# 深度神经网络的对抗攻击及防御方法综述

赵宏，常有康，王伟杰

引用本文

赵宏，常有康，王伟杰.深度神经网络的对抗攻击及防御方法综述[J]．计算机科学, 2022, 49(11A): 210900163-11.

ZHAO Hong, CHANG You-kang, WANG Wei-jie. Survey of Adversarial Attacks and Defense Methods for Deep Neural Networks [J]. Computer Science, 2022, 49(11A): 210900163-11.

---

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210900163

# 深度神经网络的对抗攻击及防御方法综述

赵 宏 常有康 王伟杰
兰州理工大学计算机与通信学院 兰州 730050
(594286500@qq.com)

**摘 要** 深度神经网络正在引领人工智能新一轮的发展高潮,在多个领域取得了令人瞩目的成就。然而,有研究指出深度神经网络容易遭受对抗攻击的影响,导致深度神经网络输出错误的结果,其安全性引起了人们极大的关注。文中从深度神经网络安全性的角度综述了对抗攻击与防御方法的研究现状。首先,围绕深度神经网络的对抗攻击问题简述了相关概念及存在性解释;其次,从基于梯度的对抗攻击、基于优化的对抗攻击、基于迁移的对抗攻击、基于GAN的对抗攻击和基于决策边界的对抗攻击的角度介绍了对抗攻击方法,分析每种攻击方法的特点;再次,从基于数据预处理、增强深度神经网络模型的鲁棒性和检测对抗样本等3个方面阐述了对抗攻击的防御方法;然后,从语义分割、音频、文本识别、目标检测、人脸识别、强化学习等领域列举了对抗攻击与防御的实例;最后,通过对对抗攻击与防御方法的分析,展望了深度神经网络中对抗攻击和防御的发展趋势。

**关键词**:人工智能;深度神经网络;神经网络安全;对抗攻击;防御方法
**中图法分类号** TP391

## Survey of Adversarial Attacks and Defense Methods for Deep Neural Networks

ZHAO Hong,CHANG You-kang and WANG Wei-jie
School of Computer and Communication,Lanzhou University of Technology,Lanzhou 730050,China

**Abstract** Deep neural networks are leading a new round of high tide of artificial intelligence development,and have made remarkable achievements in many fields. However,recent studies have pointed out that deep neural networks are vulnerable to adversarial attacks,resulting in incorrect network outputs,and their security has attracted great attention. This paper summarizes the current state of research on adversarial attacks and defense methods from the perspective of deep neural network security. Firstly,it briefly describes the related concepts and existence explanations around the adversarial attacks of deep neural networks. Secondly,it introduces adversarial attacks from the perspectives of gradient-based adversarial attacks,optimization-based adversarial attacks,migration-based adversarial attacks,GAN-based adversarial attacks and decision boundary-based adversarial attacks,and analyses the characteristics of each adversarial attack method,analyzing the characteristics of each attack method. Again,the defense methods of adversarial attacks are explained from three aspects,including data-based pre-processing,enhancing the robustness of deep neural network models and detecting adversarial samples. Then,from the fields of semantic segmentation,audio,text recognition,target detection,face recognition,reinforcement learning,examples of adversarial attacks and defenses are listed. Finally,the development trend of adversarial attacks and defenses in deep neural networks is forcasted through the analysis of adversarial attacks and defense methods.

**Keywords** Artificial intelligence,Deep neural network,Neural network security,Adversarial attacks,Defense methods

## 1 引言

随着深度神经网络的快速发展,深度神经网络正在引领人工智能走向新一轮的高潮,其在许多领域都取得了瞩目的成就。如在自动驾驶[1-2]中,利用摄像头和激光雷达等传感器采集道路环境数据,将其输入深度神经网络中进行识别预测,在没有人工干预的情况下实现自动驾驶;在医学影像分析[3-4]中,深度神经网络可以识别、分类和量化医学图像,辅助医生进行疾病的快速诊断;在图像识别[5-7]中,由于深度神经网络强大的数据学习能力,可以在ImageNet数据集上大幅度提升图像识别的准确率;在网络分析[8]中,深度神经网络通过分析并识别数据流量中的异常信息,来实现快速的网络入侵检测。

深度神经网络之所以取得瞩目的成就,首先得益于计算机计算性能的大幅提升和训练规模的增加;其次在于深度神经网络的强特征提取能力,可以将输入的高维数据在经过特征提取后降为低维数据,对低维数据变换输出最终的结果。然而,在数据从高维空间到低维空间的过程中,关键信息被保留而其他次要信息被忽略,如果对关键信息加以扰动,可能会

基金项目:国家自然科学基金(62166025);甘肃省重点研发计划(21YF5GA073);甘肃省优秀研究生"创新之星"(2021CXZX-511,2021CXZX-512)
This work was supported by the National Natural Science Foundation of China(62166025);Science and Technology Project of Gansu Province(21YF5GA073) and Outstanding Postgraduate Student "Innovation Star" in Gansu Province (2021CXZX-511,2021CXZX-512).
通信作者:常有康(2507576651@qq.com)

影响深度神经网络的正常输出[9]。

Szegedy 等[10]首先指出在干净图像中添加一些细微、精心设计、人类视觉不易观察的扰动数据后，致使神经网络分类错误。其中，添加扰动的样本称为对抗样本。随后，Goodfellow 等[11]提出线性假设来解释对抗样本的存在。Kurakin 等[12]指出，对抗样本亦存在于现实世界中，利用摄像头和其他传感器将获得的自然图像输入到 Inception_v3 分类网络中，结果显示网络遭受到了对抗样本的攻击。Joshi 等[13]提出了一种新的攻击方法，其没有在原始图像中添加扰动，而是改变了图像内容的特征属性，生成难以辨别的攻击图像，成功地欺骗了分类网络。Fan 等[14]提出了一种稀疏对抗攻击方法，将每个像素的扰动表示为扰动幅度和二进制选择因子的乘积，联合优化二进制选择因子和像素的连续扰动幅度，对像素添加扰动生成对抗样本。

对抗攻击问题自发现以来就引起了人们的广泛关注，研究人员先后提出了多种方法进行对抗攻击防御[15-19]。现阶段主要的防御方法主要有 3 类。

(1)数据预处理，如对抗样本去噪[20-21]和数据压缩[22-23]等。这些方法计算速度快，不需要修改模型的网络结构，缺点在于修改输入样本时，会丢失样本的高频信息，使得网络模型无法提取正确的特征区域，从而使神经网络分类错误。

(2)增强深度神经网络的鲁棒性，如对抗训练[24-25]、防御蒸馏方法[26]、生物启发防御方法[27]和深度压缩网络[28]等。这种增强深度神经网络鲁棒性的方法在一定程度上提高了网络模型的随机性和网络的认知性能，然而若对特定网络进行特定攻击时，其防御效率会大幅下降。

(3)检测对抗样本，如基于 Generative Adversarial Network(GAN)[29]和基于 MagNet[30]等。这类方法具有很好的泛化能力，尤其对黑盒和灰盒攻击具有良好的防御能力，然而在白盒攻击的情况下其性能会大幅度下降。

本文从深度神经网络的安全性出发，综述了对抗攻击及防御的研究现状。本文第 2 节介绍了对抗攻击的相关概念及对抗攻击方法；第 3 节介绍了对抗防御方法；第 4 节简述了对抗攻击与防御实例；第 5 节对对抗攻击与防御的研究趋势进行了展望；最后总结全文。

## 2 对抗攻击

### 2.1 相关概念

本节将介绍对抗攻击的相关概念，包括对抗样本的定义和存在性解释。

#### 2.1.1 对抗样本

对抗样本指在干净图像中添加精心制作、微小的扰动，人类视觉系统难以察觉这些扰动。若将其输入深度神经网络，神经网络会以较高的概率输出错误的分类结果，如式(1)所示：

$$F(x_{adv})=y_{adv}\neq y \tag{1}$$

其中，$F(\cdot)$表示深度神经网络；$x_{adv}$表示对抗样本；$y_{adv}$表示对抗样本的输出结果；$y$表示真实标签。

#### 2.1.2 对抗样本存在性解释

众多学者就对抗样本存在的原因提出了不同的假说，本文主要从非线性假说、线性假说、边界倾斜假说 3 个角度解释对抗样本存在的原因。

(1)非线性假说

Szegedy 等[10]指出，对抗样本的存在是因为深度神经网络的高度非线性输出。由于数据存在于概率较低的盲区中，数据采样时不能有效地覆盖这些区域，因此在测试时，神经网络不能有效地处理来自盲区的数据样本，导致神经网络的泛化能力较差，出现对抗攻击现象。盲区中的数据如图 1 中的红色圆圈位置所示。

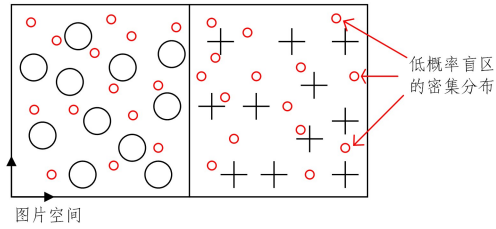


图 1 非线性假说中对抗样本存在的区域(电子版为彩图)

Fig. 1 Regions of adversarial example presence in non-linear hypothesis

(2)线性假说

Goodfellow 等[11]认为，深度神经网络的高度线性以及监督学习中模型正则化导致的过拟合产生了对抗样本。由于输入深度神经网络的数据维度较高，添加微小的扰动后输入线性模型，该模型明显地表现出了较低的鲁棒性。分析原因发现，对抗样本中微小的扰动经过深度神经网络的层层特征提取，在线性的作用下通过点乘叠加的方式放大微小的扰动，使得深度神经网络的分类结果错误。

(3)边界倾斜假说

Tanay 等[31]认为，线性假说存在局限性，即使输入样本的维度较高，也并不是所有的线性模型都会产生对抗攻击现象。他们认为，虽然在简单的 logistic 回归函数中存在对抗攻击现象，但简单的线性模型与深度神经网络生成的对抗样本存在区别，前者产生低频扰动，后者产生难以察觉的高频扰动。

基于以上结论，提出了边界倾斜假说来解释对抗样本。该假说认为，对抗样本存在于采样数据的分类边界，如图 2 所示，红色表示样本的实际分布边界，黑色表示采样数据的分类边界。实际分布边界与采样数据分布边界非常接近但不重合，由于分布边界无法保持一致，因此对抗样本可能存在于它们之间的夹缝之中，以图 2 为例，规定红色平面以上的样本为○，以下为×，将图中的○在加入微小的扰动之后，新样本穿过夹缝进入红色面以下，样本就会被分类为×。这也解释了线性和非线性模型都存在对抗样本的原因。

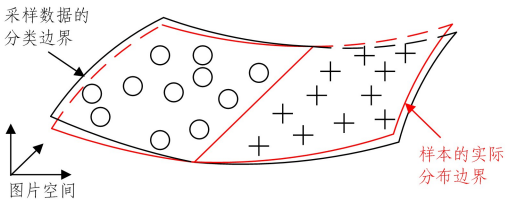


图 2 边界倾斜假说示意图

Fig. 2 Illustration of boundary tilt hypothesis

除了上述解释，Gilmer 等[32]观察到正确分类的干净样本与错误分类的对抗样本视觉效果很接近，认为神经网络遭受对抗攻击与训练过程无关，这是深度学习一个固有的弱点；

Madry 等[33]认为，神经网络在训练时没有大规模的数据集，不能训练出鲁棒性较强的分类神经网络，因此导致神经网络容易遭受对抗样本的攻击。

### 2.2　对抗攻击方法

本文根据生成特征的不同，可以将对抗攻击方法分为如下几类：

(1)基于梯度的攻击方法，利用干净图像的梯度信息生成对抗样本；

(2)基于优化的攻击方法，利用优化目标函数生成对抗样本；

(3)基于迁移的攻击方法，利用对抗攻击之间的迁移性生成对抗样本；

(4)基于 GAN 的攻击方法，利用 GAN 网络生成深度神经网络难以区分的对抗样本；

(5)基于决策边界的攻击方法，利用差分进化算法，以迭代的方式生成最佳的对抗样本。

#### 2.2.1　基于梯度的攻击方法

基于梯度的攻击方法通过反向传播算法为输入图像生成一个扰动向量。在通过网络进行反向传播时，其认为模型的参数是常数，而输入的数据是变量，因此可以计算输入图像中每个像素相对应的梯度信息，利用梯度信息获得扰动向量，生成对抗样本，并且生成的对抗样本与输入图像非常相似。

(1)FGSM[11]方法。首先利用损失函数计算干净图像的最大损失值，其次将损失值传给干净图像并计算梯度值，最后使用符号函数 sign()计算梯度方向，如式(2)所示：

$$x_{\text{adv}}=x+\varepsilon\,\text{sign}(\nabla_x J(x,y)) \tag{2}$$

其中，$x_{\text{adv}}$表示对抗样本；$\nabla_x$表示干净图像 $x$ 的梯度信息；$J$ 表示交叉熵损失函数；$y$ 表示干净图像 $x$ 的正确分类结果；$\varepsilon$ 表示对抗扰动的强度。

FGSM 方法不需要改变网络的模型结构和输入图像中特定的像素点，重点关注梯度扰动的方向，计算每张输入图像的梯度以改变图像的结构。

(2)JSMA(Jacobian-based Saliency Map Attack)[34]方法。2016 年，Papernot 等针对非循环深度神经网络提出了基于梯度的目标攻击方法——雅克比矩阵显著图攻击方法。该方法计算深度神经网络的前向导数构造雅克比矩阵，根据雅克比矩阵计算对抗性显著图，对输入图像添加扰动，生成对抗样本。前向导数表示输入图像的特征对分类结果的影响程度。

雅克比矩阵如式(3)所示：

$$J_F(x)=\frac{\partial F(X)}{\partial X}=\left[\frac{\partial F_j(X)}{\partial x_i}\right]_{i\in 1,\cdots M;j\in 1,\cdots,N} \tag{3}$$

神经网络计算每个特征的偏导数，如式(4)所示：

$$\frac{\partial F_j(X)}{\partial x_i}=\left(W_{n+1,j}\cdot\frac{\partial H_n}{\partial x_i}\right)\times\frac{\partial f_{n+1,j}}{\partial x_i}(W_{n+1,j}\cdot H_n+b_{n+1,j}) \tag{4}$$

其中，$F(\cdot)$表示神经网络；$\boldsymbol{X}$ 表示干净样本；$x_i$ 表示 $\boldsymbol{X}$ 的第 $i$ 个分量；$\boldsymbol{H}_n$ 为第 $n$ 层神经元的输出向量；$f_{n+1,j}$ 为第 $n+1$ 层输出神经元 $j$ 的激活函数；$\boldsymbol{W}$ 为权重向量；$\boldsymbol{b}$ 为偏置。

JSMA 方法只需要修改少量的像素点就可以使输入样本分类错误。该方法在样本量较小的 MNIST[35]数据集中取得了很好的攻击效果，但其是否适用于其他数据集还需要进一步研究。

(3)BIM(Basic Iterative Method)[12]方法。BIM 算法沿着梯度的方向，将扰动多次逐步添加到输入图像中，在每一次迭代之后重新计算梯度方向，如式(5)所示：

$$\begin{aligned}&x_{\text{adv}}^0=x\\&x_{\text{adv}}^{n+1}=clip_{\varepsilon,x}(x_{\text{adv}}^n+\alpha\cdot\text{sign}(\nabla_x J(x_{\text{adv}}^n,y)))\end{aligned} \tag{5}$$

其中，$clip()$是一个裁剪函数，它保证攻击样本点的像素在原始样本像素的邻域内，使得图像不会失真；$\alpha$ 为迭代步长。

BIM 方法是基于 FGSM 的迭代攻击方法，每次迭代将像素值改变 $\alpha$，弥补了 FGSM 单步攻击方法在面对白盒攻击时效率较低的缺点，然而由于倾向过拟合神经网络模型，因此在面对黑盒攻击时，防御效率有所下降。

(4) MI-FGSM(Momentum Iterative Fast Gradient Sign Method)[36]方法。MI-FGSM 方法是基于 BIM 的迭代攻击方法。在对抗样本的生成过程中，Dong 等将动量方法添加到迭代过程中，如式(6)、式(7)所示：

$$g_0=0,g_{t+1}=\mu g_t+\frac{\nabla_x J(x_t^*,y)}{\|\nabla_x J(x_t^*,y)\|_1} \tag{6}$$

$$x_t^*=x_{t-1}+clip_{\varepsilon,x}\{\alpha\cdot sign(g_{t+1})\} \tag{7}$$

其中，$x^*$ 表示对抗样本；$\mu$ 表示动量的衰减因子；$g_t$ 表示第 $t$ 步迭代时，模型对输入图像的梯度；$J$ 表示损失函数；$\nabla_x$表示梯度。

MI-FGSM 方法保证梯度的更新方向稳定，加速梯度的收敛和下降，并且在迭代的过程中避免局部最大值。

(5) DI$^2$ FGSM(Diverse-Input-Iterative FGSM) 和 MDI$^2$-FGSM(Momentum-Diverse-Input-Iterative FGSM)[37]方法。基于单步的攻击方法会产生欠拟合现象，而基于迭代的攻击方法会产生过拟合现象。为解决此问题，研究人员分别基于 BIM 和 MI-FGSM 方法，提出了 DI$^2$ FGSM 和 MDI$^2$-FGSM 方法来提高对抗攻击的能力。DI$^2$ FGSM 在生成对抗样本的过程中以概率 $p$ 对图像进行随机变换，如式(8)所示：

$$\begin{aligned}&x_{\text{adv}}^0=x\\&x_{\text{adv}}^{n+1}=clip_{\varepsilon,x}(x_{\text{adv}}^n+\alpha\cdot\text{sign}(\nabla_x J(T(x_{\text{adv}}^n;p),y_{\text{true}})))\end{aligned} \tag{8}$$

其中，$n$ 表示迭代次数；$clip_{\varepsilon,x}$表示在 $\varepsilon$ 邻域内，对输入图像 $x$ 进行裁剪；$\alpha$ 表示步长；$J$ 表示损失函数；$T(\cdot)$表示 random resize 和 random padding 变换。

MDI$^2$-FGSM 方法通过在 DI$^2$ FGSM 方法中加入动量，避免局部最大值，提升攻击效率，如式(9)所示：

$$\begin{aligned}&g_0=0\\&g_{t+1}=\mu g_t+\frac{\nabla_x J(T(x_t^*;p),y)}{\|\nabla_x J(T(x_t^*;p),y)\|_1}\end{aligned} \tag{9}$$

其中，$\mu$ 表示动量的衰减因子；$g_t$ 表示步骤 $t$ 的累积梯度。

FGSM，I-FGSM，MI-FGSM，DI$^2$ FGSM 和 MDI$^2$ FGSM 之间的关系如图 3 所示。

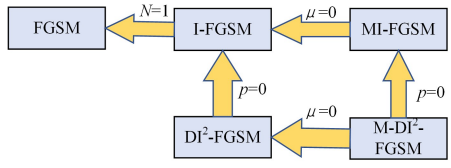


图 3　FGSM 攻击之间的关系

Fig. 3　Relationship between FGSM attacks

在 FGSM 系列方法中，通过转换概率 $p$ 来权衡白盒模型和黑盒模型的成功率。当 $p=0$ 时，DI$^2$ FGSM 退化为 I-FGSM，MDI$^2$ FGSM 退化为 MI-FGSM，导致模型过拟合；当

$p=1$ 时，$DI^2$FGSM 和 $MDI^2$FGSM 方法对黑盒模型具有较高的成功率，然而对白盒模型的成功率较低。通过控制衰减因子 $\mu$ 决定是否将动量加入到迭代的过程中，当 $\mu=0$ 时，$MDI^2$FGSM 退化为 $DI^2$FGSM，MI-FGSM 退化为 I-FGSM；此外，当迭代系数 $N=1$ 时，I-FGSM 退化为 FGSM，降低了攻击效率。

(6) AdvFlow 方法[38]。2020 年，Dolatabadi 等提出了 AdvFlow 方法。利用标准化流(Normalizing Flows，NF)生成模型对对抗样本的数据集进行概率分布建模，如式(10)所示：

$$X'=proj_s(f(z)),z\sim N(z|\mu,\sigma^2 I) \tag{10}$$

其中，$f$ 表示标准化流模型，满足可逆和可微的性质；$proj_s$ 表示映射规则，将对抗样本限制在对抗域中；$z$ 表示均值为 $\mu$、方差为 $\sigma^2$ 的正态分布。标准化流模型 $f(z)$ 的分布与干净样本的分布十分接近，生成的对抗样本分布与干净样本的分布也十分接近，因此很难检测出对抗样本。

(7) SVD-Universal 方法[39]。2020 年，Sanddesh 等提出了基于奇异值分解的泛用对抗攻击方法。该方法网络结构简单，通过对神经网络进行少量查询访问，使用很少的测试集生成泛用对抗攻击样本。SVD-Universal 方法结合了 Gradient，FGSM 和 DeepFool 梯度方向，分别给出了 SVD-Gradient，SVD-FGSM 和 SVD-DeepFool 泛用对抗攻击方法，生成了 3 种不同类型的对抗样本，将对抗样本进行归一化处理，然后组合为矩阵，对矩阵进行奇异值分解，寻找它们的相似性，以此获取攻击能力更强的泛用对抗样本。

(8) Simulator Attack[40]方法。针对当前黑盒攻击查询复杂度高，并且容易被防御和检测的问题。Ma 等将对抗攻击与元学习方法结合，提出了一种模拟器攻击方法，该方法通过收集对抗攻击过程中产生的查询序列，以多任务的形式构建训练数据，然后训练一个通用替代模型模拟器，其可以模拟任何未知目标模型的输出；其次，得益于元学习具有较强的泛化能力和鲁棒性，在元学习学习的过程中，使用均方误差损失函数最小化模拟器和网络输出之间的距离，从多个网络模型中计算和累积损失的元梯度，更新模拟器的参数以提高泛化能力；在测试时，经过训练的模拟器使用查询数据微调参数，精确模拟未知目标网络的输出。由于查询目标模型的过程被转移到模拟器中，因此降低了对抗攻击的查询复杂度。

### 2.2.2 基于优化的攻击方法

基于优化的攻击方法是通过优化目标函数，保证原始图像与对抗样本的距离最小的情况下，寻找添加的最小扰动，生成对抗样本。

(1) C&W (Carlini and Wagner)[19] 方法。该方法使用 Adam-Optimizer 优化器生成对抗样本。首先使用变量 $\omega_n$ 计算最小扰动 $r_n$，并利用 tanh() 函数将对抗样本的像素范围映射到 $(-\infty,+\infty)$，以便于目标函数的优化，如式(11)所示：

$$r_n=\frac{1}{2}(\tanh(\omega_n)+1)-x \tag{11}$$

其次，基于 $\omega_n$，C&W 算法损失函数如式(12)所示：

$$\min_{\omega_n}\| r_n \|+c\cdot f(\frac{1}{2}(\tanh(\omega_n)+1)) \tag{12}$$

其中：

$$f(x_{\text{adv}})=\max(\max\{Z(x_{adv};i\neq t\}-Z(x_{\text{adv}})_t,-k) \tag{13}$$

其中，$c$ 表示超参数，用于平衡 min 和 max 两个损失函数的相对关系；$Z(x_{\text{adv}})$ 表示样本 $x_{\text{adv}}$ 的输出向量；$k$ 表示置信度，$k$ 值越大说明模型分类错误的概率越大。

C&W 攻击在面对防御性能较强的防御蒸馏网络时，可以成功攻击深度神经网络；并且由于 C&W 方法加入的扰动较小，视觉效果比基于梯度生成的对抗样本更加自然。

(2) Attacking CNN+RNN[41]。2020 年，Xu 等研究了具有 CNN 和 RNN 结构的对抗攻击。该方法针对序列识别任务中的场景文本识别和图像字幕，通过显式捕获序列中的时间依赖性计算每个标签的梯度，从而在数值约束下迭代生成扰动。该方法通过较大的学习率和更多的训练次数，以较少的迭代次数提高攻击成功率。

### 2.2.3 基于迁移的攻击方法

由于黑盒攻击无法获得神经网络的内部结构、参数信息和模型的梯度，因此针对黑盒攻击通常是使用近似梯度生成对抗样本，近似梯度由替代模型计算得出，然后迁移至目标模型，生成对抗样本。

(1) ZOO(Zeroth Order Optimization)[42]方法。2017 年，Chen 等基于 C&W 方法提出了 ZOO 方法。与 C&W 方法不同的是，ZOO 的损失函数如式(14)所示：

$$f(x,t)=\max\{\max_{i\neq t}\log[F(x)]_i-\log[F(x)]_t,-k\} \tag{14}$$

其中，$[F(x)]_i$ 表示神经网络将样本 $x$ 分类为 $i$ 的概率；$k$ 表示调整参数。使用 log(·)函数的优点在于它是一个单调函数，对于任何 $x,y\geqslant 0$，当且仅当 $x\geqslant y$ 时，有 $\log(x)\geqslant\log(y)$，因此当满足式(15)时，分类神经网络将样本分类为 $t$ 的概率最大。

$$\max_{i\neq t}\log[F(x)]_i-\log[F(x)]_t\leqslant 0 \tag{15}$$

由于黑盒攻击无法获得模型的梯度信息，通过使用对称差分(Symmetric Difference Quotient)和黑塞(Hessian)矩阵估计黑盒模型的梯度，其次使用零阶优化方法求解梯度。对称差分的计算如式(16)所示：

$$\hat{g}_i:=\frac{\partial f(x)}{\partial x_i}\approx\frac{f(x+he_i)-f(x-he_i)}{2h} \tag{16}$$

对损失函数进行二次估计得到坐标方向的 Hessian 估计，如式(17)所示。

$$\hat{h}_i:=\frac{\partial^2 f(x)}{\partial x_{ii}^2}\approx\frac{f(x+he_i)-2f(x)+f(x-he_i)}{h^2} \tag{17}$$

其中，$\frac{\partial f(x)}{\partial x_i}$ 表示梯度估计；$x$ 表示对抗样本；$f$ 表示损失函数；$h$ 表示值为 0.0001 的常量；$e_i$ 表示基向量，第 $i$ 个分量值为 1。

ZOO 攻击使用零阶优化方法，解决了黑盒攻击不能获取模型梯度信息的问题。该方法的计算量与输入样本的大小成正比，若输入样本的尺寸较大，则需要大量的计算进行梯度估计，为解决上述问题，ZOO 攻击使用降维的方法生成对抗样本，以减小计算量，并且通过 Adam[43]优化算法加速梯度的收敛。

(2) P-RGF(Prior-guided Random Gradient-Free)[44] 方法。Cheng 等提出了一种先验引导的随机无梯度方法来改进黑盒对抗攻击，该方法将基于迁移的先验知识和查询攻击相结合，以提高梯度估计的精确度。基于迁移的先验知识由白盒替代模型的梯度计算得出，其包含关于真实梯度的丰富先验知识；同时从理论上分析了控制梯度强度的最优系数。该方法通过更少的查询次数和更高的成功率对黑盒进行对抗攻击。

(3)LBAT[45]方法。2021 年，Ding 等针对黑盒攻击中查询次数较多、成功率较低并且图像出现失真的情况，提出了一种基于可转移性的低查询黑盒对抗攻击。该方法将基于优化的方法和基于转移的方法相结合，其首先攻击代理模型，计算梯度估计的对抗向量；在生成对抗向量的过程中，LBAT 在每一步随机选择代理模型的参数，保证生成的对抗向量在每次都有较大的差异；由于黑盒攻击优化过程中代理模型和攻击模型的损失函数不同，通过增加动态系数优化白盒攻击的损失函数，来确保白盒攻击和黑盒攻击之间损失函数的一致性。LBAT 以其高效的查询能力实现了高成功率的黑盒攻击。

### 2.2.4　基于 GAN 的攻击方法

基于 GAN 的攻击方法指，利用 GAN 网络的生成器和判别器生成对抗样本，生成器生成对抗样本试图欺骗判别器；相反，判别器的目的是不被生成的对抗样本所欺骗，能够成功鉴别出对抗样本。通过联合优化生成器和判别器，直到生成器生成的对抗样本成功欺骗判别器。

(1)AdvGAN(Generating Adversarial Examples with Adversarial Network)[46]方法。2018 年，Xiao 等[47]使用生成对抗网络(Generative Adversarial Networks，GAN)生成对抗样本。AdvGAN 框架由 3 部分构成：生成器 $G$、判别器 $D$ 和目标攻击网络 $f$。首先将干净样本 $x$ 输入生成器 $G$，生成微小的扰动 $G(x)$，其次将对抗样本 $x_{adv}=x+G(x)$ 送入判别器 $D$，区分干净样本和对抗样本；最后使用对抗样本 $x_{\mathrm{adv}}$ 攻击目标网络 $f$，计算对抗样本预测类与目标类之间的距离，输出损失值 $L_{\mathrm{adv}}$，迭代此过程生成对抗样本。

AdvGAN 方法可以生成逼真自然的对抗样本，实现高效的攻击；在面对黑盒与灰盒攻击时，不需要借助可转移性可以直接攻击深度神经网络。

(2)DaST(Data-free Substitute Training)[48]方法。由于黑盒攻击需要预先训练白盒攻击生成对抗样本，其次迁移到黑盒中进行攻击，然而在现实世界中很难获得预先训练好的模型。针对此问题，Zhou 等提出了 DaST 方法，利用无数据替代训练方法获得黑盒攻击的替代模型。

DaST 利用生成对抗网络训练替代模型。其中生成模型 $G$ 为多分支结构，处理不均匀分布的样本，$G$ 从输入空间中随机采样噪声 $z$，生成对抗样本 $X_{\mathrm{adv}}=G(z)$；其次将对抗样本输入攻击模型 $T$ 中输出 $T(X_{\mathrm{adv}})$，替代模型 $D$ 由输出对($X_{\mathrm{adv}}$，$T(X_{\mathrm{adv}})$)进行训练。如果替代模型 $D$ 可以实现 $D(X_{\mathrm{adv}})=T(X_{\mathrm{adv}})$，则在不能获得攻击模型 $T$ 的梯度信息的情况下，可以利用替代模型进行攻击。

### 2.2.5　基于决策边界的攻击方法

基于决策边界的攻击方法首先寻找较大的扰动向量，然后在保持对抗性的同时逐渐减少扰动，直至网络模型分类错误。该攻击方法不需要调整超参数，不依赖于替代模型，仅依赖于模型最终决策的直接攻击。

(1)DeepFool[49]方法。Moosavi-Dezfooli 等提出了 DeepFool 方法，其基本思想基于二分类问题。对于一个超平面分类边界 $F\triangleq\{x;f(x)=0\}$，定义 $f(x)=w^{\mathrm{T}}x+b$，给定一个样本 $x_0$，如果要改变它的分类结果，最佳的方法就是将 $x_0$ 沿着垂直于超平面的方向移动到另一边，如图 4 所示。

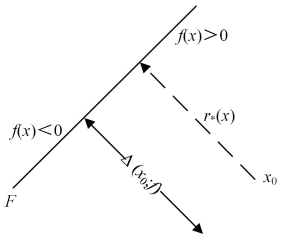


图 4　线性二分类问题中的对抗样本

Fig. 4　Adversarial example in linear binary classification problems

同理，在对抗攻击问题中，添加的最小扰动向量为 $x_0$ 到直线之间的垂直距离向量，如式(18)所示：

$$r_*(x_0)=\arg\min\|r\|_2=-\frac{f(x_0)}{\|w\|_2^2}w \tag{18}$$

其中，$f$ 表示神经网络；$\omega$ 表示权重系数；$r$ 表示扰动。

此时，$x_0$ 的鲁棒性 $\Delta(x_0;f)^2$ 可以用 $x_0$ 到边界 $F$ 的距离表示，通过迭代的方法估计鲁棒性的变化 $\Delta(x_0;f)$。在每次迭代中，$f$ 在当前点 $x_i$ 附近线性化，线性化分类器的最小扰动 $r_i$ 如式(19)所示：

$$\arg\min_{r_i}\|r_i\|_2\quad \text{s.t. } f(x_i)+\nabla f(x_i)^{\mathrm{T}}r_i=0 \tag{19}$$

其中，$r_i$ 表示 $i$ 阶段的扰动；$f$ 表示神经网络；$\nabla$表示梯度信息。

DeepFool 攻击使用迭代方式逐步生成对抗样本。与 FGSM 方法相比，该方法只需添加较小的扰动就可以生成攻击性能较好的对抗样本，然而缺点在于计算量较大，需要逐步计算添加扰动的大小。

(2)One pixel attack[50]方法。2019 年，Su 等提出了 One pixel attack 黑盒攻击方法。该方法使用差分进化算法，限制修改的像素数，在每次像素迭代的过程中将生成的子图像与父图像进行比较，如果子图像的攻击效果优于父图像，则保留子图像，实现高效的对抗攻击。

One pixel attack 攻击是一种极端的方法，不需要了解分类神经网络内部结构和参数，只需改变一个像素就可以实现有效的对抗攻击。与基于梯度的方法相比，该方法加入差分进化算法，通过比较父图像与子图像的差异，选择最优的对抗样本；差分进化算法容易找出全局最小值，避免局部最小值；此外，差分进化算法不需要关于微分的信息，因此该方法适用于网络不可微或者难以获取梯度下降的情形。

### 2.2.6　对抗攻击方法类型总结

在上述对抗攻击方法中，根据是否知道深度神经网络的模型结构，可以将对抗攻击分为白盒攻击、黑盒攻击和灰盒攻击。本节将对上述攻击方法进行总结分类，如图 5 所示。

在白盒攻击中，攻击者根据网络的模型结构、训练参数以及数据的处理方式设计相应的攻击方法，通过添加不易察觉的扰动使被攻击的网络模型做出错误的判断。上述 FGSM，JSMA，BIM，DeepFool，MI-FGSM，DI$^2$FGSM，C&W 和 SVD-Universal 属于白盒攻击。

与白盒攻击相反，黑盒攻击通过查询神经网络的输出数据生成对抗样本。2019 年，George 等[51]为黑盒攻击寻找了另外一种方法，该研究发现对抗攻击具有可转移性，即对于深度神经网络，可以通过构造白盒攻击生成对抗样本，其次转移对抗样本对黑盒进行攻击。上述 AdvFlow，ZOO，AdvGAN，One pixel，DaST 和 Simulator Attack 属于黑盒攻击。

灰盒攻击指攻击者只知道模型的一部分信息，如目标网络的模型结构或查询访问的次数。由于对抗样本具有可转移性，因此白盒攻击生成的对抗样本同样适用于灰盒攻击。其中 AdvGAN 不仅属于黑盒攻击，也属于灰盒攻击。

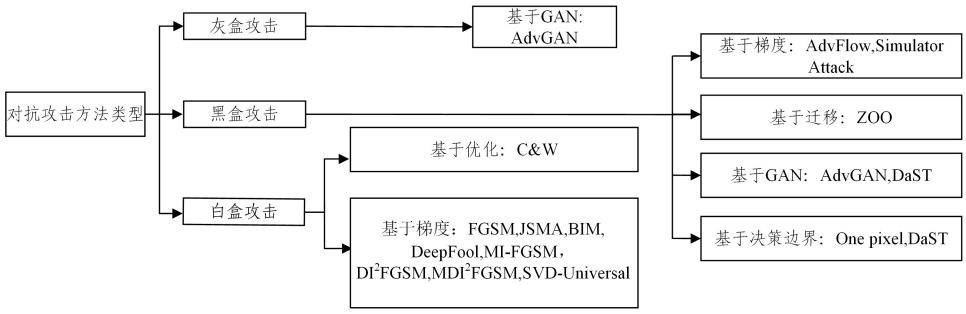


图 5 对抗攻击方法类型的总结

Fig. 5 Summary of types of adversarial attack methods

## 3 对抗攻击防御方法

针对对抗攻击带来的诸多问题，研究人员先后提出了多种方法进行防御，主要分为 3 种：1）数据预处理；2）增强神经网络的鲁棒性；3）检测对抗样本。

### 3.1 数据预处理

数据预处理指在对抗样本输入深度神经网络之前修改对抗样本，消除对抗扰动，以缓解对抗攻击带来的影响。数据预处理应用于各种类型的对抗样本，防御不同类型的攻击，在不降低模型对干净样本预测精度的情况下保持可用性。

#### 3.1.1 对抗样本去噪

Xie 等[20]通过对对抗样本特征去噪，来提高神经网络的鲁棒性。通过对非局部均值滤波器[52]、双侧过滤器[53]、均值过滤器[54]和中值过滤器[55]进行实验，来证明非局部均值和残差连接的方式可以有效地防御对抗样本。

Li 等[56]提出了 UDDN(U-Net Deep Denoising Network)深度去噪神经网络。该网络以 U-Net 为基础，添加了去噪自动编码器(Denoising Auto Encoder，DAE)。U-Net 由压缩路径和扩展路径组成，压缩路径捕获图像中的上下文信息，扩展路径通过扩展空间尺寸逐渐恢复图像的细节；将 DAE 特征图融入解码层，恢复图像的纹理信息。UDDN 通过残差学习的方式训练神经网络，去除对抗样本中的噪声，提升神经网络的鲁棒性。

Niu 等[57]提出了 ACM(Adaptive Compression and Reconstruction Model)方法。首先使用空域、频域和潜在空间的去噪方法防御对抗攻击，结果表明这些去噪策略均存在局限性。在此基础上，他们提出了自适应压缩和重构模型防御策略，首先在压缩约束下，图像的语义信息比不规则的对抗噪声更容易重构；其次，由于压缩信息使得图像发生了一定程度的变化，即使对抗扰动依然存在，也无法完成攻击。并且根据其特点，应用于图像的高频段与低频段，消除对抗扰动，提升鲁棒性效果。

#### 3.1.2 数据压缩

Das 等[22]认为，JPEG 压缩可以选择性地丢弃不易察觉的信息，在训练的过程中加入 JPEG 压缩图像，可以有效地防御 FGSM 和 DeepFool 攻击。这种方法具有便捷性，不需要知道模型的结构和攻击方式。Dziugaite 等[23]认为，在神经网络训练的过程中使用 JPG 压缩的数据集，JPG 图像空间中不会留下对抗扰动，可以有效消除对抗攻击带来的影响。

作为防御方法，数据压缩的核心是基于人类心理视觉系统，该系统使用离散余弦变换抑制高频信息，如色调和亮度的转换。实验结果表明，JPG 压缩可以有效地抵御 FGSM 攻击生成的小幅度对抗样本，然而随着对抗扰动的增加，JPG 的防御性能会降低。

#### 3.1.3 像素偏移

Prakash 等[58]提出了像素偏移防御机制。与图像去噪和超分辨率重建相比，该方法关注对抗样本中的关键信息区域，该区域是对抗扰动最显著的区域，也最影响神经网络的分类结果，同时也是最重要的图像恢复区域。通过使用像素偏移方法，从小邻域中随机选择像素点，使用 $r$ 边心距替换像素点；其次使用小波变换平滑对抗样本。两种方法的融合减小了对抗攻击的影响，这种方法在区域图像进行修复比对整个图像进行重建更容易处理。最后结论表明，此方法可以有效抵御 L-FGSM，FGSM，C&W，DeepFool，JSMA，BIM 攻击，并且不会降低干净样本的准确率。

对输入数据进行预处理的优点在于计算速度快，不需要修改网络结构，缺点在于去噪和数据压缩会造成输入数据高频信息的丢失，导致网络从错误的特征区域提取特征，使分类神经网络做出错误的判断。

### 3.2 增强神经网络的鲁棒性

增强神经网络的鲁棒性是通过在深度神经网络中添加更多的层、子网络或者增强神经网络的泛化能力来提升鲁棒性。

#### 3.2.1 对抗训练

对抗训练[16,24-25]是将对抗样本集加入到神经网络的训练集中，与干净样本一同训练。神经网络将对抗样本看作干净样本，拟合数据的分布，覆盖对抗样本的盲区，从而达到防御的效果。将对抗样本和原始数据一起训练，对抗样本产生的损失作为神经网络损失的一部分，在不修改原模型结构的情况下增加模型的损失，产生正则化的效果。对抗训练需要大量的时间，并且随着新的对抗攻击方法的提出，模型的鲁棒性不会有显著的提升。

#### 3.2.2 深度压缩网络

Gu 等[28]认为，对抗样本的存在是前向神经网络的固有属性，这与神经网络的训练过程和目标函数有关，总是可以根据反向传播误差找到对抗样本。其提出了一种端到端的深度压缩网络模型，将输入不变性传播到网络的输出。在前向神经网络中加入压缩自动编码器，利用压缩自动编码器逐层平滑的性质，最小化网络输入与输出的相对方差，在进行端到端

的训练后实现模型的平滑。这种方法增加了网络的鲁棒性，并且对干净样本不会产生较大的性能损失。

3.2.3　防御蒸馏方法

Papernot 等[59]在蒸馏方法的基础上提出了防御蒸馏方法，可以抵御 FGSM 和 JSMA 攻击，防御方法如图 6 所示。

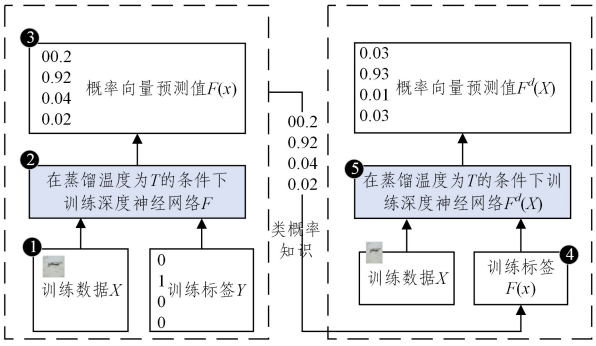


图 6　防御蒸馏网络示意图

Fig. 6　Illustration of defensive distillation network

图 6 中左边为原始网络，也称为教师网络，其网络结构复杂并且参数计算量较大；右边为蒸馏网络，也称为学生网络，其网络结构简单，具有较少的参数；参数 $T$ 称为蒸馏温度，通过调整 $T$ 得到输出概率。防御蒸馏方法首先在蒸馏温度为 $T$ 的条件下训练教师网络以获取先验知识，其次将先验知识送入蒸馏温度同样为 $T$ 的学生网络继续训练，得到最终的输出结果。防御蒸馏方法不需要平滑神经网络的输出来提升模型的鲁棒性，可以有效地防御白盒攻击，但对于黑盒攻击其防御性能较低。

通过提高模型的随机性和认知性能来提高网络的复杂性，增强神经网络的鲁棒性；但其需要重新训练网络，因此计算开销较大；并且面对精心设计的特定攻击时，防御效果欠佳。

### 3.3　检测对抗样本

检测对抗样本是使用神经网络，通过阈值策略区分干净样本和对抗样本。如果是干净样本，则直接把数据输入神经网络，否则利用增强神经网络鲁棒性的防御方法来减轻对抗样本带来的影响。该方法具有计算成本低、不需要改变或者重新训练神经网络的特点。

3.3.1　基于 Generative Adversarial Network(GAN)网络

Samangouei 等[29]提出了 Defense-GAN 网络，用于提升神经网络的鲁棒性，GAN 网络架构由生成模型和判别模型组成，生成模型 $G$ 模拟数据的分布，对抗模型 $A$ 判断输入的样本是干净样本还是对抗样本。在训练过程中，Defense-GAN 生成干净样本的模拟数据分布，当面对对抗样本时，生成模型 $G$ 学习低维到高维的映射，生成对抗样本的近似样本，其满足干净样本的数据分布。该方法将对抗样本投射到生成模型 $G$ 的范围内，减少对抗性扰动。

Defense-GAN 不改变分类神经网络的结构，可以与任何神经网络联合使用，结果表明，Defense-GAN 可以有效地抵御白盒与黑盒攻击，同时不需要重新训练神经网络，并且对干净样本的分类性能不会有显著影响。

3.3.2　基于 MagNet 网络

上述的防御方法存在只能对特定的攻击进行防御或者防御性能不佳的问题。针对此情况，Meng 等[30]提出了 MagNet 框架，该研究认为神经网络错误分类对抗样本的原因有：1）对抗样本偏离干净样本的流形区域；2）分类神经网络在流形边界附近的泛化能力较差，不能正确分类对抗样本与干净样本。MagNet 网络框架包括探测器(detector)网络和重组(reformer)网络，探测器逼近干净样本的流形，学习区分干净样本和对抗样本，从而达到检测的目的；重组网络重构对抗样本，将对抗样本移向干净样本流形，从而提升分类神经网络的分类效率。

3.3.3　基于 Defense Perturbation 方法

Nesti 等[60]提出防御扰动方法，旨在检测具有鲁棒性的对抗样本。该方法首先对对抗样本进行输入转换，评估神经网络的检测能力；其次，在具有鲁棒性的对抗样本中添加像素掩码，将其转换为非鲁棒性的对抗样本进行检测；进一步地，通过投票算法将多种网络体系结构进行网络组合，进一步提高检测性能；最终将所提出的方法结合起来，设计一个有效的架构检测鲁棒性的对抗样本。该方法检测对抗样本简单并且计算成本较低，不需要考虑复杂的模型。

## 4　对抗攻击与防御实例

随着对抗攻击研究的深入，研究人员发现对抗攻击不仅应用于图像领域，在音频识别[61]、目标检测[62]、语义分割[63]、人脸识别[64]和强化学习[65]领域也有着广泛的应用。本小节将介绍对抗攻击及防御方法的实际应用。

(1)语义分割中的对抗攻击与防御

Fischer 等[66]使用基于梯度的攻击方法，将对抗攻击成功地应用在语义分割中，使深度神经网络输出错误的语义结果，通过将像素分配到其最接近的错误类别生成对抗样本。Arnab 等[67]使用大规模 Pascal VOC[68]和 Cityscapes[69]数据集，分析不同的网络架构、模型容量、图像变换和多尺度处理的影响，并表明对抗攻击的迁移能力并不总是可以迁移到更复杂的模型中。

针对语义分割领域的对抗攻击，研究人员从频域的角度分析指出神经网络的鲁棒性与输入样本中高频信息的敏感性有关。Kapoor 等[70]对数据进行预处理，使用 Wiener filters 在频域中抑制对抗样本的噪声，从而提升神经网络的鲁棒性；同时表明 Wiener filters 对不可见的攻击具有较强的泛化特性。

(2)音频中的对抗攻击与防御

在音频领域同样可以实现对抗攻击，对于给定的音频 $x$，生成扰动 $\delta$，使得对抗样本 $x_{adv}=x+\delta$ 能够被深度神经网络识别为错误的标签。2018 年，Carlini 等[71]使用优化的攻击方法，首次对语音识别进行了对抗攻击，他们指出对于给定的音频波形，加上扰动之后，对抗样本可以输出为任何短语，并且对抗样本与干净样本的波形相似度超过 99.9%。

针对音频领域的攻击，Kwon 等[72]提出了一种用于检测音频对抗样本的声学诱饵防御方法，其依赖于原始样本分类结果与防御后对抗样本分类结果之间的差异。该方法使用低通滤波器从音频中去除高频范围，保证了与原始样本较高的相似率，同时消除了对抗样本中的高频噪声。实验结果表明，这种方法不会影响原始样本的分类结果，而对抗样本的检测成功率达到了 97%。

(3)文本识别中的对抗攻击与防御

文本领域中的对抗攻击指通过改变输入文本的几个

字符，神经网络就可以错误地输出文本表达的意思，然而人们的判断并不会因此受到影响。Li 等[73]基于雅克比矩阵提出了 TextBugger 方法，该方法通过提取文本的关键词，采用插入、删除、字符交换、字符替换和单词替换 5 种方法对关键词进行扰动生成对抗样本，并选择最优的对抗样本，大大地提升了攻击效率。

针对文本领域的对抗攻击，研究人员首先提出对抗训练[74-76]方法进行防御；针对同义词替换的攻击，Wang 等[77]提出了 SEM(Synonym Encoding Method)方法。该方法首先在网络模型的输入层之前插入一个编码器，为对抗样本进行数据预处理，然后将同义词进行聚类，为每个同义词进行编码，并为每个聚类分配一个唯一的标识符以防御对抗攻击。实验结果表明，SEM 方法降低了对抗攻击的可转移性，有效地防御了对抗攻击，同时保证了在干净样本上的分类精度。Zhou 等[78]提出了 DISP(Discriminate Perturbations)方法，该方法由 perturbation discriminator，embedding estimator 和 hierarchical navigable small world graphs 组成，通过在神经网络中添加模块的方法来增强神经网络的鲁棒性。

(4)目标检测中的对抗攻击与防御

目标检测领域中的对抗攻击指，深度神经网络检测错误或者不能正确检测出目标。Lu 等[79]提出了一种对抗攻击方法，该方法通过梯度下降的优化方法成功地欺骗了 Faster RCNN，并且对抗样本可以迁移到 YOLO9000 模型中。Liu 等[80]提出了 DPatch 方法，该方法在输入样本中添加一个随机初始化补丁块，将其输入目标检测网络得到检测结果，根据损失函数反向更新补丁块中的像素值，通过迭代初始化补丁块像素值实现对抗攻击。

针对目标检测领域的对抗攻击，Li 等[81]提出用增强神经网络鲁棒性的鲁棒显著(Robust Salient，ROSA) 目标检测框架进行防御。该框架由分段屏蔽器、全卷积神经网络和上下文感知恢复组件组成，分段屏蔽器首先引入通用噪声破坏对抗样本中的扰动，之后将对抗样本随机划分，并将每个划分区域中的像素随机排列，这不仅能限制对抗噪声，还可以去除新引入的噪声；然后将分段屏蔽器输出的图像输入全卷积神经网络，得到粗糙的显著图；最后利用上下文感知恢复组件通过全局对比度建模完善显著性图。ROSA 方法的结果更清晰并且显著增强了主干网络的鲁棒性。

(5)人脸识别中的对抗攻击与防御

Dong 等[82]针对人脸识别系统，首次提出了基于决策边界的黑盒攻击，这是一种进化的攻击算法，可以对搜索方向的局部几何形状进行建模，减少搜索空间的维度，实验结果证明，该方法的扰动较小，查询次数较少，并且在现实生活中有很强的适用性。Komkov 等[83]提出了一种基于 FGSM 的 AdvHat 方法，该方法首先对矩形图像计算位置和形态的扰动，然后进行非平面变换，最后将带有扰动的矩形图像粘贴在额头上，使得 Face ID 系统判别错误。

针对人脸识别领域中的攻击，腾讯优图从两个方向进行攻击防御：

1)对抗样本的检测，如果检测出对抗样本，则拒绝识别模型；

2)增强神经网络的对抗攻击能力，即输入图像为对抗样本时，神经网络也能识别正确。

(6)强化学习中的对抗攻击与防御

Xiang 等[84]关注强化学习中 Q 学习的对抗攻击，通过计算因子构建一个线性模型，使用主成分分析计算权重参数拟合因子，最后基于影响因子的概率输出模型和相应的权重，预测对抗样本。Qu 等[85]提出了 Minimalistic Attacks 方法，定义了 3 种方法来寻找最小化攻击：

1)攻击者仅能访问强化学习中游戏策略的输入状态和输出动作概率；

2)只有几个像素被攻击，极端情况下进行单像素攻击；

3)选择关键帧进行攻击。

通过限定 3 种方法生成对抗样本，发现仅仅攻击单个像素，训练成功的游戏策略可以被欺骗；并且攻击大约 1% 的帧，Deep Q Network(DQN)训练的游戏策略在某些游戏上完全被欺骗。

针对强化学习的对抗攻击，Behzadan 等[86]研究了非连续对抗扰动下 DQN 的鲁棒性。该方法将对抗样本以概率 $p$ 加入到对抗训练过程中，当 $p=0.2$ 和 $p=0.4$ 时，DQN 能够保持对抗样本的鲁棒性。Havens 等[87]提出了 MLAH(Meta-Learned Advantage Hierarchy)方法，该方法通过处理决策空间的攻击来减小对抗引入的学习偏差。该方法与攻击模型无关，在监督主代理的监督下，子策略函数检测对抗样本的存在。实验结果表明，与最新的策略学习方法相比，所提方法能够以较低的偏差进行策略学习。

## 5 对抗攻击与防御的未来发展趋势

自对抗攻击提出以来，就引起了人们广泛的关注，研究人员提出了众多对抗攻击方法和防御方法。随着对抗攻击及防御领域的技术更迭，未来对抗攻击的发展趋势可以总结为以下几方面：

(1)目前针对黑盒攻击的方法主要是利用对抗攻击的可转移性或者以查询的方式进行攻击，然而这两种方法在实际应用时会遇到很多问题。因此，在今后的研究中，需要设计了出快速高效、转移性能较强且查询次数较少的攻击模型。

(2)针对白盒、黑盒和灰盒攻击，对于不同类型的攻击，设计了一种高效的通用攻击方法，在减小对抗攻击资源开销的同时提升了攻击效率。

(3)在今后的研究中，应该不仅限于在输入图像中添加扰动，应该设计新型的扰动方法，加入更小的扰动并保持较好的攻击效果。

未来防御方法的发展趋势总结为以下几个方面：

(1)在对黑盒攻击进行防御时，探索防御方法之间的可转移性是以后研究的重点。利用可转移性，减少目标网络参数的计算量和复杂性，快速有效地防御对抗攻击。

(2)现有的对抗攻击防御方法大都是针对某一种特定的攻击方法而设计的，因此在以后的研究中，如何设计一种普遍通用的防御方法以抵御不同种类的攻击方法是对抗样本领域值得研究的方向。

(3)目前为止，对于对抗样本存在的原因并没有一个公认的解释，在以后的研究过程中，应从存在的原因作为切入点，解释对抗样本的形成，在深入了解存在的原因之后，研究设计鲁棒性较好的网络模型。

**结束语** 深度神经网络已经广泛应用于人工智能领域，

但其安全问题也逐渐引起了人们的关注。本文综述了对抗攻击领域的相关知识，首先介绍了对抗攻击的相关概念以及存在性解释；接下来从不同的角度讲述了对抗攻击方法；在此基础上介绍了对抗攻击的防御方法，从不同的防御类型阐述了不同的方法；回顾了对抗攻击与防御实例；最后分析对抗攻击及防御方法的未来发展趋势。应当指出，对抗攻击不是对深度学习的否定，而是通过对抗攻击提升神经网络的鲁棒性，构建一个安全、可控的人工智能环境。

## 参 考 文 献


[1] TIAN Y，PEI K，JANA S，et al. Deeptest：Automated testing of deep-neural-network-driven autonomous cars[C]// Proceedings of the 40th International Conference on Software Engineering. New York，2018：303-314.

[2] CHEN C，SEFF A，KORNHAUSER A，et al. Deepdriving：Learning affordance for direct perception in autonomous driving [C]// Proceedings of the IEEE International Conference on Computer Vision. Piscataway，2015：2722-2730.

[3] LITJENS G，KOOI T，BEJNORDI B E，et al. A survey on deep learning in medical image analysis[J]. Medical Image Analysis，2017，42：60-88.

[4] SHEN D，WU G，SUK H I. Deep learning in medical image analysis[J]. Annual Review of Biomedical Engineering，2017，19：221-248.

[5] HE K，ZHANG X，REN S，et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscata-way，2016：770-778.

[6] HUANG G，LIU Z，VAN DER MAATEN L，et al. Densely connected convolutional networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway，2017：4700-4708.

[7] SIMONYAN K，ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [OL]. (2015-04-10) [2021-08-12]. https://arxiv. org/pdf/1409. 1556. pdf.

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

[9] YUFEI C，CHAO S，QIAN W，et al. Security and Privacy Risks in Artificial Intelligence Systems[J]. Journal of Computer Research and Development，2019，56(10)：2135.

[10] SZEGEDY C，ZAREMBA W，SUTSKEVER I，et al. Intriguing properties of neural networks[OL]. (2014-02-19)[2021-08-12]. https://arxiv. org/pdf/1312. 6199. pdf.

[11] GOODFELLOW I J，SHLENS J，SZEGEDY C. Explaining and harnessing adversarial examples[OL]. (2015-02-25) [2021-08-12]. https://arxiv. org/pdf/1412. 6572. pdf.

[12] KURAKIN A，GOODFELLOW I，BENGIO S. Adversarial examples in the physical world[OL]. (2017-02-11)[2021-08-12]. https://arxiv. org/pdf/1607. 02533. pdf.

[13] JOSHI A，MUKHERJEE A，SARKAR S，et al. Semantic adversarial attacks：Parametric transformations that fool deep classifiers[C]// Proceedings of the IEEE International Conference on Computer Vision. Piscataway，2019：4773-4783.

[14] FAN Y，WU B，LI T，et al. Sparse adversarial attack via perturbation factorization[C]// Proceedings of European Conference on Computer Vision. Cham，2020：.

[15] GUO M，YANG Y，XU R，et al. When NAS Meets Robustness：In Search of Robust Architectures against Adversarial Attacks [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway，2020：631-640.

[16] ZHANG H，WANG J. Defense against adversarial attacks using feature scattering-based adversarial training[C]// Advances in Neural Information Processing Systems. Vancouver，2019：1831-1841.

[17] PANG T，DU C，ZHU J. Robust deep learning via reverse cross-entropy training and thresholding test [OL]. (2018-11-07) [2021-08-12]. https://arxiv. org/pdf/1706. 00633. pdf. 2021.

[18] METZEN J H，GENEWEIN T，FISCHER V，et al. On detecting adversarial perturbations [OL]. (2017-02-21) [2021-08-12]. https://arxiv. org/pdf/1702. 04267. pdf，2021.

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

[20] 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. Piscataway，2019：501-509.

[21] 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. Piscataway，2018：1778-1787.

[22] DAS N，SHANBHOGUE M，CHEN S-T，et al. Keeping the bad guys out：Protecting and vaccinating deep learning with jpeg compression[OL]. (2017-05-08)[2021-08-12]. https://arxiv. org/pdf/1705. 02900. pdf. 2021. 8.

[23] DZIUGAITE G K，GHAHRAMANI Z，ROY D M. A study of the effect of jpg compression on adversarial images[OL]. (2016-08-02)[2021-08-12]. https://arxiv. org/pdf/1608. 00853. pdf.

[24] TRAMÈR F，KURAKIN A，PAPERNOT N，et al. Ensemble adversarial training：Attacks and defenses[OL]. (2020-04-26) [2021-08-12]. https://arxiv. org/pdf/1705. 07204. pdf.

[25] KURAKIN A，GOODFELLOW I，BENGIO S. Adversarial machine learning at scale [OL]. (2017-02-11) [2021-08-12]. https://arxiv. org/pdf/1611. 01236. pdf.

[26] PAPERNOT N，MCDANIEL P. On the effectiveness of defensive distillation[OL]. (2016-07-18) [2021-08-12]. https://arxiv. org/pdf/1607. 05113. pdf.

[27] NAYEBI A，GANGULI S. Biologically inspired protection of deep networks from adversarial attacks [OL]. (2017-03-27) [2021-08-12]. https://arxiv. org/pdf/1703. 09202. pdf.

[28] GU S，RIGAZIO L. Towards deep neural network architectures robust to adversarial examples[OL]. (2015-04-09) [2021-08-12]. https://arxiv. org/pdf/1412. 5068. pdf.

[29] SAMANGOUEI P，KABKAB M，CHELLAPPA R. Defense-gan：Protecting classifiers against adversarial attacks using generative models[OL]. (2015-05-18) [2021-08-12]. https://arxiv. org/pdf/1805. 06605. pdf.

[30] 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. New York, 2017: 135-147.
[31] TANAY T, GRIFFIN L. A boundary tilting persepective on the phenomenon of adversarial examples[OL]. (2016-08-27)[2021-08-12]. https://arxiv.org/pdf/1608.07690.pdf.
[32] GILMER J, METZ L, FAGHRI F, et al. Adversarial spheres [OL]. (2018-09-10) [2021-08-12]. https://arxiv.org/pdf/1801.02774.pdf.
[33] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[OL]. (2019-09-04) [2021-08-12]. https://arxiv.org/pdf/1706.06083.pdf.
[34] 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). IEEE, Piscataway, 2016: 372-387.
[35] LECUN Y. The MNIST database of handwritten digits[OL]. [2021-08-12]. http://yann.lecun.com/exdb/mnist/.
[36] DONG Y, LIAO F, PANG T, et al. Boosting adversarial attacks with momentum[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. Piscataway, 2018: 9185-9193.
[37] XIE C, ZHANG Z, ZHOU Y, et al. Improving transferability of adversarial examples with input diversity[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, 2019: 2730-2739.
[38] DOLATABADI H M, ERFANI S, LECKIE C. AdvFlow: Inconspicuous Black-box Adversarial Attacks using Normalizing Flows[OL]. (2020-10-23) [2021-08-12]. https://arxiv.org/pdf/2007.07435.pdf. 2021.8.
[39] KAMATH S, DESHPANDE A, SUBRAHMANYAM K. Universalization of any adversarial attack using very few test examples[OL]. (2020-05-18)[2021-08-12]. https://arxiv.org/pdf/2005.08632.pdf.
[40] MA C, CHEN L, YONG J H. Simulating Unknown Target Models for Query-Efficient Black-box Attacks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Piscataway. 2021: 11835-11844.
[41] XU X, CHEN J, XIAO J, et al. Learning optimization-based adversarial perturbations for attacking sequential recognition models[C]// Proceedings of the 28th ACM International Conference on Multimedia. New York, 2020: 2802-2822.
[42] 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. New York, 2017: 15-26.
[43] KINGMA D P, BA J. Adam: A method for stochastic optimization[OL]. (2017-01-30) [2021-08-12]. https://arxiv.org/pdf/1412.6980.pdf.
[44] CHENG S, DONG Y, PANG T, et al. Improving black-box adversarial attacks with a transfer-based prior[OL]. (2020-07-26) [2021-08-12]. https://arxiv.org/pdf/1906.06919.pdf.
[45] DING K, LIU X, NIU W, et al. A low-query black-box adversarial attack based on transferability[J]. Knowledge-Based Systems, 2021, 226: 107102.
[46] XIAO C, LI B, ZHU J Y, et al. Generating adversarial examples with adversarial networks[OL]. (2019-02-14) [2021-08-12]. https://arxiv.org/pdf/1801.02610.pdf.
[47] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[J]. Advances in Neural Information Processing Systems, 2014, 27: 2672-2680.
[48] ZHOU M, WU J, LIU Y, et al. Dast: Data-free substitute training for adversarial attacks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, 2020: 234-243.
[49] 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. Piscataway, 2016: 2574-2582.
[50] SU J, VARGAS D V, SAKURAI K. One pixel attack for fooling deep neural networks[J]. IEEE Transactions on Evolutionary Computation, 2019, 23(5): 828-841.
[51] ADAM G, SMIRNOV P, HAIBE-KAINS B, et al. Reducing adversarial example transferability using gradient regularization [OL]. (2019-04-16) [2021-08-12]. https://arxiv.org/pdf/1904.07980.pdf.
[52] 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). Piscataway. 2005: 60-65.
[53] TOMASI C, MANDUCHI R. Bilateral filtering for gray and color images[C]// Sixth international conference on computer vision (IEEE Cat. No. 98CH36271). Piscataway, 1998: 839-846.
[54] ZHANG P, LI F. A new adaptive weighted mean filter for removing salt-and-pepper noise[J]. IEEE Signal Processing Letters, 2014, 21(10): 1280-1283.
[55] IBRAHIM H, KONG N S P, NG T F. Simple adaptive median filter for the removal of impulse noise from highly corrupted images[J]. IEEE Transactions on Consumer Electronics, 2008, 54(4): 1920-1927.
[56] LI Y, WANG Y. Defense against adversarial attacks in deep learning[J]. Applied Sciences, 2019, 9(1): 76.
[57] NIU Z, CHEN Z, LI L, et al. On the Limitations of Denoising Strategies as Adversarial Defenses[OL]. (2020-12-17)[2021-08-12]. https://arxiv.org/pdf/2012.09384.pdf.
[58] 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. Piscataway, 2018: 8571-8580.
[59] 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). Piscataway, 2016: 582-597.
[60] NESTI F, BIONDI A, BUTTAZZO G. Detecting Adversarial Examples by Input Transformations, Defense Perturbations, and Voting[OL]. (2021-01-27) [2021-08-12]. https://arxiv.org/pdf/2101.11466.pdf.
[61] LEE H, PHAM P, LARGMAN Y, et al. Unsupervised feature learning for audio classification using convolutional deep belief networks[J]. Advances in Neural Information Processing Systems, 2009, 22: 1096-1104.

[62] WEI X，LIANG S，CHEN N，et al. Transferable adversarial attacks for image and video object detection[OL]. (2019-05-13) [2021-08-12]. https://arxiv. org/pdf/1811. 12641. pdf. 2021. 8.

[63] 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. Piscataway，2017：2755-2764.

[64] 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. New York，2016：1528-1540.

[65] MNIH V，KAVUKCUOGLU K，SILVER D，et al. Human-level control through deep reinforcement learning[J]. Nature，2015，518(7540)：529-533.

[66] FISCHER V，KUMAR M C，METZEN J H，et al. Adversarial examples for semantic image segmentation[OL]. (2017-03-03) [2021-08-12]. https://arxiv. org/pdf/1703. 01101. pdf.

[67] ARNAB A，MIKSIK O，TORR P H. On the robustness of semantic segmentation models to adversarial attacks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway，2018：888-897.

[68] EVERINGHAM M，VAN GOOL L，WILLIAMS C K，et al. The pascal visual object classes (voc) challenge[J]. International Journal of Computer Vision，2010，88(2)：303-338

[69] CORDTS M，OMRAN M，RAMOS S，et al. The cityscapes dataset for semantic urban scene understanding[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway，2016：3213-3223.

[70] KAPOOR N，BÄR A，VARGHESE S，et al. From a Fourier-Domain Perspective on Adversarial Examples to a Wiener Filter Defense for Semantic Segmentation[OL]. (2021-04-21) [2021-08-12]. https://arxiv. org/pdf/2012. 01558. pdf.

[71] CARLINI N，WAGNER D. Audio adversarial examples：Targeted attacks on speech-to-text[C]// 2018 IEEE Security and Privacy Workshops(SPW). Piscataway，2018：1-7.

[72] KWON H，YOON H，PARK K W. Acoustic-decoy：Detection of adversarial examples through audio modification on speech recognition system[J]. Neurocomputing，2020，417：357-370.

[73] LI J，JI S，DU T，et al. Textbugger：Generating adversarial text against real-world applications[OL]. (2018-12-13) [2021-08-12]. https://arxiv. org/pdf/1812. 05271. pdf.

[74] LI L，MA R，GUO Q，et al. Bert-attack：Adversarial attack against bert using bert[OL]. (2020-10-02) [2021-08-12]. https://arxiv. org/pdf/2004. 09984. pdf.

[75] ZANG Y，QI F，YANG C，et al. Word-level textual adversarial attacking as combinatorial optimization[OL]. (2020-12-09) [2021-08-12]. https://arxiv. org/pdf/1910. 12196. pdf.

[76] JIN D，JIN Z，ZHOU J T，et al. Is bert really robust? a strong baseline for natural language attack on text classification and entailment[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park，2020：8018-8025.

[77] WANG X，JIN H，HE K. Natural language adversarial attacks and defenses in word level[OL]. (2021-06-15) [2021-08-12]. https://arxiv. org/pdf/1909. 06723. pdf.

[78] ZHOU Y，JIANG J Y，CHANG K W，et al. Learning to discriminate perturbations for blocking adversarial attacks in text classification[OL]. (2019-09-06) [2021-08-12]. https://arxiv. org/pdf/1909. 03084. pdf.

[79] LU J，SIBAI H，FABRY E. Adversarial examples that fool detectors[OL]. (2017-12-07) [2021-08-12]. https://arxiv. org/pdf/1712. 02494. pdf.

[80] LIU X，YANG H，LIU Z，et al. Dpatch：An adversarial patch attack on object detectors[OL]. (2019-04-23) [2021-08-12]. https://arxiv. org/pdf/1806. 02299. pdf.

[81] LI H，LI G，YU Y. ROSA：Robust salient object detection against adversarial attacks[J]. IEEE Transactions on Cybernetics，2019，50(11)：4835-4847.

[82] DONG Y，SU H，WU B，et al. Efficient decision-based black-box adversarial attacks on face recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway，2019：7714-7722.

[83] KOMKOV S，PETIUSHKO A. Advhat：Real-world adversarial attack on arcface face id system[OL]. (2019-08-23) [2021-08-12]. https://arxiv. org/pdf/1908. 08705. pdf.

[84] XIANG Y，NIU W，LIU J，et al. A PCA-based model to predict adversarial examples on Q-learning of path finding[C]// 2018 IEEE Third International Conference on Data Science in Cyberspace(DSC). Piscataway，2018：773-780.

[85] QU X，SUN Z，ONG Y S，et al. Minimalistic Attacks：How Little it Takes to Fool Deep Reinforcement Learning Policies[J]. IEEE Transactions on Cognitive and Developmental Systems，2020，13(4)：806-817.

[86] BEHZADAN V，MUNIR A. Whatever does not kill deep reinforcement learning，makes it stronger[OL]. (2017-12-23) [2021-08-12]. https://arxiv. org/pdf/1712. 09344. pdf.

[87] HAVENS A J，JIANG Z，SARKAR S. Online robust policy learning in the presence of unknown adversaries[OL]. (2018-07-16) [2021-08-12]. https://arxiv. org/pdf/1807. 06064. pdf.

**ZHAO Hong**，born in 1971. Ph. D，professor，Ph. D supervisor，is a senior member of China Computer Federation. His main research interests include system modeling and simulation，deep learning，natural language processing，and computer vision.

**CHANG You-kang**，born in 1994，Ph. D candidate. His main research interests include adversarial attacks and defense methods and deep learning.