深度学习的对抗攻击方法综述

张嘉楠，王逸翔，刘博，常晓林
(北京交通大学智能交通数据安全与隐私保护技术北京市重点实验室，北京 100044)

摘 要：随着大数据时代的到来，深度学习已经成为当前计算机领域研究和应用最广泛的技术之一，成功应用于数据挖掘、计算机视觉、自然语言处理等领域。虽然深度学习已经在解决复杂问题方面取得了的成功，但是研究表明，其容易受到对抗样本的攻击，导致模型产生不正确的输出，进而影响到实际应用系统的可靠性和安全性。文章回顾了有关深度学习的对抗样本的最新发现，总结了生成对抗样本的攻击方法，最后给出了对抗攻击的未来研究方向。

关键词：深度学习；对抗样本；安全威胁；防御技术
中图分类号：TP309.2 **文献标识码：**A

Survey of adversarial attacks of deep learning

Zhang Jianan, Wang Yixiang, Liu Bo, Chang Xiaolin
(*Beijing Key Laboratory of Security and Privacy in Intelligent Transportation, Beijing Jiaotong University, Beijing 100044*)

Abstract: With the arrival of big data, deep learning has become one of the most widely studied and used technologies in the field of computer, and has been widely applied in data mining, computer vision, natural language processing and other fields. Although deep learning makes great success in solving complex problems, recent studies have shown that it is vulnerable to adversarial examples, resulting in incorrect outputs of deep learning models and affecting the reliability and security of practical application systems based on deep learning eventually. In this article, we review the latest findings of adversarial examples on deep learning, summarize the algorithms to generate adversarial examples, and finally, look forward to the next step of adversarial attacks of deep learning.

Key words: deep learning; adversarial examples; security threat; defense technology

1 引言

深度学习又迎来新的一波发展热潮，推进人工智能向前迈进一大步，并在广泛的应用中取得了卓越的进展，例如生物科学[1]、计算机视觉、语音识别[2]、自然语言理解[3]和恶意软件检测[4]等。

2013年，Szegedy等人[5]首先通过添加轻微扰动来干扰输入样本，使基于深度神经网络（Deep Neural Network, DNN）的图片识别系统输出攻击者想要的任意错误结果。研究人员称，这类使模型错误的输入样本为对抗样本（Adversarial Example），此过程称对抗攻击（Adversarial Attack）。研究人员表明，现代深度神经网络模型极易受到人类视觉系统几乎无法察觉的微小扰动的对抗攻击。这种攻击可以造成神经网络分类器对原始图像进行错误预测。更糟糕的是，受到攻击的模型对输出的错误预测结果表示高度信任，而且同样的图像扰动可以欺骗多个不同的分类器。研究表明，对

抗样本可以应用于现实世界。例如，敌手可以构建物理对抗样本，并通过操纵交通标志识别系统中的停车标志或在物体识别系统中移除行人来欺骗自动驾驶车辆。随着深度学习应用领域的不断深入和扩大，深度学习暴露的安全性问题受到了更为广泛的关注。

2018年，N. Akhtar等人[6]对计算机视觉中使用深度学习所面临的对抗攻击进行了较为详尽的研究。.本文以该综述为蓝图，对后续的研究成果进行整理总结和补充，提出了新的依据白盒和黑盒环境划分对抗样本攻击方法的分类方式，总结出了深度学习系统的安全威胁的现状。早期的研究主要针对传统机器学习（例如支持向量机、朴素贝叶斯等学习方法），但目前大量的安全威胁主要针对DNN模型。本文首先简要地回顾了针对传统机器学习的对抗攻击，然后总结了针对深度学习的攻击，根据敌手知识和对抗特异性，对对抗攻击方法进行了细致的分类，分析总结了各种攻击技术的研究思路和进展。已有文章证明，针对DNN模型生成的对抗样本在传统机器学习模型中仍然有效[7]。

2 背景

2.1 深度学习技术

机器学习是一门多领域交叉学科，主要研究如何更好地让计算机模拟和实现人类的学习行为，从而实现知识的自动获取和产生。机器学习解决问题的过程分为训练和预测两个阶段。

深度学习是机器学习中一种基于对数据进行表征学习的方法，其动机在于建立、模拟人脑进行分析学习的神经网络，它模仿人脑的机制来解释数据。

深度神经网络（DNN）是典型的深度学习模型，其强表达能力使其在语音识别、面部识别和计算机视觉等领域取得了巨大的成功。卷积神经网络（Convolutional Neural Network, CNN）和循环神经网络（Recurrent Neural Network, RNN）在DNN基础上扩展而来。

深度学习解决某些复杂问题的能力已经超出了人类水平，但研究表明，深度学习技术也面临着多种安全性威胁。自Szegedy等人[5]首先通过对输入样本添加轻微扰动来欺骗DNN网络以来，越来越多的研究发现，除了DNN模型之外，生成的对抗样本同样能成功地攻击强化学习模型、循环神经网络（RNN）模型等其他深度学习模型。

2.2 术语定义

本节描述了在面向深度学习系统的对抗攻击相关文献中使用的常用技术术语。

（1）对抗性扰动：添加到原始样本中使其成为对抗样本的噪声。

（2）欺骗率：指一个经过训练的模型在受到干扰后改变其预测标签的图像百分比。

（3）可迁移性：对抗样本可以对生成模型以外的模型进行有效的攻击。

（4）普遍扰动：能够高概率地在任意图像上欺骗给定模型。

（5）敌手知识（Adversary’s Nnowledge）：敌手更多的是指生成对抗样本的代理人。在某些情况下，这个对抗样本本身也被称为敌手。敌手知识包括模型的训练数据、特征集合、模型结构及参数、学习算法及决策函数、目标模型中可用的反馈信息等。根据敌手掌握机器学习模型信息的多少可将攻击分为白盒攻击和黑盒攻击。

1）白盒攻击：攻击者完全了解目标模型，包括模型的结构及参数值、特征集合、训练方法，在某些情况下还包括其训练数据。

2）黑盒攻击：攻击者在不知道机器学习模型的内部结构、训练参数和算法的情况下，通过传入输入数据来观察输出、判断输出与模型进行交互。在一些情况下，假设敌手具有对模型的有限知识（例如其训练过程或其架构），仅了解模型的一部分，但并不知道模型参数。

（6）对抗特异性（Adversarial Specificity）：按照攻击的专一性及目的，可以将对抗样本的攻击分为针对目标攻击和非针对目标攻击。

1）针对目标攻击：攻击者在构造对抗样本时欺骗目标模型，将对抗样本错分到指定分类类

别。针对目标攻击通常发生在多类分类问题中。

2）非针对目标攻击：对抗样本的预测标记是不相关的，只需让目标模型将其错误分类，即除了原始类别，对抗类输出可以是任意的。

（7）扰动测量（Perturbation Measurement）：优化扰动是指将扰动设置为优化问题，旨在最小化扰动，使得人类无法识别扰动。

1）L_p通过p-norm距离估计对抗样本与原样本的差距。L_p的定义如下所示：

$$\|x\|_p = \left(\sum_{i=1}^{n} \|x_i\|_p\right)^{\frac{1}{p}}$$

A. $L_\infty : \|x\|_\infty = \max_i |x_i|$

L_∞范数表示对抗样本中所有像素的最大变化。

B. $L_2 : \|x\|_2 = \sqrt{\sum_i x_i^2}$

L_2范数用于测量对抗样本与原始样本的欧几里德距离。当对许多像素应用许多小的改变时，该距离可以保持很小。

C. $L_0 : \|x\|_0 = \#\{i | x_i \neq 0\}$

L_0计算样本中更改的像素数，而不是扰动量。

L_p距离越小，表明对抗样本与原样本的差距越小，对抗样本中扰动越不易察觉。

2）PASS（Psychometric Perceptual Adversarial Similarity Score）值是Hot/Cold方法[8]定义的一个新的度量标准，用来衡量攻击前后样本的差异。

（8）攻击频率（Attack Frequency）：按照算法是否需要迭代地求解对抗样本，可将攻击算法分为单步攻击和迭代攻击。

1）单步攻击只需一次即可生成对抗样本。

2）迭代攻击通过迭代生成对抗扰动。与单步攻击相比，迭代攻击通常能生成更好的对抗样本，但需要与目标分类器进行更多交互（更多查询），需要更多的计算时间。

3 对抗攻击

对抗样本最早由Szegedy等人[5]提出，在数据集中通过添加轻微扰动来干扰输入样本，导致模型以高置信度给出一个错误的输出。模型在这个输入点x'的输出与附近的数据点x不同。在许多情况下，x'与x非常近似，人类不会察觉原始样本和对抗样本之间的差异，但是网络会做出非常不同的预测。

3.1 传统机器学习中的对抗样本

早期的研究主要针对传统机器学习模型中的对抗样本，例如垃圾邮件过滤器、入侵检测、生物识别身份验证和欺诈检测[9]，垃圾邮件通常通过添加字符以避免检测[10~12]。

Dalvi等人[10]首先讨论了对抗样本，指出对抗样本的攻击和防御是攻击者与防御者的一种迭代游戏。Biggio等人[13]首先尝试了基于梯度的方法来生成针对线性分类器、支持向量机（Support Vector Machine, SVM）和神经网络的对抗样本。与深度学习产生对抗样本的方法相比，允许更自由地修改数据。Roli等人[14]审查了几个主动防御，并讨论了改善机器学习模型安全性的防御方法。

Barreno等人[9,15]对机器学习的安全问题进行了初步调查，比较针对SpamBayes垃圾邮件过滤器和防御的攻击作为研究案例。但是，它们主要关注二分类问题，如病毒检测系统、入侵检测和防御系统。

传统机器学习中的对抗样本需要提取特征知识，而深度学习通常仅需要原始数据输入。Papernot等人[16]全面概述了机器学习中的安全问题以及深度学习中的最新发现，并建立了统一的威胁模型。

3.2 深度学习攻击分类

本节将回顾深度学习中的文献，介绍深度学习中产生对抗样本的代表性方法。根据攻击者掌握深度学习目标模型的背景知识，将对抗攻击分为白盒攻击和黑盒攻击。考虑到对抗攻击的特异性，根据生成的对抗样本是否需要定位为某一特定类别，进一步可将攻击算法分为针对目标攻击和非针对目标攻击。

3.2.1 白盒攻击

白盒攻击是指攻击者在完全了解神经网络模型的网络结构以及模型参数的情况下，针对该神经网络生成对抗样本。攻击者在产生对抗攻击数据时与机器学习系统有所交互。

（1）非针对目标攻击

1）FGSM

FGSM[17]即快速梯度迭代法（Fast Gradient Sign Method），是一种常见的白盒攻击算法。通过采用在模型输出与目标类别的误差函数对输入向量的梯度方向添加扰动得到对抗性扰动，然后将对抗性扰动添加到原始样本中生成对抗样本。该方法证实深度神经网络在高维空间中的线性特征足以产生对抗样本。

Tramèr等人[18]提出R+FGSM方法，沿着梯度的反方向添加扰动，以此拉大对抗样本与原始输入样本的距离。该方法在构造对抗样本时添加了随机攻击。

2）BIM&ILCM

Kurakin等人将对抗样本应用于现实物理世界，改进了FGSM方法，沿梯度方向采用多个较小的输入变化参数进行迭代攻击，从而提出了Basic Iterative Methods（BIM）[19]。Least Likely Class Iterative Methods（ILCM）[19]是BIM方法的变体，通过用识别概率最不可能的类别（目标类别）代替对抗扰动中的类别变量来生成对抗样本。经证实，由该方法生成的对抗样本能够让Inception v3模型受到严重影响。

3）PGD

PGD（Project Gradient Descent）[20]攻击，即投影梯度下降方法，是FGSM的变体。PGD算法首先在原图附近允许的范围内（球形噪声区域）进行随机初始化搜索，然后进行多次迭代产生对抗样本。PGD是一种典型的一阶攻击，如果防御方法对PGD攻击有效，则该防御方法对其他的一阶攻击也有着很好的防御效果。

Zheng等人[21]在PGD攻击的基础上，给出了PGD的分布优化视图，提出了分布对抗攻击（Distributionally Adversarial Attack, DAA）方法，通过学习最大程度地增加模型泛化风险的对抗性数据分布。实验证明，DAA方法在可证明的防御模型上取得了较好的攻击结果。

4）CPPN EA Fool

Nguyen等人[22]发现了一种新型的攻击，即组合模式生成网络编码的进化算法（Compositional Pattern Producing Network Encoded Evolutionary Algorithms, CPPN EA），该算法生成的对抗样本人类无法识别，但被DNN模型以高置信度（99%）将其错误分类。研究者将此类攻击归类为假正性攻击（False-positive Attack）。如图1所示，显示了假正性（False-positive）对抗样本。Nguyen指出，对于许多对抗图像，CPPN可以像JSMA一样找到改变DNN输出的关键特征。

5）DeepFool

DeepFool算法[23]是一种基于超平面分类思想的对抗样本生成算法，目的是以迭代方式计算可以使分类产生误判的最小扰动。在每次迭代时，算法通过小矢量扰动图像，逐步将位于分类边界内的图像推到边界外，直到图像被误判，累积每次迭代中添加到图像中的扰动以计算最终扰动。分别在MNIST，CIFAR-10和ILSVRC2012等数据集上进行了对比实验。实验表明，与FGSM和JSMA算法相比，DeepFool算法生成的扰动更小，计算时长更短。

6）通用对抗扰动（UAP）

诸如FGSM[17]、DeepFool[23]等方法，只能针对不同的对抗样本添加依赖某一特定样本特征的噪音，生成单张图像的对抗扰动，而通用对抗扰动（Universal Adversarial Perturbations, UAP）[24]可以使添加该扰动的原始图像被误分类为其他类别，生成对任何图像有攻击能力的扰动。Khrulkov等人[25]将通用对抗扰动作为网络的特征映射的雅可比矩阵的奇异向量，这使得仅使用少量图像就可以实现较高的欺骗率。通用对抗扰动在当前流行的深度学习模型中得到了很好的推广。

（2）针对目标攻击

1）FGSM

Kurakin等人[26]提出了FGSM[17]的变体，其攻击目标是使模型输出为原始预测最不可能的类别。对抗扰动中的类别变量用识别概率最小的目标类别代替，再将原始图像减去该扰动，原始图像就变成了对抗样本，并能输出目标类别。

2）L-BFGS

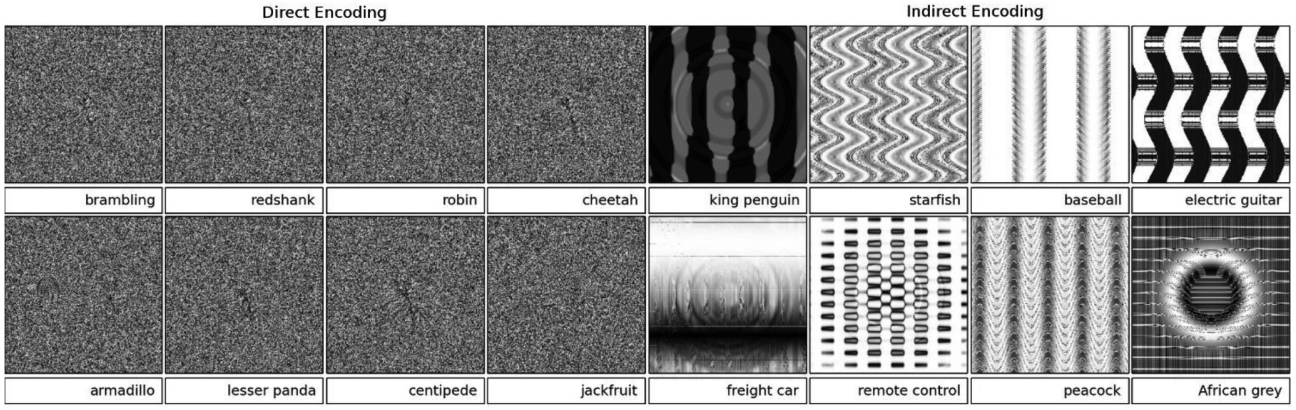

图1 假正性对抗样本

Szegedy等人[5]提出L-BFGS方法，通过计算添加到原始图像上引起神经网络错误分类的扰动来构造对抗样本。Szegedy[5]等人表示，L-BFGS算法生成的对抗样本可以应用到不同的模型和训练数据集中。

3）JSMA

JSMA[27]即Jacobian-based Saliency Map Attack。通过限制扰动的l_0范数来进行对抗攻击。JSMA算法中最重要的两个要素是雅各比矩阵和显著图。计算给定样本χ的雅各比矩阵，由下式给出：

$$J_F(x)==\frac{\partial F_j(x)}{\partial x}$$

该类攻击需要访问可微分的模型，因此属于白盒攻击。

4）C&W

Carlini和Wagner[28]对L-BFGS攻击进一步改善，提出了C&W攻击。该攻击是一种基于迭代优化的低扰动方法，通过限制$\ell_0,\ell_2,\ell_\infty$范数使得扰动无法被察觉。该算法改进其迭代攻击中的目标优化函数，使其逐步收紧对扰动幅度的限制，减少对抗样本的扰动幅度，进而使对抗样本更加难以察觉。实验证明，这三种攻击可以有效地攻击经过“蒸馏”（Defensive Distillation）的网络。

5）Hot/Cold

Rozsa等人[8]提出了一种Hot/Cold方法，该算法可以对每个输入图像生成多个对抗样本。Hot/Cold方法定义了一个新的度量标准PASS来衡量攻击前后样本的差异。PASS包括两个阶段：第一阶段将修改后的图像与原始图像对齐；第二阶段测量对齐后的修改图像与原始图像之间的相似性。其中，Hot和Cold分别代表目标类别和原始类别，在每次迭代后，算法都将样本移向Hot类，远离Cold类。该算法与FGSM相比，其生成的对抗样本具有多样性。

6）对抗转换网络（ATNs）

Baluja和Fischer[29]训练一个生成模型去生成对抗样本。训练的模型被称为对抗转换网络（Adversarial Transformation Networks，ATNs）。这些网络产生的对抗样本是通过最小化由两部分组成的联合损失函数来计算的。联合损失函数的第一部分使对抗样本与原始图像具有感知相似性，第二部分使对抗样本被目标模型错误分类。值得注意的是，ATN可以是针对目标攻击，也可以是非针对目标攻击，并以白盒或黑盒的方式进行训练。Balujar等人[29]的论文专注于有针对性的白盒攻击。

3.2.2 黑盒攻击

本文介绍的攻击方法，攻击者都需要完全了解目标模型的结构和参数。黑盒攻击则代表了更为一般的场景，攻击者可能无法获取到目标模型的全部信息，但是可以利用对抗样本在神经网络模型之间的可迁移性来进行黑盒攻击。本文介绍的FSGM、JSMA、UAP等方法，也可以在黑盒场景下进行有效攻击。

（1）非针对目标攻击

1）FGSM和UAP

Papernot等人[7]观察到对抗样本在模型之间的迁移性，提出使用FGSM算法对未知目标模型

进行黑盒攻击。黑盒攻击依赖于对抗样本的可迁移性，即使是具有不同架构的两个分类器，在其中一个分类器中产生的对抗样本也可能导致另一个分类器对该对抗样本以高置信度做出误判。基于此，Papernot等人[30]训练了一个代理模型来进行分类任务，对代理模型进行白盒攻击构造对抗样本，再使用所生成的对抗样本对目标模型进行黑盒攻击。Dong等人[31]提出U-MI-FGSM方法，将FGSM攻击迭代为多个小步骤扰动，并在每个扰动之后调整扰动方向以达到攻击目的。

Moosavi-dezfooli等人[24]用ImageNet训练了不同的模型，表明使用通用对抗扰动（UAP）算法产生的对抗样本在不同神经网络中具有有效的攻击能力。Li等人[32]通过观察现有的对抗噪声，发现针对对抗训练后的模型设计产生的对抗噪声具有很强的局部相关性，文章提出利用保持对抗噪声的局部相关性来提升对抗样本的攻击性能。文章还指出一个简单的通用扰动可以欺骗一系列最先进的防御，其产生的对抗样本可以很好地在不同的视觉任务中传递，在黑盒环境下取得了良好的效果。

2）单像素攻击

Su等人[33]提出单像素攻击（One Pixel），其目标是在输入图像中选定一个像素，更改其数值产生对抗图像，使神经网络模型对对抗图像错误分类。实验证明，该攻击能够在70.97%的测试图像上成功愚弄三种不同的网络模型。单像素攻击不需要知道网络参数或梯度的任何信息，因此为黑盒攻击。单像素攻击可以是针对目标攻击，也可以是非针对目标攻击。如图2所示，给出使用单像素攻击的攻击示例。

3）ZOO攻击

Chen等人[34]提出ZOO-based（Zeroth Order Optimization）攻击，该算法是一个典型的黑盒攻击算法，可以在没有模型转移的情况下直接部署在黑盒攻击中。通过采用梯度和Hessian矩阵的梯度估计，不需要获取目标模型的梯度信息。然而，它需要昂贵的计算来查询和估计梯度。实验表明，ZOO攻击实现了与C&W攻击相当的性能。值得注意的是，ZOO攻击已被证明在针对目标攻击上也取得了很好的效果。

图2 单像素对抗攻击示例（括号里给出的是模型预测标签）

4）结合GAN的对抗攻击

Zhao等人[35]将生成对抗网络（Generative Adversarial Networks, GAN）的理念结合到对抗样本的生成中，并将此方法命名为Natural GAN。首先在数据集上训练了WGAN模型，其中生成器G将随机噪声映射到输入域。还训练了一个“转换器”来将输入数据映射到密集的内部表示，该方法通过最小化诸如“敌手”之类的内部表示的距离来产生对抗性噪声。由于Natural GAN不需要了解原始神经网络的梯度，因此它可以应用于黑盒攻击。

大多数的黑盒攻击策略是基于对抗样本的可迁移性质。在这些方法中，攻击者首先训练一个本地模型，由本地模型模拟被攻击的模型来生成对抗样本。Xiao等人[36]提出AdvGAN方法，能够不依赖对抗样本的可迁移性进行黑盒攻击。对于AdvGAN，当网络架构中的生成器训练完毕后，对于任何的输入图像，都可以高效地生成对抗样本。应用AdvGAN方法到Madry的MNIST对抗样本生成挑战中，产生的对抗样本在半白盒攻击和黑盒攻击下，分别实现了88.93%和92.76%的攻击成功率。

（2）针对目标攻击

1）FGSM

Dong等人[31]在FGSM算法基础上，提出基于动量的迭代算法T-MI-FGSM来增强对抗攻击。为了进一步提高对抗样本的可迁移性，提高黑盒攻击的成功率，将动量迭代算法应用于一组模型，并证明了具有较强防御能力的神经

网络模型也容易受到此类算法的黑箱攻击。基于动量的迭代算法已被证明在针对目标攻击上也取得了很好的效果。

2）JSMA和ATNs

Papernot等人[7]观察到对抗样本在模型之间的迁移性,提出使用JSMA算法对未知目标模型进行黑盒攻击。

Baluja和Fischer[29]训练ATN来生成对抗样本，沿着同一方向，Hayex和Danezis[37]使用ATN构造对抗样本来进行黑盒攻击。实验结果表明，其生成的对抗样本具有较高的欺骗率。

3）UPSET和ANGRI

Sarkar等人[38]提出了2个黑盒攻击算法，UPSET（Universal Perturbations for Steering to Exact Targets）和ANGRI（Antagonistic Network for Generating Rogue Image）。UPSET使用残差梯度网络，为特定的目标类别产生对抗扰动构造对抗样本，使得分类器将对抗样本分类成目标类别。ANGRI算法生成的是“图像特定”的扰动，其产生的扰动也获得了高欺骗率。

4）Houdini

Cisse等人[39]提出了Houdini算法，这是一种用于欺骗基于梯度的机器学习算法的攻击方法。生成对抗样本的典型算法是使用网络模型的损失函数的梯度来计算扰动，但是任务损失往往不适合这种方法。Houdini算法用来解决组合不可分解的问题，例如语音识别、语义分割等。除了成功生成对抗图像外，Houdini算法还被证明能够成功攻击流行的自动语音识别系统。他们通过在黑盒攻击场景中愚弄Google Voice来证明语音识别中攻击的可转移性。

5）EAD

Chen等人[40]提出一种基于弹性网络正则化的攻击算法（Elastic-net Attacks to DNNs, EAD）。该算法将对抗样本攻击DNN的过程形式化为弹性网络正则化的优化问题。在MNIST、CIFAR10和ImageNet上的实验结果表明，EAD算法可以生成具有很小 失真的对抗样本，并且其对抗样本具有显著增强的攻击可迁移性，能在不同攻击场景中实现与当前最佳方法匹敌的攻击成功率。实验证明了EAD算法的有效性，对基于 的对抗样本和深度神经网络的安全性应用方面提供了新的线索。

6）基于模型的集成攻击

Liu等人[41]在ImageNet数据集的不同模型上进行了可迁移性研究，并研究了非针对性和针对性的对抗样本。与非针对目标对抗样本相比，有针对性的对抗样本更难以在模型之间进行转移。本文提出了基于模型的集成攻击（Model-based Ensembling Attack）以生成可转移的对抗样本，使大部分针对目标对抗样本可以在不用的模型间转递，以此来攻击黑盒模型。

Liu等人[41]提出模型集成的概念，使用联合分类器生成对抗样本，并通过对比实验总结出模型集成方法可以有效提高对抗样本的泛化能力。结果表明，基于模型的集成攻击可以生成可转移的针对目标对抗图像，这增强了黑盒攻击中对抗样本的能力。他们还证明，与以前的方法相比，这种方法在生成非针对目标对抗样本方面表现更好。

如表1所示，总结了生成对抗样本的算法，其中规定了是否为针对目标或非针对目标攻击、扰动测量衡量标准和攻击强度。

4 结束语

本文对深度学习中的对抗样本展开深入调查，根据敌手知识和对抗特异性，对对抗攻击方法进行了细致的分类，清晰地展示了研究人员的研究进展和研究思路。在面向深度学习模型的对抗样本的研究中，对抗攻击的未来研究角度可以由两个方面展开。

（1）研究并设计更有攻击性的对抗样本，作为神经网络鲁棒性的评估标准，可以为对抗防御展开新的研究思路，提高神经网络模型的稳健性。

（2）对抗场景下的现实应用。研究人员将深度学习应用在现实世界中，其实用性和普及性有了巨大提升，也引起了安全领域的极大关注。对抗样本应用在分类器、目标检测器等方面取得了良好的攻击效果，但其在物理世界下的应用还没有得到有效性验证，相关的工作需要在更大的数据集、更真实的场景中进行验证

表1 对抗攻击方法

攻击方式	对抗特异性	攻击算法	扰动测量	攻击强度	文献
白盒攻击	非针对目标攻击	FGSM	ℓ_∞	**	[17]
		R+FGSM	ℓ_∞	***	[18]
		BIM&ILCM	ℓ_∞	****	[19]
		PGD	ℓ_∞	*****	[20]
		CPPN EA Fool	None	***	[22]
		DeepFool	ℓ_2,ℓ_∞	****	[23]
		UAP	ℓ_2,ℓ_∞	*****	[24]
		L-BFGS	ℓ_∞	***	[5]
	针对目标攻击	FGSM	ℓ_∞	**	[26]
		L-BFGS	ℓ_∞	***	[5]
		C&W	$\ell_0,\ell_2,\ell_\infty$	*****	[28]
		Hot/Cold	PASS	****	[8]
		ATNs	ℓ_∞	****	[29]
黑盒攻击	非针对目标攻击	FGSM	ℓ_∞	***	[30]
		U-MI-FGSM	ℓ_∞	****	[31]
		UAP	ℓ_2,ℓ_∞	*****	[24,32]
		One Pixel	ℓ_0	**	[33]
		ZOO	ℓ_2	*****	[34]
		结合GAN的对抗攻击	ℓ_2	*****	[35,36]
	针对目标攻击	JSMA	ℓ_0	***	[7]
		T-MI-FGSM	ℓ_∞	****	[31]
		JSMA	ℓ_0	***	[7]
		One Pixel	ℓ_0	**	[33]
		ZOO	ℓ_2	*****	[34]
		UPSET	ℓ_∞	****	[38]
		ANGRI	ℓ_∞	****	[38]
		Houdini	ℓ_2,ℓ_∞	****	[39]
		EAD	ℓ_1	*****	[40]
		基于模型的集成攻击	ℓ_2	*****	[41]

和完善。

随着深度学习在图像处理、自然语言处理、语音识别、医疗诊断等多个领域的深入应用，深度学习模型面临的安全威胁也日趋严重，现有的深度学习模型极易受到对抗攻击[42]。对抗样本揭示了神经网络的脆弱性和不可解释性，但另一方面，对抗样本的存在也可以激发更多关于对抗防御的研究，从而获得鲁棒性更好的深度学习模型。

基金项目：

国家自然科学基金项目（项目编号：U1836105）。

参考文献

[1] Helmstaedter M,Briggman K L,Turaga S C,Jain V,Seung H S,Denk W.Connectomic reconstruction of the inner plexiform layer in the mouse retina[J]. Nature,2013,500(7461):168.

[2] Hinton G,Deng L,Yu D,Dahl G,Mohamed A R,Jaitly N,Senior A,Vanhoucke V,Kingsbury B,Sainath T. Deep neural networks for acoustic modeling in speech recognition[J].IEEE Signal processing magazine,2012, 29(6): 82-97.

[3] Sutskever I,Vinyals O,Le Q V.Sequence to sequence learning with neural networks[C].Advances in neural information processing systems.2014: 3104-3112.

[4] 刘金鹏.基于机器学习技术的网络安全防护[J].网络空间安全,2018,9(09):96-102.

[5] Szegedy C,Zaremba W, Sutskever I,Bruna J,Erhan D,Goodfellow I,Fergus R.Intriguing properties of neural networks[C].ICLR (Poster).2014.

[6] Akhtar N Mian A.Threat of adversarial attacks on deep learning in computer vision: A survey[J]. IEEE Access,2018,6:14410-14430.

[7] Papernot N,McDaniel P,Goodfellow I. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples[J].arXiv preprint arXiv:1605.07277,2016.

[8] Rozsa A, Rudd E M, Boult T E.Adversarial diversity and hard positive generation[C].Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops.2016:25-32.

[9] Barreno M,Nelson B,Joseph A D,Tygar J D.The security of machine learning[J].Machine Learning, 2010,81(2): 121-148.

[10] Dalvi N, Domingos P,Sanghai S, Verma D.Adversarial classification[C].Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining.ACM, 2004: 99-108.

[11] Lowd D,Meek C.Adversarial learning[C].Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining.ACM, 2005: 641-647.

[12] Biggio B,Fumera G,Roli F.Multiple classifier systems for robust classifier design in adversarial environments[J]. International Journal of Machine Learning and Cybernetics,2010, 1(1-4): 27-41.

[13] Biggio B,Corona I,Maiorca D,Nelson B,Srndic N,Laskov P,Giacinto G,Roli F.Evasion attacks against machine learning at test time[C]. Joint European conference on machine learning and knowledge discovery in databases. Springer,Berlin,Heidelberg,2013: 387-402.

[14] Roli F,Biggio B,Fumera G.Pattern recognition systems under attack[C].Iberoamerican Congress on Pattern Recognition.Springer,Berlin,Heidelberg,2013:1-8.

[15] Barreno M,Nelson B,Sears R,Joseph A D,Tygar J D.Can machine learning be secure?[C].Proceedings of the 2006 ACM Symposium on Information,computer and communications security.ACM,2006:16-25.

[16] Papernot N, McDaniel P,Sinha A, Wellman M.Towards the science of security and privacy in machine learning[J].arXiv preprint arXiv:1611.03814,2016.

[17] Goodfellow I J,Shlens J,Szegedy C.Explaining and harnessing adversarial examples[C].ICLR (Poster).2015.

[18] Tramèr F,Kurakin A,Papernot N,Goodfellow I,Boneh D, McDaniel P.Ensemble adversarial training: Attacks and defenses[C].ICLR (Poster).2018.

[19] Kurakin A,Goodfellow I,Bengio S.Adversarial examples in the physical world[C].ICLR (Workshop). 2017

[20] Madry A,Makelov A,Schmidt L,Tsipras D,Vladu A.Towards deep learning models resistant to adversarial attacks[C].ICLR (Poster).2018.

[21] Zheng T,Chen C,Ren K. Distributionally adversarial attack[C].Proceedings of the AAAI Conference on Artificial Intelligence. 2019,33: 2253-2260.

[22] Nguyen A,Yosinski J,Clune J.Deep neural networks are easily fooled: High confidence predictions for unrecognizable images[C].Proceedings of the IEEE conference on computer vision and pattern recognition.2015: 427-436.

[23] Moosavi-Dezfooli S M,Fawzi A, Frossard P.Deepfool:a simple and accurate method to fool deep neural networks[C].Proceedings of the IEEE conference on computer vision and pattern recognition.2016:2574-2582.

[24] Moosavi-Dezfooli S M,Fawzi A, Fawzi O,Frossard P.Universal adversarial perturbations[C].Proceedings of the IEEE conference on computer vision and pattern recognition. 2017:1765-1773.

[25] Khrulkov V,Oseledets I.Art of singular vectors and universal adversarial perturbations[C].Proceedings of the IEEE Conference on Computer Vision and P attern Recognition. 2018:8562-8570.

[26] Kurakin A,Goodfellow I,Bengio S.Adversarial machine learning at scale[J].arXiv preprint arXiv:1611.01236,2016.

[27] 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, 2016: 372-387.

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

[29] Baluja S,Fischer I. Adversarial transformation networks: Learning to generate adversarial examples[J].arXiv preprint arXiv:1703.09387, 2017.

[30] Papernot N,McDaniel P,Goodfellow I,Jha S,Celik Z B, Swami A.Practical black-box attacks against machine learning[C].Proceedings of the 2017 ACM on Asia conference on computer and communications security. ACM, 2017: 506-519.

[31] Dong Y,Liao F,Pang T,Su H,Zhu J,Hu X,Li J.Boosting adversarial attacks with momentum[C].Proceedings of the IEEE conference on computer vision and pattern recognition.2018: 9185-9193.

[32] Li Y,Bai S, Xie C,Liao Z,Shen X,Yuille A L.Regional Homogeneity: Towards Learning Transferable Universal Adversarial Perturbations Against Defenses[J].arXiv preprint arXiv:1904.00979, 2019.

[33] Su J,Vargas D V,Sakurai K.One pixel attack for fooling deep neural networks[J].IEEE Transactions on Evolutionary Computation, 2019.

[34] Chen P Y,Zhang H,Sharma Y, Yi J, Hsieh C J.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. ACM, 2017: 15-26.

[35] Zhao Z,Dua D,Singh S.Generating natural adversarial examples[J].arXiv preprint arXiv:1710.11342, 2017.

[36] Xiao C,Li B,Zhu J Y,He W,Liu M,Song D.Generating adversarial examples with adversarial networks[J].arXiv preprint arXiv:1801.02610,2018.

[37] Hayes J,Danezis G. Machine learning as an adversarial service: Learning black-box adversarial examples[J]. arXiv preprint arXiv:1708.05207,2017.

[38] Sarkar S,Bansal A, Mahbub U,Chellappa R.UPSET and ANGRI:breaking high performance image classifiers[J]. arXiv preprint arXiv:1707.01159,2017.

[39] Cisse M,Adi Y,Neverova N,Keshet J.Houdini: Fooling deep structured prediction models[J].arXiv preprint arXiv:1707.05373,2017.

[40] Chen P Y,Sharma Y,Zhang H,Hsieh C J.Ead: elastic-net attacks to deep neural networks via adversarial examples[C].Thirty-second AAAI conference on artificial intelligence.2018.

[41] Liu Y,Chen X,Liu C,Song D.Delving into transferable adversarial examples and black-box attacks[J].arXiv preprint arXiv:1611.02770,2016.

[42] 李盼,赵文涛,刘强,崔建京,殷建平.机器学习安全性问题及其防御技术研究综述[J].计算机科学与探索, 2018,12(2): 171-184.

作者简介:

张嘉楠（1996-），女，汉族，山东菏泽人，北京交通大学，在读硕士；主要研究方向和关注领域：网络空间安全。

王逸翔（1995-），男，汉族，山西运城人，北京交通大学，在读博士；主要研究方向和关注领域：网络空间安全。

刘博（1985-），男，汉族，河北新城人，北京交通大学，硕士，北京应用技术研究所，工程师；主要研究方向和关注领域：信息安全、保密技术。

常晓林（1971-），女，汉族，福建福清人，香港科技大学，博士，北京交通大学，教授；主要研究方向和关注领域：可信智能软件、网络安全、云边计算。