攻击方面达到了最先进的性能，但在L_∞的攻击下其性能要比PGD对抗训练的模型稍差。

5.8. 基于一致性的防御

对于诸如音频识别和图像语义分割之类的机器学习任务，可以应用一致性信息来区分良性样本和对抗性样本。Xiao等[102]发现对于语义分割任务，对抗性干扰一个像素也会影响其周围像素的预测。因此对单个区块进行干扰也会破坏其附近元素之间的空间一致性。这种一致性信息可以用于区分良性和对抗性图像。作者用自适应攻击方式对这种基于一致性的防御进行了评估，最终证明了这种防御比其他异常检测系统具有更好的性能。对于音频识别任务，Yang等[103]探索了音频信号的时间一致性，并且还发现对抗性扰动将破坏时间一致性。对于对抗性音频信号，信号的一部分平移与整个信号的平移不一致。文献[103]显示基于一致性测试的检测可以在对抗性音频信号上实现90%以上的检测率。

6. 讨论

6.1. 白盒与黑盒攻击

从攻击者的角度来看，白盒攻击和黑盒攻击的主要区别在于它们对目标模型的访问权限。在白盒攻击中，攻击者可以访问模型的结构和权重，以便他们可以通过文献[17]中的方法计算真实的模型梯度或近似梯度，此外攻击者还可以根据防御方法和参数调整其攻击方法。在这种情况下，以前引入的大多数启发式防御实际上无法抵御这种强大的自适应攻击者。在黑盒攻击中，模型结构和权重不会被攻击者知道，在这种情况下，为了使用上述基于梯度的攻击算法，对手必须从有限的信息中推断出模型的梯度。在没有任何特定模型信息的情况下，对模型梯度的无偏估计就是对具有不同随机种子的一组预训练模型梯度的期望。文献[18]使用基于动量梯度的方法进行此梯度估计，并在NIPS2017挑战赛（在黑盒设置下）中获得第一名。Chen等[104]研究了另一种黑盒攻击方法，该方法可以向攻击者授予额外的查询访问权限。因此如果给定精心设计的输入，攻击者可以从目标模型的输出推断出梯度。在这种设置下，可以应用零阶方法来更好地估计模型梯度[104]。但是此方法的缺点是需要进行大量的查询操作，其查询量与数据维度成比例。

6.2. 对抗攻击与防御研究趋势之间的差异

对抗攻击的研究趋势主要包括两个方向。第一个是设计更有效、更强大的攻击用来评估新兴的防御系统，这个方向的重要性很直观，因为我们希望在潜在对手面前预先了解所有的风险。第二个是实现物理世界中的对抗攻击。以前对该研究主题的主要疑问是那些对抗性攻击是否会对物理世界形成真正的威胁。一些研究人员怀疑由于某些环境因素的影响，最初在数字空间中设计的对抗性攻击将无效。Kurakin等[6]首先通过使用模型梯度相对于输入的期望值并加上环境因素引起的随机噪声来实现物理世界中的对抗攻击。Eykholt等[33]进一步考虑了掩模和制造误差，从而实现了交通标志的对抗性扰动。最近Cao等[105]成功生成的对抗目标可以用来欺骗基于激光雷达的检测系统，这些都验证了物理对抗样本的存在。在防御方面，由于大多数启发式防御都无法防御自适应白盒攻击，因此研究者开始关注可证明的防御，这种防御是指无论攻击者采用哪种攻击方式，可证明防御都可以在一定程度下保证防御的性能。但是到目前为止，可扩展性是目前大多数可证明防御所普遍具有的问题。例如，区间界分析是最近流行的证明式防御方法，但是它不能扩展到非常深的神经网络和大型数据集。由此可见，与攻击相比，防御系统的发展面临着更多的挑战。这主要是因为一次攻击只能针对一类防御，所以防御机理急需被证明，这样某种防御在某些情况下对所有可能的攻击才能都有效。

6.3. 模型鲁棒性分析的最新进展

由于DNN具有复杂的非凸性，理论上很难对其进行分析，因此人们开始分析一些简单的ML模型，如KNN和线性分类器的鲁棒性。Wang等[97]指出KNN的稳健性在很大程度上依赖于参数k、数据维数d和数据大小n。k必须非常大才能保证KNN像贝叶斯优化分类器一样具有渐近稳健性。Fawzi等[106]分析了线性和二阶分类器稳健性的理论框架，其中模型稳健性由能够引起样本类别变化的扰动范数平均值来定义。在适用于包括线性和二阶在内的大量分类器假设下，模型鲁棒性的上界得到了证明。研究表明，与均匀随机噪声的鲁棒性相比，对抗鲁棒性的尺度为$O(\sqrt{1/d})$。最近，MLP、CNN和ResNet的鲁棒性也被广泛研究，研究者通过抽象的区间界分析试图在给定扰动大小的情况下逐层约束输出。我们在本次调查中没有详细分析，请感兴趣的读者参考文献[107–109]。

6.4. 未解决的主要挑战

（1）对抗样本背后的因果关系。虽然提出了许多对抗攻击方式，但是对抗样本的因果关系仍不清楚。早期对这一问题的研究将对抗样本的出现归功于模型结构和学习方法，研究者认为适当的策略和网络结构将显著提高对抗样本的鲁棒性。研究者沿着这种思路尝试过一些探索，特别是与产生模糊梯度相关的研究，然而实际上这可能是一种不太合理的研究方向[2]。相反，最近的研究发现对抗性的出现更可能是高维数据几何[110–112]和训练数据不足[113]的结果。具体来说，文献[110–112]证明了对抗性扰动在几个概念验证数据集（如$\{0, 1\}^n$和同心n维球体）上按比例$O(\sqrt{1/d})$放缩，其中d是数据维度。Ludwig等[113]表明对抗性强的任务比普通的ML任务需要更多的数据，并且所需的数据大小可能以比例$O(\sqrt{1/d})$放缩。

（2）通用鲁棒决策边界的存在。由于在不同度量标准下定义了许多对抗攻击方法，一个自然的问题是是否存在由特定训练策略的某种DNN来学习的通用鲁棒决策边界。当前，这个问题的答案是“否”。尽管PGD对抗训练对各种L_∞攻击表现出显著的抵抗力，但文献[59]表明它仍然容易受到其他L_p范数的对抗攻击，如EAD和CW_2。Khoury等[111]证明了2-同心球面数据集的最优L_2和L_∞决策边界是不同的，它们的差异随着数据集的共维（即数据流形的维数与整个数据空间的维数之差）而增大。

（3）有效防御白盒攻击。我们仍然没有看到一种能够很好地平衡效果和效率的防御。在有效性方面，对抗性训练表现出最好的性能，但计算成本很高。在效率方面，许多基于随机和去噪的防御/检测系统的配置只需几秒钟。然而，最近的许多论文[17,84,114,115]表明这些防御方法并没有他们声称的那样有效。这些研究可证明防御理论为实现对抗防御指明了一条道路，但其准确性和有效性都远远不能满足实际要求。

7. 结论

本文综述了近年来具有代表性的对抗攻击与对抗防御方法。我们研究了这些方法的思想，还根据最新进展对对抗防御的有效性进行了评论。我们的结论是：近两年来新的对抗攻击和防御迅速发展，同时针对对抗样本的因果关系、一般鲁棒边界的存在等基本问题还需要深入研究。此外我们还没有看到一种有效的对抗防御方

法，目前最有效的防御是对抗性训练，但其在实际部署中计算成本太高。许多启发式防御都声称其是有效的，但这类防御目前还不能抵御自适应性白盒攻击者的攻击。简而言之，要达到有效防御的目标似乎还有很长的路要走。

致谢

本工作得到了“浙江大学-蚂蚁金服数据安全实验室”的支持。

Compliance with ethics guidelines

Kui Ren, Tianhang Zheng, Zhan Qin, and Xue Liu declare that they have no conflict of interest or financial conflicts to disclose.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.eng.2019.12.012.

References

[1] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Proceedings of the 26th Conference on Neural Information Processing Systems; 2012 Dec 3–6; Nevada, USA; 2012. p. 1097–105.
[2] Cho K, van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. 2014. arXiv:1406.1078.
[3] Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, et al. Mastering the game of Go with deep neural networks and tree search. Nature 2016;529(7587):484–9.
[4] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, et al. Intriguing properties of neural networks. 2013. arXiv:1312.6199.
[5] Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. 2014. arXiv:1412.6572.
[6] Kurakin A, Goodfellow I, Bengio S. Adversarial examples in the physical world. 2016. arXiv:1607.02533.
[7] Zheng T, Chen C, Ren K. Distributionally adversarial attack. 2018. arXiv:1808.05537.
[8] Carlini N, Wagner D. Towards evaluating the robustness of neural networks. In: Proceedings of the 2017 IEEE Symposium on Security and Privacy; 2017 May 22–26; San Jose, CA, USA; 2017. p. 39–57.
[9] Papernot N, McDaniel P, Jha S, Fredrikson M, Celik ZB, Swami A. The limitations of deep learning in adversarial settings. In: Proceedings of the 2016 IEEE European Symposium on Security and Privacy; 2016 Mar 21–24; Saarbrucken, Germany; 2016. p. 372–87.
[10] Moosavi-Dezfooli SM, Fawzi A, Frossard P. DeepFool: a simple and accurate method to fool deep neural networks. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 27–30; Las Vegas, NV, USA; 2016. p. 2574–82.
[11] Papernot N, McDaniel P, Goodfellow I. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. 2016. arXiv:1605.07277.
[12] Liu Y, Chen X, Liu C, Song D. Delving into transferable adversarial examples and black-box attacks. 2016. arXiv:1611.02770.
[13] Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. 2017. arXiv: 1706.06083.
[14] Xie C, Wu Y, van der Maaten L, Yuille A, He K. Feature denoising for improving adversarial robustness. 2018. arXiv:1812.03411.
[15] Zheng T, Chen C, Yuan J, Li B, Ren K. PointCloud saliency maps. 2018. arXiv:1812.01687.
[16] Li J, Ji S, Du T, Li B, Wang T. TextBugger: generating adversarial text against real-world applications. 2018. arXiv:1812.05271.
[17] Athalye A, Carlini N, Wagner D. Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples. 2018. arXiv:1802.00420.
[18] Dong Y, Liao F, Pang T, Su H, Zhu J, Hu X, et al. Boosting adversarial attacks with momentum. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA; 2018. p. 9185–193.
[19] Chen PY, Sharma Y, Zhang H, Yi J, Hsieh CJ. EAD: elastic-net attacks to deep neural networks via adversarial examples. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence; 2018 Feb 2–7; New Orleans, LA, USA; 2018.
[20] Moosavi-Dezfooli SM, Fawzi A, Fawzi O, Frossard P. Universal adversarial perturbations. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition; 2017 Jul 21–26; Honolulu, HI, USA; 2017. p. 1765–73.
[21] Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, et al. Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM International Conference on Multimedia; 2014 Nov 3–7; Orlando, FL, USA; 2014. p. 675–8.
[22] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition; 2015 Jun 7–12; Boston, MA, USA; 2015. p. 1– 9.
[23] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. arXiv:1409.1556.
[24] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 27–30; Las Vegas, NV, USA; 2016. p. 770–8.
[25] Sharif M, Bhagavatula S, Bauer L, Reiter MK. Accessorize to a crime: real and stealthy attacks on state-of-the-art face recognition. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security; 2016 Oct 24–28; Vienna, Austria; 2016. p. 1528–40.
[26] Parkhi OM, Vedaldi A, Zisserman A. Deep face recognition. In: Proceedings of British Machine Vision Conference; 2017 Sep 7–10; Swansea, UK; 2015.
[27] Brown TB, Mané D, Roy A, Abadi M, Gilmer J. Adversarial patch. 2017. arXiv:1712.09665.
[28] Athalye A, Engstrom L, Ilya A, Kwok K. Synthesizing robust adversarial examples. 2017. arXiv:1707.07397.
[29] Liu Y, Ma S, Aafer Y, Lee WC, Zhai J, Wang W, et al. Trojaning attack on neural networks. In: Proceedings of Network and Distributed Systems Security Symposium; 2018 Feb 18–21; San Diego, CA,USA; 2018.
[30] Xiao C, Li B, Zhu JY, He W, Liu M, Song D. Generating adversarial examples with adversarial networks. 2018. arXiv:1801.02610.
[31] Song Y, Shu R, Kushman N, Ermon S. Constructing unrestricted adversarial examples with generative models. In: Proceedings of the 32nd Conference on Neural Information Processing Systems; 2018 Dec 3–8; Montréal, Canada; 2018. p. 8312–23.
[32] Odena A, Olah C, Shlens J. Conditional image synthesis with auxiliary classifier GANs. In: Proceedings of the 34th International Conference on Machine Learning; 2017 Aug 6–11; Sydney, Austrila; 2017. p. 2642–51.
[33] Eykholt K, Evtimov I, Fernandes E, Li B, Rahmati A, Xiao C, et al. Robust physical-world attacks on deep learning visual classification. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA; 2018. p. 1625–34.
[34] Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention; 2015 Oct 5–9; Munich, Germany; 2015. p. 234–41.
[35] Grundmann M, Kwatra V, Han M, Essa I. Efficient hierarchical graph-based video segmentation. In: Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition; 2010 Jun 13–18; San Francisco, CA, USA; 2010. p. 2141–8.
[36] Su H, Maji S, Kalogerakis E, Learned-Miller E. Multi-view convolutional neural networks for 3D shape recognition. In: Proceedings of the IEEE International Conference on Computer Vision; 2015 Dec 7–13; Santiago, Chile; 2015. p. 945–53.
[37] Qi CR, Su H, Mo K, Guibas LJ. Pointnet: deep learning on point sets for 3D classification and segmentation. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition; 2017 Jul 21–26; Honolulu, HI, USA; 2017. p. 652–60.
[38] Lee H, Pham P, Largman Y, Ng AY. Unsupervised feature learning for audio classification using convolutional deep belief networks. In: Proceedings of the 23rd Conference on Neural Information Processing Systems; 2009 Dec 7–10; Vancouver, Canada; 2009. p. 1096–104.
[39] Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, et al. Human-level control through deep reinforcement learning. Nature 2015;518 (7540):529–33.
[40] Xie C, Wang J, Zhang Z, Zhou Y, Xie L, Yuille A. Adversarial examples for semantic segmentation and object detection. In: Proceedings of the 2017 IEEE International Conference on Computer Vision; 2017 Oct 22–29; Venice,

Italy; 2017. p. 1369–78.
[41] Cisse M, Adi Y, Neverova N, Keshet J. Houdini: fooling deep structured prediction models. 2017. arXiv:1707.05373.
[42] Qi CR, Yi L, Su H, Guibas LJ. Pointnet++: deep hierarchical feature learning on point sets in a metric space. In: Proceedings of the 31st Conference on Neural Information Processing Systems; 2017 Dec 4–9; Long Beach, CA, USA; 2017. p. 5099–108.
[43] Wang Y, Sun Y, Liu Z, Sarma SE, Bronstein MM, Solomon JM. Dynamic graph CNN for learning on point clouds. 2018. arXiv:1801.07829.
[44] Xiang C, Qi CR, Li B. Generating 3D adversarial point clouds. 2018. arXiv:1809.07016.
[45] Liu D, Yu R, Su H. Extending adversarial attacks and defenses to deep 3D point cloud classifiers. 2019. arXiv:1901.03006.
[46] Xiao C, Yang D, Li B, Deng J, Liu M. MeshAdv: adversarial meshes for visual recognition. 2018. arXiv:1810.05206v2.
[47] Carlini N, Wagner D. Audio adversarial examples: targeted attacks on speechto- text. In: Proceedings of 2018 IEEE Security and Privacy Workshops; 2018 May 24; San Francisco, CA, USA; 2018. p. 1–7.
[48] Hannun A, Case C, Casper J, Catanzaro B, Diamos G, Elsen E, et al. Deep speech: scaling up end-to-end speech recognition. 2014. arXiv:1412.5567.
[49] Yakura H, Sakuma J. Robust audio adversarial example for a physical attack. 2018. arXiv:1810.11793.
[50] Liang B, Li H, Su M, Bian P, Li X, Shi W. Deep text classification can be fooled. 2017. arXiv:1704.08006.
[51] Huang S, Papernot N, Goodfellow I, Duan Y, Abbeel P. Adversarial attacks on neural network policies. 2017. arXiv:1702.02284.
[52] Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, et al. Playing Atari with deep reinforcement learning. 2013. arXiv:1312.5602.
[53] Mnih V, Badia AP, Mirza M, Graves A, Harley T, Lillicrap TP, et al. Asynchronous methods for deep reinforcement learning. In: Proceedings of the 33rd International Conference on Machine Learning; 2016 Jun 19–24; New York, NY, USA; 2016. p. 1928–37.
[54] Schulman J, Levine S, Moritz P, Jordan M, Abbeel P. Trust region policy optimization. In: Proceedings of the 32nd International Conference on Machine Learning; 2015 Jul 6–11; Lille, France; 2015. p. 1889–97.
[55] Behzadan V, Munir A. Vulnerability of deep reinforcement learning to policy induction attacks. In: Proceedings of the International Conference on Machine Learning and Data Mining in Pattern Recognition; 2017 Jul 15–20; New York, NY, USA; 2017. p. 262–75.
[56] Lin YC, Hong ZW, Liao YH, Shih ML, Liu MY, Sun M. Tactics of adversarial attack on deep reinforcement learning agents. 2017. arXiv:1703.06748.
[57] Carlini N, Katz G, Barrett C, Dill DL. Ground-truth adversarial examples. In: ICLR 2018 Conference; 2018 Apr 30; Vancouver, BC, Canada; 2018.
[58] Papernot N, Faghri F, Carlini N, Goodfellow I, Feinman R, Kurakin A, et al. Technical report on the CleverHans v2.1.0 adversarial examples library. 2016. arXiv:1610.00768v6.
[59] Sharma Y, Chen PY. Attacking the Madry defense model with L1-based adversarial examples. 2017. arXiv:1710.10733v4.
[60] Kurakin A, Goodfellow I, Bengio S. Adversarial machine learning at scale. 2016. arXiv: 1611.01236.
[61] Tramèr F, Kurakin A, Papernot N, Goodfellow I, Boneh D, McDaniel P. Ensemble adversarial training: attacks and defenses. 2017. arXiv:1705.07204.
[62] Kannan H, Kurakin A, Goodfellow I. Adversarial logit pairing. 2018. arXiv:1803.06373.
[63] Zheng S, Song Y, Leung T, Goodfellow I. Improving the robustness of deep neural networks via stability training. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 27–30; Las Vegas, NV, USA; 2016. p. 4480–8.
[64] Engstrom L, Ilyas A, Athalye A. Evaluating and understanding the robustness of adversarial logit pairing. 2018. arXiv: 1807.10272.
[65] Lee H, Han S, Lee J. Generative adversarial trainer: defense to adversarial perturbations with GAN. 2017. arXiv: 1705.03387.
[66] Liu X, Hsieh CJ. Rob-GAN: generator, discriminator, and adversarial attacker. 2018. arXiv:1807.10454v3.
[67] Xie C, Wang J, Zhang Z, Ren Z, Yuille A. Mitigating adversarial effects through randomization. 2017. arXiv: 1711.01991.
[68] Guo C, Rana M, Cisse M, van der Maaten L. Countering adversarial images using input transformations. 2017. arXiv: 1711.00117.
[69] Liu X, Cheng M, Zhang H, Hsieh CJ. Towards robust neural networks via random self-ensemble. In: Proceedings of the 2018 European Conference on Computer Vision; 2018 Sep 8–14; Munich, Germany; 2018. p. 369–85.
[70] Lecuyer M, Atlidakis V, Geambasu R, Hsu D, Jana S. Certified robustness to adversarial examples with differential privacy. 2018, arXiv:1802.03471v4.
[71] Dwork C, Lei J. Differential privacy and robust statistics. In: Proceedings of the 41st Annual ACM Symposium on Theory of Computing; 2009 May 31–Jun 2; Bethesda, MD, USA; 2009. p. 371–80.
[72] Li B, Chen C, Wang W, Carin L. Certified adversarial robustness with additive noise. 2018. arXiv: 1809.03113v6.
[73] Dhillon GS, Azizzadenesheli K, Lipton ZC, Bernstein J, Kossaifi J, Khanna A, et al. Stochastic activation pruning for robust adversarial defense. 2018. arXiv: 1803.01442.
[74] Luo T, Cai T, Zhang M, Chen S, Wang L. Random mask: towards robust convolutional neural networks. In: ICLR 2019 Conference; 2019 Apr 30; New Orleans, LA, USA; 2019.
[75] Xu W, Evans D, Qi Y. Feature squeezing: detecting adversarial examples in deep neural networks. 2017. arXiv: 1704.01155.
[76] Xu W, Evans D, Qi Y. Feature squeezing mitigates and detects Carlini/Wagner adversarial examples. 2017. arXiv: 1705.10686.
[77] Xu W, Evans D, Qi Y. Feature squeezing mitigates and detects Carlini/Wagner adversarial examples. 2017. arXiv: 1705.10686.
[78] Sharma Y, Chen PY. Bypassing feature squeezing by increasing adversary strength. 2018. arXiv:1803.09868.
[79] Samangouei P, Kabkab M, Chellappa R. Defense-GAN: protecting classifiers against adversarial attacks using generative models. 2018. arXiv:1805.06605.
[80] Shen S, Jin G, Gao K, Zhang Y. APE-GAN: adversarial perturbation elimination with GAN. 2017. arXiv: 1707.05474.
[81] Carlini N, Wagner D. MagNet and "efficient defenses against adversarial attacks" are not robust to adversarial examples. 2017. arXiv:1711.08478.
[82] Meng D, Chen H. MagNet: a two-pronged defense against adversarial examples. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security; 2017 Oct 30–Nov 3; New York, NY, USA; 2017. p. 135–47.
[83] Liao F, Liang M, Dong Y, Pang T, Hu X, Zhu J. Defense against adversarial attacks using high-level representation guided denoiser. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA; 2018. p. 1778–87.
[84] Athalye A, Carlini N. On the robustness of the CVPR 2018 white-box adversarial example defenses. 2018. arXiv:1804.03286.
[85] Raghunathan A, Steinhardt J, Liang P. Certified defenses against adversarial examples. 2018. arXiv:1801.09344.
[86] Raghunathan A, Steinhardt J, Liang P. Semidefinite relaxations for certifying robustness to adversarial examples. In: Proceedings of the 32nd Conference on Neural Information Processing Systems; 2018 Dec 3–8; Montréal, Canada; 2018. p. 10877–87.
[87] Wong E, Kolter JZ. Provable defenses against adversarial examples via the convex outer adversarial polytope. In: Proceedings of the 31st Conference on Neural Information Processing Systems; 2017 Dec 4–9; Long Beach, CA, USA; 2017.
[88] Wong E, Schmidt FR, Metzen JH, Kolter JZ. Scaling provable adversarial defenses. 2018. arXiv:1805.12514.
[89] Sinha A, Namkoong H, Duchi J. Certifying some distributional robustness with principled adversarial training. 2017. arXiv:1710.10571.
[90] Namkoong H, Duchi JC. Stochastic gradient methods for distributionally robust optimization with f-divergences. In: Proceedings of the 30th Conference on Neural Information Processing Systems; 2016 Dec 5–10; Barcelona, Spain; 2016. p. 2208–16.
[91] Gao R, Kleywegt AJ. Distributionally robust stochastic optimization with Wasserstein distance. 2016. arXiv:1604.02199.
[92] Guo Y, Zhang C, Zhang C, Chen Y. Sparse DNNs with improved adversarial robustness. In: Proceedings of the 32nd Conference on Neural Information Processing Systems; 2018 Dec 3–8; Montréal, Canada; 2018. p. 242–51.
[93] Hein M, Andriushchenko M. Formal guarantees on the robustness of a classifier against adversarial manipulation. In: Proceedings of the 31st Conference on Neural Information Processing Systems; 2017 Dec 4–9; Long Beach, CA, USA; 2017. p. 2266–76.
[94] Weng TW, Zhang H, Chen PY, Yi J, Su D, Gao Y, et al. Evaluating the robustness of neural networks: an extreme value theory approach. 2018. arXiv:1801.10578.
[95] Xiao KY, Tjeng V, Shafiullah NM, Madry A. Training for faster adversarial robustness verification via inducing ReLU stability. 2018. arXiv:1809.03008.
[96] Katz G, Barrett C, Dill DL, Julian K, Kochenderfer MJ. Reluplex: an efficient SMT solver for verifying deep neural networks. In: Proceedings of the International Conference on Computer Aided Verification; 2017 Jul 24–28; Heidelberg, Germany; 2017. p. 97–117.
[97] Wang Y, Jha S, Chaudhuri K. Analyzing the robustness of nearest neighbors to adversarial examples. 2017. arXiv: 1706.03922.
[98] Papernot N, McDaniel P. Deep k-nearest neighbors: towards confident, interpretable and robust deep learning. 2018. arXiv:1803.04765.
[99] Liu X, Li Y, Wu C, Hsieh C. Adv-BNN: improved adversarial defense through robust Bayesian neural network. 2018. arXiv:1810.01279.
[100] Neal RM. Bayesian learning for neural networks. New York: Springer Science & Business Media; 2012.
[101] Schott L, Rauber J, Bethge M, Brendel W. Towards the first adversarially robust neural network model on MNIST. 2018. arXiv:1805.09190.
[102] Xiao C, Deng R, Li B, Yu F, Liu M, Song D. Characterizing adversarial examples based on spatial consistency information for semantic segmentation. In: Proceedings of the European Conference on Computer Vision; 2018 Sep 8–14; Munich, Germany; 2018. p. 217–34.
[103] Yang Z, Li B, Chen PY, Song D. Characterizing audio adversarial examples using temporal dependency. 2018. arXiv:1809.10875.
[104] Chen PY, Zhang H, Sharma Y, Yi J, Hsieh CJ. Zoo: zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security; 2017 Nov 3; Dalas, TX, USA; 2017. p. 15–26.
[105] Cao Y, Xiao C, Yang D, Fang J, Yang R, Liu M, et al. Adversarial objects against LiDAR-based autonomous driving systems. 2019. arXiv:1907.05418.
[106] Fawzi A, Fawzi O, Frossard P. Analysis of classifiers' robustness to adversarial perturbations. Mach Learn 2018;107(3):481–508.
[107] Mirman M, Gehr T, Vechev M. Differentiable abstract interpretation for provably robust neural networks. In: Proceedings of the 35th International

Conference on Machine Learning; 2018 Jul 10–15; Stockholm, Sweden; 2018. p. 3578–86.
[108] Singh G, Gehr T, Mirman M, Puschel M, Vechev M. Fast and effective robustness certification. In: Proceedings of the 32nd Conference on Neural Information Processing Systems; 2018 Dec 3–8; Montréal, Canada; 2018. p. 10802–13.
[109] Gowal S, Dvijotham K, Stanforth R, Bunel R, Qin C, Uesato J, et al. On the effectiveness of interval bound propagation for training verifiably robust models. 2018. arXiv:1810.12715.
[110] Dube S. High dimensional spaces, deep learning and adversarial examples. 2018. arXiv:1801.00634.
[111] Khoury M, Hadfield-Menell D. On the geometry of adversarial examples. 2018. arXiv:1811.00525.
[112] Gilmer J, Metz L, Faghri F, Schoenholz SS, Raghu M, Watterberg M, et al. Adversarial spheres. 2018. arXiv:1801.02774.
[113] Schmidt L, Santurkar S, Tsipras D, Talwar K, Madry A. Adversarially robust generalization requires more data. 2018. arXiv:1804.11285.
[114] Carlini N, Wagner D. Adversarial examples are not easily detected: bypassing ten detection methods. In: Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security; 2017 Nov 3; Dalas, TX, USA; 2017. p. 3–14.
[115] Carlini N. Is AmI (attacks meet interpretability) robust to adversarial examples? 2019. arXiv:1902.02322v1.