DOI:10.12132/ISSN.1673-5048.2020.0012

针对神经网络的对抗攻击及其防御

何正保，黄晓霖*

（上海交通大学 自动化系，上海　200240）

摘　要：随着深度学习和神经网络的不断发展，深度神经网络已经广泛应用于多个领域，其安全性也日渐受到人们的关注。对抗攻击和对抗样本作为神经网络最大的威胁之一，近年来成为研究的热点。同时，对抗攻击及其防御的研究也对神经网络认知能力的提升具有重要的意义。本文围绕对抗攻击及其防御，介绍了基本原理和一些经典算法，并就对抗攻击的意义与作用、发展趋势进行了阐述。

关键词：深度学习；神经网络；对抗攻击；对抗样本；防御算法；人工智能

中图分类号：TJ760；TP18　　**文献标识码：**A　　**文章编号：**1673-5048(2020)03-0011-09

0 引　　言

随着深度学习的发展，深度神经网络已经广泛应用于图像识别[1-3]、医学影像分析[4-5]、自动驾驶[6-7]等领域，并且在许多复杂任务上的表现都超过了人类水平。但神经网络同样有很多问题，阻碍其进一步发展与应用。神经网络在对抗攻击下的脆弱性就是其中一个重要的问题，例如，对输入图片添加人眼不可见的扰动，就能使分类网络以高置信度将修改后的图片错误分类[8-10]。神经网络的脆弱性制约了其在复杂、未知、多变环境的应用，也制约了其在需要高可靠性的任务中的推广。这种脆弱性与神经网络在认知能力方面的缺乏紧密相关。近年来，针对研究神经网络弱点的对抗攻击及其防御，逐渐成为新的研究重点，其目的既在于增强神经网络的可靠性，也在于通过对攻击及其防御的迭代研究，提升神经网络的认知能力。

本文综述了对抗攻击的基本概念和原理，梳理经典的对抗攻击算法和相应的防御机制，同时探讨对抗攻击及其防御的发展趋势和其对人工智能的推动作用。

1 对抗攻击

1.1 数学描述与基本原理

对抗攻击是指对原始数据添加特定的扰动得到对抗样本，使得神经网络产生错误的输出。从深度学习的机理上理解，神经网络的训练是在训练数据集上进行的，而训练数据只是真实数据中的一部分，无法完全表示真实数据的分布特性。因此，总可以寻找到训练数据无法覆盖的空间使得神经网络发生错误。从神经网络的结构进行理解，由于深度神经网络是含有多个隐含层的高度非线性函数，输入数据的一些细小变化，都可能逐层传递、放大，并对最终输入结果产生巨大的影响。

对抗攻击的本质是寻找神经网络与人类认知的差异。其差异首先表现在神经网络会对特定的扰动过于敏感。以图像分类器为例，利用神经网络过于敏感的特性，可以对原始图像添加较小的扰动，使分类器将图片错误分类。上述攻击可表述为

$$\begin{aligned}&\text{Find}\quad \boldsymbol{x}'\\&\text{s.t.}\quad f(\boldsymbol{x}')\neq f(\boldsymbol{x})\\&\qquad \|\boldsymbol{x}'-\boldsymbol{x}\|\leqslant\epsilon\end{aligned}\tag{1}$$

式中：$\boldsymbol{x}$ 为原始样本；$\boldsymbol{x}'$ 为所生成的对抗样本；f 为被攻击的神经网络；ϵ 为事先设定的扰动裕度。

图 1 显示了对抗攻击与传统目标隐藏之间的区别（图片均下载于互联网）。训练集中的一个目标样本如图 1(a)所示。传统上，为了使该型飞机不被识别，需要设计迷彩涂装用以欺骗人类的视觉系统。但针对人类设计的迷彩未必能欺骗机器视觉，相反，机器视觉系统对于

收稿日期：2020-01-19

基金项目：国家重点研发项目(2018AAA0100702)；国家自然科学基金项目(61977046)

作者简介：何正保(1999-)，男，河南人，研究方向是深度学习的对抗攻击及其防御。

*通讯作者：黄晓霖(1983-)，男，江西人，工学博士，副教授，青年千人，研究方向是稳健机器学习的理论与方法。
E-mail：xiaolinhuang@sjtu.edu.cn

引用格式：何正保，黄晓霖　针对神经网络的对抗攻击及其防御[J]. 航空兵器，2020，27(3)：11-19.

He Zhengbao, Huang Xiaolin. Adversarial Attacks and Defenses Against Neural Networks [J]. Aero Weaponry, 2020, 27(3): 11-19. (in Chinese)

对抗攻击十分敏感。如图 1(c)所示，虽然图像和原始图像在视觉上没有显著的区别，但机器视觉系统却做出了错误的判断。这个例子展示了人类视觉系统和机器视觉系统的差异，也显示了在人工智能广泛应用的今天，航空兵器这样存在强对抗的领域，研究其对抗攻击及其防御的必要性。

(a) 训练集中的目标样本　(b) 改变视觉效果欺骗人类视觉系统　(c) 添加对抗噪声欺骗人工神经网络

图 1　Type Ⅱ对抗攻击示例

Fig. 1　An example of Type Ⅱ adversarial attack

与过于敏感相对应，神经网络与人类感知的不一致性还表现在某些情况下过于迟钝，即对原始图像添加较大的扰动，而分类器仍然以较高的置信度将对抗样本分类为原始类别，其数学表述如下：

$$\begin{aligned} &\text{Find} \quad \boldsymbol{x}' \\ &\text{s.t.} \quad f(\boldsymbol{x}') = f(\boldsymbol{x}) \\ &\qquad \| \boldsymbol{x}' - \boldsymbol{x} \| \geqslant \epsilon \end{aligned} \tag{2}$$

目前已有的对抗攻击集中于式(1)所描述的情况[8-10]，关于式(2)的攻击可见文献[11-12]；根据所对应的统计误差的分类，这两类对抗攻击被分别称为 Type Ⅱ和 Type Ⅰ对抗攻击。

1.2　特征的不一致

对抗样本的存在证实了神经网络与人类认知的差异。从特征的角度考察这种差异会发现，当深度神经网络所学习到的数据特征空间与真实数据特征空间不一致时，就会出现对抗样本。

当神经网络学习到数据中的冗余特征(一般是数据中的噪声特征)时，网络就会对这些冗余特征较为敏感(也因此，冗余特征又被称为非稳健特征)。如果在冗余特征空间中对输入做一定的扰动，由于人类的认知中没有考虑这类扰动而无法观察到显著的变化，但由于这类特征被神经网络识别并纳入决策体系之中，其微小的变化将使得神经网络的输出发生巨大的变化，即遭受 Type Ⅱ攻击。相应地，如果神经网络学习到的数据特征空间较小时，会出现一些网络没有学习到的缺失特征。这部分特征为人类所重视，但是网络并不利用这部分特征进行决策，因此，缺失特征上较大的扰动能为人类所观测，但不会引起网络输出的相应变化，即遭受 Type Ⅰ攻击。文献[13]给出了一个很有趣的例子。在这个看似简单的内外两个球面数据的分类问题中，如果特征数量与真实系统不一致(包括冗余特征[13]和缺失特征[12])，神经网络都会被对抗样本所攻击。

近年来，有很多研究者从理论的角度分析特征稳健性与对抗样本。文献[14]认为想要学习一个鲁棒的模型比学习一个标准模型需要更多的数据；文献[13,15-16]认为对抗样本在某些情况下是不可避免的，无论是由于计算的限制还是数据本身的特性等；文献[17]则认为对抗样本作为神经网络学习到的非稳健特征之一，有助于模型的泛化，只是这种特征不易被人类察觉，这种观点认为对抗样本只是一种“以人为中心”的现象。

1.3　防御策略的基本原理

对抗攻击会极大地降低神经网络的准确率，并且指出了神经网络的弱点。因此，人们希望设计针对对抗攻击的防御方法以增强神经网络的性能。防御策略一般可以分为四类：对图像进行滤波[18-23]、修改模型结构[24-28]、对抗训练[29-32]以及特征与网络分析[33-34]。

1.3.1　图像滤波

常见的对抗攻击方法是通过在原始图像上添加精心设计的扰动实现的。由于这种扰动在某种程度上表现得像噪声，因此，可以通过对对抗样本进行去噪，使其更接近于原始样本，即去掉生成对抗样本过程中加入的噪声，将其尽可能恢复成原始样本，从而实现对对抗样本的准确分类。从流形学习的角度理解，滤波防御是通过去噪试图将对抗样本拉回到干净样本所在的子空间。

文献[18-21]通过对图像进行压缩以达到去噪目的；文献[22]通过一个去噪网络消除对抗样本中的扰动；文献[23]通过构造一个低维拟自然图像空间将对抗样本投影到自然图像空间中。图像滤波的方法不改变网络自身的结构，不需要重训练，但图像滤波没有在本质上提升神经网络的认知能力，因此，其防御效果有限。当攻击在图像上的变化幅度较小或者高频信息较小的时候，单纯的滤波难以区分图像细节与对抗噪声，使得这类防御方法会影响网络的识别精度。

1.3.2　修改模型结构

由于大多数对抗攻击算法是基于梯度来生成对抗样本，因此通过修改模型隐藏或限制网络的梯度是一种有效的防御方法。文献[24-26]通过添加新的单元或在网络中引入随机性以隐藏网络的梯度；文献[27-28]通过知识蒸馏和梯度正则化等方法限制网络的梯度，给对抗样本的生成带来困难。对模型的修改在某种程度上提升了网络的认知性能，但存在需要针对特定网络特定攻击进行重训练的问题，其防御效率有待进一步提升。

1.3.3　对抗训练

通过利用对抗样本进行对抗训练以增加网络的鲁棒性是一种直接而有效的方式。按照对抗样本的获取方式，对抗训练可以分成直接训练[29]、集成训练[30]、生成模型训练[31-32]等。对抗训练对于增强网络性能具有重要的意义。由于通过对抗攻击，能够生成错误样本，从而可以更高效地提升已得到神经网络的性能，因此在小数据学习方面具有很好的应用前景。

1.3.4　特征与网络分析

原始样本与对抗样本在图像或网络响应方面特征的不同可以用来检测对抗样本。文献[33]提出一种利用隐写分析来检测对抗样本的方法；文献[34]提出了基于有效路径的对抗样本检测方法。这类方法与神经网络的分析相结合，有可能从本质上提升网络的性能，但目前的

分析往往只在较为简单的网络和特定的攻击起作用，对于各类新型攻击的防御效果需要进一步研究。

2 不同的对抗攻击任务

在前述对抗攻击描述的基础上，本节将根据不同的被攻击对象，介绍对分类器的攻击、对检测器的攻击、对编码器的攻击，以及针对小样本学习和在线学习的数据投毒攻击方式。

2.1 对分类器的攻击

针对分类器的攻击一般是通过修改原始图像使分类器产生错误的分类结果以达到攻击目的，其中又分为目标攻击与非目标攻击。目标攻击指的是使分类器将对抗样本错误分类至指定的类别，而非目标攻击只需要分类器分类错误即可，对具体类别没有要求。根据是否知道被攻击对象的信息，对抗攻击分为白盒攻击和黑盒攻击，其中白盒攻击指完全知道网络的结构、参数等信息，而黑盒攻击指不知道网络的具体信息，而直接使用对抗样本欺骗神经网络。

2.1.1 FGSM 及其变种

FGSM[8]是一种基于梯度生成对抗样本的算法，通过最大化损失函数以获取对抗样本，沿着梯度增加的方向进一步生成对抗样本：

$$\boldsymbol{x}^{*}=\boldsymbol{x}+\eta \operatorname{sgn}\left(\nabla_{x} J(\boldsymbol{x}, \boldsymbol{y})\right) \tag{3}$$

式中：J 为分类算法中衡量分类误差的损失函数；$\boldsymbol{x}$ 为原始样本；$\boldsymbol{y}$ 为原始样本对应的正确分类；$\operatorname{sgn}(\cdot)$ 为符号函数；η 为攻击步长。最大化 J 使得添加噪声后的样本偏离 $\boldsymbol{y}$ 类，由此完成非目标攻击。

单纯的 FGSM 仅考虑导数的符号，且只进行一次的扰动，其对模型的攻击效果往往较为有限，Basic Iterative Method[35]基于 FGSM 进行改进，本质上是对前述算法的多次应用，使用一个小的步长进行多次迭代。

文献[36]借鉴优化的思想，在梯度迭代的基础上引入了动量，通过将动量项整合到攻击的迭代过程中，可以稳定更新方向，并在迭代过程中摆脱不良的局部最大值，以获得具有更好迁移性的对抗样本。

2.1.2 Jacobian - based Saliency Map Attack

在对抗攻击相关文献中，为保证添加的扰动不会被人察觉，通常会使用 l_{∞} 范数或 l_2 范数限制扰动的大小，文献[9]提出的 JSMA 方法使用 l_0 范数约束添加扰动的大小，即只改变几个像素的值，而不是对整张图像进行修改。通过计算图像中每个像素的导数，可以找到对于模型判断影响较大的像素点，从而可以更改较少的像素点完成攻击任务。

2.1.3 Deepfool

Deepfool[10]是一种基于超平面分类思想的生成对抗样本的方法。在二分类问题中，超平面是实现分类的基础，若需要改变分类器对某个样本的分类结果，最小的扰动就是将该样本移至超平面上，这种操作的距离代价最小，对于多分类问题也是如此。

在计算对抗样本过程中，Deepfool 将位于分类边界内的图像逐步推到边界外，直至分类结果出现错误，相较于 FGSM，该算法可以通过更小的扰动达到对抗攻击的效果。

2.1.4 C&W 攻击

C&W[37]基于优化如下的目标函数实现攻击：

$$\min_{w_n}\|\boldsymbol{r}_n\|+cf\left(\frac{1}{2}(\tanh(\boldsymbol{w}_n)+1)\right)$$

$$\boldsymbol{r}_n=\frac{1}{2}(\tanh(\boldsymbol{w}_n+1))-\boldsymbol{X}_n$$

$$f(\boldsymbol{x}')=\max(\max\{Z(\boldsymbol{x}')_i: i\neq t\}-Z(\boldsymbol{x}')_t,-k) \tag{4}$$

式中：$\boldsymbol{r}_n$为添加的扰动，通过将对抗样本映射到空间，使其可以在 $-\infty$ 至 $+\infty$ 做变换，更有利于优化；优化目标函数的第二部分中，$Z(\boldsymbol{x})$为样本 $\boldsymbol{x}$ 通过模型未经过 Softmax 的输出向量，其最大值对应的就是该样本分类的类别；k 为置信度，越大的 k 代表模型以越高的置信度识别错误；超参数 c 用来平衡两个损失函数之间的相对关系。通过最小化该损失函数即可将分类类别拉至目标类别，从而实现目标攻击。

作为基于优化的攻击方法，C&W 攻击通过改变可以调节置信度，同时对添加扰动的大小进行抑制，生成的扰动更小，但该方法的速度较慢。

2.1.5 Zeroth Order Optimization(ZOO)

ZOO[38]是一种经典的黑盒攻击，无需知晓网络内部参数，通过对图像的像素点逐步添加一个小的扰动，根据模型输出的逻辑值的变化估计其对每个像素的梯度。在估计所得梯度的基础上，直接使用白盒的 C&W 进行攻击。

2.1.6 One pixel attack

作为一种极端的对抗攻击方法，One pixel attack 仅改变图像中的一个像素值以实现对抗攻击[39]。该算法采用差分进化算法，针对每个像素迭代地修改生成子图像，并与母图像对比，根据选择标准保留攻击效果最好的子图像，从而实现对抗攻击。该攻击无需获得网络结构与内部参数或梯度的任何信息，属于黑盒攻击。

2.1.7 仅基于分类结果的攻击

文献[40]提出了一种只基于样本的类别标签边界攻击(boundary attack)。在算法迭代过程中，样本由初始化的图像逐渐向原始样本靠近，直至寻找到决策边界，并在决策边界上找到与原始样本最近的对抗样本。这种只根据输入图像的标签信息并基于决策边界的攻击方法相较于其他方法更为简单，但是由于该方法对模型的访问次数巨大，使该方法耗时过长。针对这一局限，boundary attack + +对该算法进行了优化，大大降低该算法的时间成本。

除此之外，还有很多优秀的攻击算法[41-43]，在此不再详细介绍。总体而言，对分类器的攻击研究最为充分，也往往是实现其他攻击的基础。

2.2 对检测器的攻击

目前很多分类器是基于卷积神经网络(CNN)构建的，而许多优秀的目标检测算法[44-47]同样使用 CNN 网

络作为前层特征提取网络。因此，许多针对分类器的攻击算法也能够有效地攻击检测器[48-50]。

图 2 是文献[50]中给出的一张对检测器的攻击效果图。左侧是原始图像的分割结果和检测结果(紫色区域为狗)，右侧是对抗样本的分割结果和检测结果(浅绿色区域误认为是火车，粉红色区域误认为是人)。可以看出，尽管人类看不出任何区别，检测器却以很高的置信度分割并检测错误。

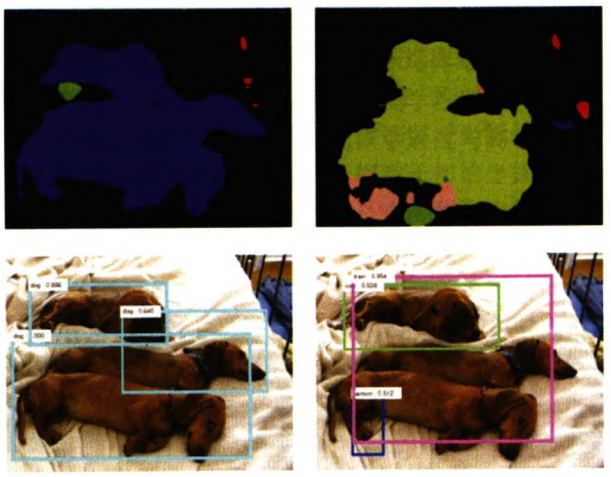


图 2　检测器攻击示例[50]

Fig. 2　An example of attacks on detectors[50]

2.3　对编码器的攻击

自动编码器[51-52]能够将输入数据压缩为维度较小的向量，传递之后可以使用解码器近似地重建原始数据。尽管相对于分类器和检测器，编码器攻击难度较大，但其仍然受到来自对抗样本的威胁。例如，文献[53]对编码器隐变量进行了攻击，使得原始样本与对抗样本的表示向量相似。这是一种 Type Ⅱ攻击示例。

同样地，Type I 对抗攻击[12]能够使对抗样本相对于原始样本产生很大的变化，而其重建之后的结果却与原始样本相似。图 3 显示了在人脸数据库上对编码器的攻击效果[54]：虽然两张图的隐变量差异非常大(这里 *Dev* 表示的是两者之间每个维度上的平均相对差值)，但其解码后的结果却非常相似。换言之，虽然相应的人脸很

(a) *Dev*=156%

(b) *Dev*=361%

图 3　编码器 Type I 攻击示例

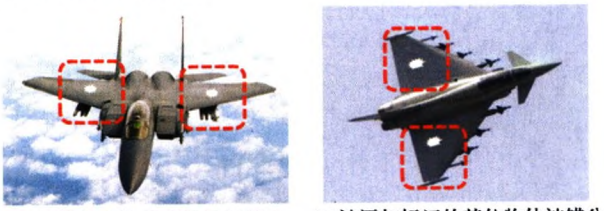
Fig. 3　An example of Type I attacks on encoders

像，但其编码却丧失了相似性，使得编码空间的分类器失效。

2.4　数据投毒

与前面的攻击方式不同，数据投毒[55](也称为特洛伊攻击)主要针对网络的训练过程。具体地，在网络的训练过程中，通过在训练数据的某几个样本中加入小的不易察觉的标记，引导对方分类器以此类标记作为特征进行识别，丧失真实的判别能力。这类攻击尤其对于小样本学习和在线学习等具有很强的破坏作用。

以图 4 展示的机型识别任务为例，数据投毒类攻击在训练样本中添加特别设计的标记，如图 4(a)红框所示。将被投毒的样本送入训练样本库后，会诱导神经网络将该标志物作为样本的重要特征，进而影响其判别能力。当投毒成功后，在被检测物体上添加类似标志，如图 4(b)所示，即使机型已经发生了显著的变化，但这个标志仍然会诱导检测器将其识别为特定的类别。为了视觉效果，图 4 展示的标记较为明显，在实际攻击中，这个标记可以小到不易察觉。投毒攻击可以和对抗攻击(被称为后门攻击)联合使用，即可以通过向训练集投毒增强对抗攻击的成功率。



(a)训练集中被投毒的目标样本　(b)被添加标记的其他物体被错分

图 4　数据投毒示例

Fig. 4　An example of data poison

3　生成攻击的发展趋势

对抗攻击这一概念被提出后，就成为热门研究领域[56]，涌现出许多有前景的研究。时至今日，对抗攻击有以下几个发展趋势：从对图像的攻击到对特征的攻击、从白盒攻击到黑盒攻击、从数字攻击到物理攻击。

3.1　从对图像的攻击到对特征的攻击

目前，大多数对抗攻击算法都是集中在图像空间中，通过在图像上叠加噪声来欺骗神经网络，以此为基础的许多攻击以及防御算法都取得了很好的效果。不过，将对抗攻击与图像特征和语义信息联系起来，将有助于更好地分析图像特征和网络机理[12,57]，由此提出了一种新的攻击思路——基于特征的攻击。基于特征的攻击并不是简单地在图像上添加随机噪声，而是添加具有语义的扰动，从而更难被目前主流的防御算法所防御。

3.2　从白盒攻击到黑盒攻击

尽管目前对抗攻击对深度神经网络提出了严峻挑战，但在实际应用中人们却并不是那么担心，因为目前的大多数攻击算法为白盒攻击算法，需要获得网络的结构、梯度等信息来生成对抗样本，而在实际应用中这几乎是不可能的，因此，白盒攻击并不会造成较大的威胁，攻击者需要借助于黑盒攻击才有可能实现稳定的攻击。

黑盒攻击指的是攻击者不需要获得被攻击对象的具体信息，而直接进行攻击。具体可以分为基于查询的攻击[58-59]和基于迁移的攻击[60-61]。基于查询的攻击需要多次访问被攻击的网络以估计梯度从而实现攻击，但在实际应用中这种查询很容易被防御系统检测到。基于迁移的攻击先在一个参考网络上实现白盒攻击，生成对抗样本，再迁移到被攻击网络上。目前这种方式往往依赖于参考网络和被攻击网络的相似性，较高的迁移率需要二者有较高的相似度，然而这与黑盒攻击的思想相违背。黑盒攻击这一任务不仅是对抗攻击的发展趋势之一，同时也十分具有挑战性。

3.3 从数字攻击到物理攻击

即使实现了黑盒攻击，对抗攻击若想真正威胁到现实中的神经网络还需要突破最后一个障碍，那就是从数字攻击转变为物理攻击[35, 62-63]。图5展示了文献[63]中实现的一个物理攻击的例子。左侧的人被成功检测，而右侧贴有对抗噪声的人则成功“隐形”。与数字攻击不同，这里并不是在图片上施加攻击，而是真实打印出了用于攻击的图案。

图5 物理攻击示例[63]

Fig. 5 An example of physical attacks[63]

目前的绝大多数对抗攻击是将获取的原始图片转换为对抗样本，即图片的数字信息被改变。然而在实际应用，例如安保检测中，入侵监控摄像头更改被攻击系统的数字信息十分困难，并且被检测物体往往处在移动之中，这对对抗攻击的稳定性也提出了挑战。除此之外，对抗图片相对于原始图片的噪声不仅仅添加在被检测物体上，同时也会添加在背景中，在具体的攻击场景中，为整个环境添加噪声几乎是不可能完成的任务。物理攻击的实现会对目前许多深度学习的应用产生巨大的威胁，除了人们日常生活中的应用之外，对高可靠性的军事应用威胁更甚。例如，在目标检测中，通过添加对抗噪声来实现在神经网络眼中的“隐形战机”或者对地面目标进行“隐形”，或者通过数据投毒等方式破解高保密性的人脸识别系统等。因此，如何在物理世界中实现对神经网络的高效攻击是未来对抗攻击的研究热点之一。

4 防御方法及对神经网络的提升

4.1 预处理与图像压缩

对于高维的分类任务，训练集数据往往处于一个复杂的低维子空间中，而对抗样本则不处于该子空间内，如果可以将这些对抗样本映射到该子空间中，也就可以实现防御对抗攻击的效果。然而在实际任务中，往往很难确定该子空间，但可以尝试一个找到包含该子空间的低维空间。

4.1.1 图像压缩

文献[18-19]指出，JPEG空间是一个有效的低维空间，将对抗样本转换为JPEG格式可以一定程度上抵御对抗攻击；文献[20]提出了面向深度神经网络的JPEG压缩方法来抵御对抗样本(“特征蒸馏”)，通过重新设计标准的JPEG压缩算法，以达到最大化提高防御效率同时保证DNN测试准确率的效果；文献[21]提出了ComDefend防御方法，利用图像压缩网络来消除对抗扰动或打破对抗扰动的结构。ComDefend使用两个网络先将图片进行压缩以去除对抗样本中的噪声信息，再重建以获得清晰的图片。ComDefend针对清晰图像进行训练，在训练阶段不需要对抗样本，因此降低了计算成本。

4.1.2 去噪网络

文献[22]通过添加外部模型作为附加网络来针对对抗样本进行去噪预处理，论文首先提出一种以像素为导向的去噪器(pixel guided denoiser, PGD)，希望最小化对抗样本去噪后的图像与原始样本之间的差异。但由于去噪器难以完全消除扰动，剩下的微小扰动仍然会逐层放大，最终导致网络的错误输出。针对此问题，论文又提出了一种以高级表示为导向的去噪器(HGD)，与PGD不同，将去噪后的图片与原始图片都输入到预训练好的深度神经网络模型中，将最后几层的高级特征的差异作为损失函数来训练去噪器，有效避免了PGD的扰动逐层放大的问题。

4.1.3 卷积稀疏编码

文献[23]在卷积稀疏编码的基础上，构造了一个分层的低维拟自然图像空间，该空间在消除对抗扰动的同时逼近自然图像空间。通过在输入图像和神经网络第一层之间引入一种新的稀疏变换层(Sparse Transformation Layer, STL)，可以有效将对抗样本投影到拟自然图像空间中。

4.2 修改模型

4.2.1 在原模型上添加新的单元

文献[24]在初始网络结构的基础上，添加一个利用非局部平均(non-local means)与滤波器对特征进行降噪，利用对抗训练实现对对抗样本的防御。

4.2.2 引入随机性

文献[25]提出一种随机多样化机制作为防御对抗攻击的策略。该方法在网络中引入了一种多通道的结构，各个通道在训练与测试阶段采用不同的随机策略，以达到防御对抗攻击的目的。

文献[26]提出了PNI(Parametric Noise Injection)方法，通过将高斯噪声注入到神经网络每一层的激活和权重中提高网络的随机性。但在网络训练的过程中，除了训练每一层的权重，还要训练噪声的参数，加重了训练负担。

4.2.3　防御性蒸馏

文献[27]基于训练深度神经网络的蒸馏法提出了防御性蒸馏，以提高模型鲁棒性。该方法希望将训练好的复杂模型学习到的“知识”迁移到一个结构更为简单的网络中，或者通过简单的网络去学习复杂模型中的“知识”。其具体思路是：首先根据原始训练样本 X 和标签 Y 训练一个初始的深度神经网络 $F(X)$，然后利用样本 X 与 $F(X)$作为新的标签训练一个蒸馏网络，得到新的概率分布 $F^d(X)$，最终利用整个网络进行分类或预测。这样可以使网络的决策边界更加平滑，有效防御基于梯度产生的对抗样本。

4.2.4　梯度正则化

神经网络的输出对输入的梯度幅度过大是造成其过于敏感的原因，因此文献[28]使用梯度正则化来提升网络的对抗鲁棒性。在训练深度神经网络的过程中，惩罚输出相对输入的变化幅度，使输出对于输入的敏感性降低，从而达到隐藏梯度的效果，但带来了更大的计算量。使用梯度正则化的逻辑基础在于输入和输出之间的连续性，由于分类问题的标签是量化后的结果，已经丧失了连续性。因此，直接使用梯度正则化对分类任务并不特别合适，而是适用于编码器等本身输出与输入之间连续性较好的任务。图 6(a) 显示了对编码器进行攻击的结果，虽然其输入图像视觉效果保持不变，但其重建图像发生了显著的变化，如图 6(b) 所示。在编码器训练时加入梯度正则化可以显著增强所得结果的稳健性[64]。此时，为达到图 6(b) 所示的重建目标，输入图像本身就已经变得和目标类别很相近(如图 6(c) 所示)，避免了对抗攻击。

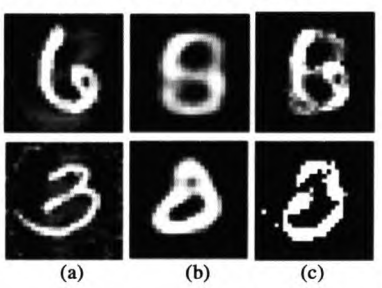


图 6　梯度正则化能够提升编码器稳健性

Fig. 6　Double-BP for robustness encoder

4.3　对抗训练

对抗训练指的是利用对抗样本对网络进行对抗训练。随着网络所接受的对抗样本数量的增加，网络对于对抗样本的鲁棒性也就越强，同时网络对于干净样本的分类正确率也往往会增加[29]。

最直接的方法是使用针对训练网络产生的对抗样本进行训练[29]。另外，还可以采用集成训练方法[30]，使用其他模型产生的对抗样本来扩充本模型的训练集，从而增强训练模型的黑盒鲁棒性。除此之外，还可以采用生成模型来产生对抗样本进行训练，文献[31]基于 AC - GAN 产生无约束(非噪声)的对抗性样本；文献[32]提出了 AdvGAN 模型，使用生成对抗网络对图片生成对抗扰动。采用生成模型进行对抗训练的方法受限于生成模型对数据集的拟合能力与生成图像的能力，并且生成模型仍然是学习训练集的分布，无法突破训练集的约束，对未知对抗样本的鲁棒性仍有待考证。

虽然对抗训练取得了一定效果，但是对抗训练在训练过程中不仅需要干净的训练样本，同时也需要大量的对抗样本，极大地增加了所需的计算资源。同时，文献[65]指出，即使是经过对抗训练的网络，也能有效计算出针对该网络的新对抗样本。

4.4　特征分析与网络分析

除了以上方法，还可以分析图像特征或者分析网络响应，进而对对抗样本进行检测。

隐写分析是指在已知或未知嵌入算法的情况下，从观察到的数据检测判断其中是否存在秘密信息，分析数据量的大小和数据嵌入的未知，并最终破解嵌入内容的过程。针对图像的对抗攻击与图像上的隐写术都是在像素值上进行扰动，而隐写分析可以有效地检测通过隐写术模拟真实图像中相邻像素之间的依赖关系进行的修改，所以也可以利用隐写分析来识别对抗攻击造成的偏差。文献[33]提出一种利用隐写分析来检测对抗样本的方法，根据隐写分析对图像进行特征的提取，再利用线性分类器对样本是否为对抗样本进行分类。

文献[34]提出了基于有效路径的对抗样本检测方法，其理论基础在于深度神经网络工作时，并不是所有的神经元都会被激活，正常样本和攻击样本所激活的神经元会有所不同，即有效路径不同，由此可以从有效路径的角度出发来分析神经网络。

文献[17]认为，数据中的特征分为稳健特征和非稳健特征，而对抗样本的产生归因于非稳健特征的出现：某些来自数据分布模式的特征具备高度预测性，但对于人类而言是脆弱且难以理解的，模型在对抗攻击下表现的脆弱性是模型对数据中泛化较好的特征具备敏感性的直接结果，这种对抗脆弱性完全是以一种“以人为中心”的现象，因为从标准监督学习的角度来看，非稳健特征和稳健特征具备同等的重要性。因此，如果希望获得具有鲁棒性、解释性的模型，就需要将人类先验知识更好地引入训练过程，而仅通过模型训练难以获得。

4.5　类脑计算、因果计算及其他

一些观点认为，神经网络虽然受人类神经元的启发，但其并不能真正模仿人类思考的过程。因此，一些研究致力于开发新的计算方式以增强模型的泛化能力。

文献[66]提出了一种通用似然比方法，该方法能够使用一些类似于大脑的生物机制来训练人工神经网络；文献[67]受神经回路中非线性树突计算基础的生物物理原理启发，证明了神经网络对于高度非线性激活的对抗攻击具有天然的鲁棒性；文献[68]分析了 DNN 和人类对模式进行分类的方式之间的差异，提出了密集关联记忆(DAM)模型，神经元之间具有更高阶的相互作用，更能准确地模仿人类的感知；文献[69 - 71]受人类神经元中电脉冲信号的启发，提出脉冲神经元，其输入输出全部都是脉冲信号(例如 010100…)，更接近于真实的生物

神经模型；文献[72－74]则从贝叶斯推理、图卷积神经网络等角度，重新考虑了数据之间的因果关系，希望神经网络能够具有推理能力，以训练出更加鲁棒的模型。

5 结 论

本文介绍了针对神经网络的对抗攻击的基本概念和原理，梳理了经典的对抗攻击算法和防御算法，分析了对抗攻击的未来发展趋势。对抗攻击在近年来得到了很多关注，在通过攻击发现神经网络共有缺陷、通过防御增强神经网络的性能等方面仍然有许多值得探索的问题。对抗攻击并不是为了否定深度学习和神经网络，而是为了帮助神经网络抵御未知的恶意攻击，帮助人类更好地了解神经网络背后的数学原理，揭示其工作过程，进而训练出更稳健、更具解释性的模型，以达到提升神经网络认知水平的最终目的。

参考文献：

[1] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large－Scale Image Recognition [EB/OL]. (2015－04－10)[2020－01－19]. https://arxiv.xilesou.top/pdf/1409.1556.pdf.

[2] He K M, Zhang X Y, Ren S Q, et al. Deep Residual Learning for Image Recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 770－778.

[3] Huang G, Liu Z, Van Der Maaten L, et al. Densely Connected Convolutional Networks[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 4700－4708.

[4] 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.

[5] Shen D G, Wu G R, Suk H I. Deep Learning in Medical Image Analysis[J]. Annual Review of Biomedical Engineering, 2017, 19: 221－248.

[6] Bojarski M, Del Testa D, Dworakowski D, et al. End to End Learning for Self－Driving Cars[EB/OL]. (2014－04－25)[2020－01－19]. https://arxiv.xilesou.top/pdf/1604.07316.pdf.

[7] Tian Y C, Pei K X, Jana S, et al. Deeptest: Automated Testing of Deep－Neural－Network－Driven Autonomous Cars[C]//Proceedings of the 40th International Conference on Software Engineering, 2018: 303－314.

[8] Goodfellow I, Shlens J, Szegedy C. Explaining and Harnessing Adversarial Examples[C]//International Conference on Learning Representation(ICLR), 2015.

[9] Papernot N, McDaniel P, Jha S, et al. The Limitations of Deep Learning in Adversarial Settings[C]//IEEE European Symposium on Security and Privacy (EuroS&P), 2016: 372－387.

[10] Moosavi－Dezfooli S M, Fawzi A, Frossard P. Deepfool: A Simple and Accurate Method to Fool Deep Neural Networks[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 2574－2582.

[11] Nguyen A, Yosinski J, Clune J. Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015: 427－436.

[12] Tang S L, Huang X L, Chen M J, et al. Adversarial Attack Type I: Cheat Classifiers by Significant Changes[J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence. DOI: 10.1109/TPAMI.2019.2936378.

[13] Gilmer J, Metz L, Faghri F, et al. Adversarial Spheres[EB/OL]. (2019－08－12)[2020－01－19]. https://arxiv.org/pdf/1801.02774v2.pdf.

[14] Schmidt L, Santurkar S, Tsipras D, et al. Adversarially Robust Generalization Requires More Data[C]//Conference on Neural Information Processing Systems(NIPS), 2018: 5014－5026.

[15] Bubeck S, Price E, Razenshteyn I. Adversarial Examples from Computational Constraints[EB/OL]. (2018－05－25)[2020－01－19]. https://arxiv.org/pdf/1805.10204.pdf.

[16] Shafahi A, Huang W R, Studer C, et al. Are Adversarial Examples Inevitable? [EB/OL]. (2018－09－06)[2020－01－19]. https://arxiv.org/pdf/1809.02104v1.pdf.

[17] Ilyas A, Santurkar S, Tsipras D, et al. Adversarial Examples are Not Bugs, They are Features[EB/OL]. (2019－08－12)[2020－01－19]. https://arxiv.org/pdf/1905.02175v3.pdf.

[18] Dziugaite G K, Ghahramani Z, Roy D M. A Study of the Effect of JPG Compression on Adversarial Images[EB/OL]. (2016－08－02)[2020－01－19]. https://arxiv.org/pdf/1608.00853.pdf.

[19] Das N, Shanbhogue M, Chen S T, et al. Keeping the Bad Guys Out: Protecting and Vaccinating Deep Learning with JPEG Compression[EB/OL]. (2017－05－08)[2020－01－19]. https://arxiv.org/pdf/1705.02900.pdf.

[20] Liu Z H, Liu Q, Liu T, et al. Feature Distillation: DNN－Oriented JPEG Compression Against Adversarial Examples [EB/OL]. (2018－03－14)[2020－01－19]. https://arxiv.org/pdf/1803.05787.pdf.

[21] Jia X J, Wei X X, Cao X C, et al. ComDefend: An Efficient Image Compression Model to Defend Adversarial Examples[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 6084－6092.

[22] Liao F Z, Liang M, Dong Y P, 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.

[23] Sun B, Tsai N, Liu F C, et al. Adversarial Defense by Stratified Convolutional Sparse Coding[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 11447－11456.

[24] Xie C H, Wu Y X, Van Der Maaten L, et al. Feature Denoising for Improving Adversarial Robustness[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 501－509.

[25] Taran O, Rezaeifar S, Holotyak T, et al. Defending Against Adversarial Attacks by Randomized Diversification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 11226－11233.

[26] Rakin A S, He Z, Fan D. Parametric Noise Injection: Trainable Randomness to Improve Deep Neural Network Robustness Against Adversarial Attack[EB/OL]. (2018－11－22)[2020－01－

19]. https://arxiv.org/pdf/1811.09310.pdf.

[27] Papernot N, McDaniel P, Wu X, et al. Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks[C]// IEEE Symposium on Security and Privacy (SP), San Jose, USA, 2016: 582 - 597.

[28] Ross A S, Doshi - Velez F. Improving the Adversarial Robustness and Interpretability of Deep Neural Networks by Regularizing Their Input Gradients[C]//Thirty - Second AAAI Conference on Artificial Intelligence, 2018: 1660 - 1669.

[29] Sharif M, Bauer L, Reiter M K. On the Suitability of L_p - Norms for Creating and Preventing Adversarial Examples[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018: 1605 - 1613.

[30] Tramèr F, Kurakin A, Papernot N, et al. Ensemble Adversarial Training: Attacks and Defenses[EB/OL]. (2017 - 05 - 30)[2020 - 01 - 19]. https://arxiv.org/pdf/1705.07204.pdf.

[31] Song Y, Shu R, Kushman N, et al. Constructing Unrestricted Adversarial Examples with Generative Models[C]// Conference on Neural Information Processing Systems (NIPS), 2018: 8312 - 8323.

[32] Xiao C W, Li B, Zhu J Y, et al. Generating Adversarial Examples with Adversarial Networks[EB/OL]. (2018 - 01 - 15)[2020 - 01 - 19]. https://arxiv.org/pdf/1801.02610.pdf.

[33] Liu J Y, Zhang W M, Zhang Y W, et al. Detection Based Defense Against Adversarial Examples from the Steganalysis Point of View[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2019: 4825 - 4834.

[34] Qiu Y X, Leng J W, Guo C, et al. Adversarial Defense Through Network Profiling Based Path Extraction[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019: 4777 - 4786.

[35] Kurakin A, Goodfellow I, Bengio S. Adversarial Examples in the Physical World[EB/OL]. (2016 - 11 - 01)[2020 - 01 - 19]. https://arxiv.org/pdf/1607.02533.pdf.

[36] Dong Y P, Liao F Z, Pang T Y, et al. Boosting Adversarial Attacks with Momentum[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 9185 - 9193.

[37] Carlini N, Wagner D. Towards Evaluating the Robustness of Neural Networks[C]// IEEE Symposium on Security and Privacy (SP), 2017: 39 - 57.

[38] 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.

[39] Su J W, Vargas D V, Sakurai K. One Pixel Attack for Fooling Deep Neural Networks[J]. IEEE Transactions on Evolutionary Computation, 2019, 23(5):828 - 841.

[40] Brendel W, Rauber J, Bethge M. Decision - Based Adversarial Attacks: Reliable Attacks Against Black - Box Machine Learning Models[EB/OL]. (2017 - 11 - 12)[2020 - 01 - 19]. https://arxiv.org/pdf/1712.04248.pdf.

[41] Baluja S, Fischer I. Adversarial Transformation Networks: Learning to Generate Adversarial Examples[EB/OL]. (2017 - 03 - 28)[2020 - 01 - 19]. https://arxiv.org/pdf/1703.09387.pdf.

[42] Cisse M, Adi Y, Neverova N, et al. Houdini: Fooling Deep Structured Prediction Models[EB/OL]. (2017 - 07 - 17)[2020 - 01 - 19]. https://arxiv.org/pdf/1707.05373.pdf.

[43] Han J F, Dong X Y, Zhang R M, et al. Once a Man: Towards Multi - Target Attack via Learning Multi - Target Adversarial Network Once[C]//Proceedings of the IEEE International Conference on Computer Vision, 2019: 5158 - 5167.

[44] Girshick R, Donahue J, Darrell T, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014: 580 - 587.

[45] Girshick R. Fast R - CNN[C]//Proceedings of the IEEE International Conference on Computer Vision, 2015: 1440 - 1448.

[46] Ren S Q, He K M, Girshick R, et al. Faster R - CNN: Towards Real - Time Object Detection with Region Proposal Networks[C]// Conference on Neural Information Processing Systems (NIPS), 2015: 91 - 99.

[47] Long J, Shelhamer E, Darrell T. Fully Convolutional Networks for Semantic Segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015: 3431 - 3440.

[48] Arnab A, Miksik O, Torr P H S. On the Robustness of Semantic Segmentation Models to Adversarial Attacks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018: 888 - 897.

[49] Metzen J H, Kumar M C, Brox T, et al. Universal Adversarial Perturbations Against Semantic Image Segmentation[C]// Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017: 2774 - 2783.

[50] Xie C H, Wang J Y, Zhang Z S, et al. Adversarial Examples for Semantic Segmentation and Object Detection[C]// Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017: 1369 - 1378.

[51] Bengio Y. Learning Deep Architectures for AI[J]. Famdations and Trends in Machme Learning, 2009, 2(1): 1 - 55.

[52] Doersch C. Tutorial on Variational Autoencoders[EB/OL]. (2016 - 08 - 13)[2020 - 01 - 19]. https://arxiv.org/pdf/1606.05908.pdf.

[53] Tabacof P, Tavares J, Valle E. Adversarial Images for Variational Autoencoders[EB/OL]. (2016 - 11 - 01)[2020 - 01 - 19]. https://arxiv.org/pdf/1612.00155.pdf.

[54] Sun C J, Chen S Z, Cai J, et al. Type I Attack for Generative Models[EB/OL]. (2020 - 03 - 04) [2020 - 03 - 04]. https://arxiv.org/pdf/2003.01872.pdf.

[55] Gu T Y, Dolan - Gavitt B, Garg S. Badnets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain[EB/OL]. (2017 - 08 - 22)[2020 - 01 - 19]. https://arxiv.org/pdf/1708.06733.pdf.

[56] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing Properties of Neural Networks[EB/OL]. (2013 - 11 - 21)[2020 - 01 - 19]. https://arxiv.org/pdf/1312.6199.pdf.

[57] Song Y, Shu R, Kushman N, et al. Constructing Unrestricted Adversarial Examples with Generative Model[C]// Conference on Neural Information Processing Systems (NIPS), 2018: 8312 - 8323.

[58] Cheng S Y, Dong Y P, Pang T Y, et al. Improving Black - Box

Adversarial Attacks with a Transfer - Based Prior [EB/OL]. (2019 - 10 - 30) [2020 - 01 - 19]. http://export.arxiv.org/pdf/1906.06919.

[59] Ilyas A, Engstrom L, Madry A. Prior Convictions: Black - Box Adversarial Attacks with Bandits and Priors[EB/OL]. (2018 - 07 - 20) [2020 - 01 - 19]. https://arxiv.org/pdf/1807.07978v1.pdf.

[60] Papernot N, McDaniel P, Goodfellow I, et al. Practical Black - Box Attacks Against Machine Learning[C]// Proceedings of the ACM Asia Conference on Computer and Communications Security, 2017: 506 - 519.

[61] Dong Y P, Pang T Y, Su H, et al. Evading Defenses to Transferable Adversarial Examples by Translation - Invariant Attacks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2019: 4312 - 4321.

[62] Eykholt K, Evtimov I, Fernandes E, et al. Robust Physical - World Attacks on Deep Learning Models[EB/OL]. (2017 - 09 - 13) [2020 - 01 - 19]. https://arxiv.org/pdf/1707.08945.pdf.

[63] 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.

[64] Sun C J, Chen S, Huang X L. Double Backpropagation for Training Autoencoders Against Adversarial Attack[EB/OL]. (2020 - 03 - 04) [2020 - 03 - 04]. https://arxiv.org/pdf/2003.01895.pdf.

[65] 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 (CVPR), 2017: 1765 - 1773.

[66] Xiao L, Peng Y J, Hong J, et al. Training Artificial Neural Networks by Generalized Likelihood Ratio Method: Exploring Brain - Like Learning to Improve Adversarial Defensiveness [EB/OL]. (2019 - 07 - 11) [2020 - 01 - 19]. https://arxiv.org/pdf/1902.00358.pdf.

[67] Nayebi A, Ganguli S. Biologically Inspired Protection of Deep Networks from Adversarial Attacks[EB/OL]. (2017 - 03 - 27) [2020 - 01 - 19]. https://arxiv.org/pdf/1703.09202.pdf.

[68] Krotov D, Hopfield J. Dense Associative Memory is Robust to Adversarial Inputs[J]. Neural Computation, 2018, 30(12): 3151 - 3167.

[69] Ghosh - Dastidar S, Adeli H. Spiking Neural Networks[J]. International Journal of Neural Systems, 2009, 19(4): 295 - 308.

[70] Van Gerven M, Bohte S. Artificial Neural Networks as Models of Neural Information Processing [J]. Frontiers in Computational Neuroscience, 2017(11): 114.

[71] Xin J G, Embrechts M J. Supervised Learning with Spiking Neural Networks[C]// Proceedings of IEEE International Joint Conference on Neural Networks, 2001: 1772 - 1777.

[72] Kipf T N, Welling M. Semi - Supervised Classification with Graph Convolutional Networks[EB/OL]. (2016 - 11 - 03) [2020 - 01 - 19]. https://arxiv.org/pdf/1609.02907.pdf.

[73] Battaglia P W, Hamrick J B, Bapst V, et al. Relational Inductive Biases, Deep Learning, and Graph Networks[EB/OL]. (2018 - 11 - 17) [2020 - 01 - 19]. https://arxiv.org/pdf/1806.01261.pdf.

[74] Xu K Y L, Hu W H, Leskovec J, et al. How Powerful are Graph Neural Networks? [EB/OL]. (2018 - 12 - 26) [2020 - 01 - 19]. https://arxiv.org/pdf/1810.00826.pdf.

Adversarial Attacks and Defenses Against Neural Networks

He Zhengbao, Huang Xiaolin*

(Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, China)

Abstract: With continuous development of deep learning and neural network, deep neural networks have been widely used in many fields and its security has been paid more and more attention. Adversarial attacks and adversarial samples, as the biggest threats to neural networks, have become a hot topic in recent years. The research on adversarial attacks and its defenses is also helpful to understand and improve the cognitive ability of neural networks. This article introduces the basic principles and some classic algorithms around the adversarial attack and its defense, and explains the significance and role of the adversarial attack and the development trend.

Key words: deep learning; neural network; adversarial attack; adversarial sample; defense algorithm; artificial intelligence