深度学习中的对抗样本问题

张思思 左 信 刘建伟

(中国石油大学(北京)自动化系 北京 102249)

摘 要 对抗样本是深度学习在安全领域中的热点问题,对抗样本的特性、生成、攻击方式以及如何防御对抗样本的攻击是当前研究对抗样本的重点问题.该文从对抗样本的概念、出现对抗样本的原因、对抗样本的攻击方式及原因阐述对抗样本的关键技术问题,对抗样本的概念主要是对对抗样本、对抗目标、对抗攻击所需知识的定义.该文列出了产生对抗样本的可能原因,目前,针对对抗样本出现的原因主要有三种观点:流形中的低概率区域解释,线性解释,此外,还有一种观点认为线性解释存在局限性,即当前的猜想都不能令人信服,进一步研究对抗样本出现的原因是未来重要的研究内容.并详细分析了对抗样本的几种典型生成方式:F-BFGS 法、FGS 法、迭代法、迭代最小可能类法及其它方法.并指出了其优缺点和适用的场景,比较了几种主要生成方式的不同之处.此外,对抗样本的攻击方式从应用场景上看主要分为两种,一种是白盒攻击,一种是黑盒攻击.对抗样本具有迁移性是对抗样本攻击的原因,该属性意味着攻击者可以不用直接接触基础模型,而选择攻击一个机器学习模型使样本被错误分类.针对对抗样本的攻击方式及原因,列出了目前深度学习中针对对抗样本的几种主要的防御技术:基于正则化方法、对抗性的预处理训练方法,蒸馏方法、拒绝分类方法等其它方法.指出了不同防御措施的适用场景与不足,阐释了上述防御措施均不能完全避免对抗样本的攻击.该文进一步探讨了对抗样本的应用,目前为止,对抗样本的应用主要是用在对抗评估及对抗训练上.最后,对对抗样本的未来研究方向进行了总体展望,彻底解决对抗攻击问题,仍有大量的理论和实践问题需要解决.找出对抗样本的特性,给出其具有实际应用前景的数学描述,探讨普适性的对抗样本生成方法,对抗样本的生成机理及对抗样本的攻击方式是研究对抗样本的重点问题,探索不同对抗性样本攻击的防御算法是主要目标,将这两部分结合解决对抗样本的攻击是今后对抗样本的主要研究方向.

关键词 对抗样本;特性;对抗样本生成;攻击方式;防御技术;深度学习;机器学习

中图法分类号 TP181 **DOI 号** 10.11897/SP.J.1016.2019.01886

The Problem of the Adversarial Examples in Deep Learning

ZHANG Si-Si ZUO Xin LIU Jian-Wei

(*Department of Automation, China University of Petroleum, Beijing* 102249)

Abstract Adversarial examples are a hot issue in the field of deep learning security. The characteristic, generation and attack mode are the key problems to be solved against the adversarial examples. This paper expounds the key technical problems of adversarial examples from the concept of adversarial examples, the causes of adversarial examples, the attacking ways and reasons of adversarial examples. The concept of adversarial examples is mainly about the definition of adversarial examples, adversarial examples' targets and the knowledge of counter attack. This paper lists possible reasons for the causes of adversarial examples, at present, there are three main viewpoints on the causes of adversarial examples: low probability region interpretation in manifolds, linear interpretation, in addition, there is another viewpoint that linear interpretation

收稿日期:2018-01-11;在线出版日期:2018-10-25.本课题得到国家重点研发计划项目(2016YFC0303703)、中国石油大学(北京)年度前瞻导向及培育项目(2462018QZDX02)资助.**张思思**,博士研究生,主要研究方向为机器学习.E-mail:sisisizhang@foxmail.com.**左 信**,教授,主要研究领域为过程控制与实时优化、可靠性分析.**刘建伟**,博士,副研究员,主要研究方向为机器学习、智能信息处理、复杂系统的分析、预测与控制、算法分析与设计.

has limitations, that is, the current conjecture can not be convincing. Further research on the causes of adversarial examples is an important research content in the future. Moreover, the main generation ways of adversarial examples analyzed are: F-BFGS method, FGS method, iterative method, iterative minimum possible class method and others. Meanwhile, their advantages and disadvantages and applicable scenarios are pointed out. Furthermore, the differences between several main ways of formation are compared. In addition, there are mainly two kinds of attacks of adversarial examples from the application scenario, one is white-box attack, and the other is black-box attack. The migration of adversarial examples is the reason for adversarial examples' attack. And this attribute means that an attacker can choose to attack a machine learning model without directly touching the underlying model to misclassify the examples. In addition, according to ways and causes of adversarial examples, the main defensive techniques against the examples are elaborated, including regularization method and preconditioning training based on antagonism, distillation, denial of classification and other methods, the application scenarios and shortcomings of different defense measures are pointed out, and it is explained that the above defense measures can not completely avoid the attack against the adversarial examples. Then, the application of adversarial examples has been discussed. So far, the application of adversarial examples is mainly used in confrontation evaluation and confrontation training. Finally, the future research direction of the adversarial examples is prospected. There are still a lot of theoretical and practical problems to solve on the problem of adversarial attack thoroughly. Finding out the characteristics of adversarial examples, considering the mathematical description of its practical application, discussing a universal method for generating adversarial examples, and investigating the generation mechanism and the attacking ways of adversarial examples are the key problems in the future study. It is the main goal to explore the defense algorithms against the attack of different adversarial examples. Combining these two parts to solve the attack of adversarial examples is the main research direction in the future.

Keywords　adversarial examples; characteristic; generation; attack method; defense against adversarial examples; deep learning; machine learning

1　引　言

2012 年，在 ImageNet 大规模视觉识别挑战中，深度学习开始崭露头角[1]. 近年来，深度学习发展迅速，其应用范围进一步扩大[1-2]，网络结构更加复杂[2-3]，此外，其训练方法有所改善，一些重要技巧的应用进一步提高了分类性能、减少了训练时间[1-5]. 例如，在图像识别领域，一些标准测试集上的实验结果表明，深度模型的识别能力已经可以达到人类的智力水平. 然而，在深度学习带给人们巨大便利的同时，其本身也存在一些安全性问题. 对于一个非正常的输入，深度模型是否依然能够得出满意的结果，其中隐含的安全问题也渐渐引起安全专家们的关注，因此，很多学者开始关注深度模型的抗干扰能力，也就是本文提出的关于深度学习对抗样本问题的研究.

早期在垃圾邮件检测系统和入侵检测系统等应用深度学习算法的安全领域中就发现了针对系统模型特点来逃避检测的问题，给深度学习在安全检测领域带来了很大的挑战. 迄今为止，越来越多威胁深度学习安全的问题被发现，有针对面部识别系统(Face Recognition System，FRS)缺陷来模仿受害者身份的非法认证危害，也有涉及医疗数据、人物图片数据的隐私窃取危害，更有针对自动驾驶汽车、语音控制系统的恶意控制危害[6].

因此，深度学习中对抗样本问题越来越值得关注，对抗样本出现的原因及生成方式是研究对抗样本的关键问题，利用对抗样本进行对抗训练及提高系统鲁棒性和深度学习的安全性迫在眉睫.

虽然人们曾猜测对抗样本出现的原因是深度神经网络的高度非线性特征以及监督学习中模型

平均不充分和正则化不充分，但 Goodfellow 等人指出高维空间的线性而非非线性才是对抗样本存在的真正原因[7]. 目前，对抗样本的生成方式主要有：L-BFGS(Limited-memory Broyden-Fletcher-Goldfarb-Shanno)方法、FGS(Fast Gradient Sign)方法、迭代方法、迭代最小可能类方法及其它方法.

找出对抗样本的攻击特性及攻击方式是我们解决问题的核心. 对抗样本的攻击方式从应用场景上看主要分为两种，一种是黑盒攻击，一种是白盒攻击[2]. 对抗样本具有黑盒攻击的能力是因为对抗样本的可迁移性[5]，Goodfellow 等人提出对抗样本在不同模型的泛化能力是对抗干扰与模型权值的高度一致造成的. 所以，当训练同一任务时，对抗样本可以在不同的模型上学习相似的函数，即对同一个输入-输出对，拟合多个相似的函数[7]. 探索对抗性样本攻击的不同的防御算法是研究对抗样本的主要目标. 目前，针对对抗样本的防御技术主要可以分为五类：基于正则化方法、对抗性的预处理训练方法、蒸馏法、拒绝分类方法和其它方法. 虽然通过这些方法在一定程度上可以抵御对抗样本的攻击，但这些方法并不能应用在所有模型上，所以研究更有力的防御对抗样本攻击的算法是未来的主要研究方向.

2 对抗样本的关键技术问题

2.1 对抗样本的概念

定义 1. 对抗样本. 通过故意对数据集中输入样例添加难以察觉的摄动使模型以高置信度给出一个错误的输出. 即只需要在一张图片上做微小的扰动，分类器以很高的置信度将图片错误分类，甚至被分类成一个指定的标签(不是图片正确所属的标签)[5].

样本通常包含样例-类标签对，由于对抗样本在生成时已经预知其类标签，故对抗样本仿佛是隐含有类标签的. 所以，本文中称对抗样例为对抗样本以强调对抗样例的特性.

简单来说，有一个学习系统 M 及干净的输入样本(没有添加噪声的样本)C，我们假设样本 C 被学习系统正确地分类，即 $M(C)=y_{\text{true}}$，建立一个几乎与样本 C 相同但是却被错误分类的样本 A，使 $M(A)\neq y_{\text{true}}$，这样的样本 A 我们称之为对抗样本.

虽然对抗干扰的存在机率远比噪声干扰要小，但被分类器误分的概率却远比噪声干扰高. 此外，在训练集的不同子集上训练得到具有不同结构的模型都会对相同的对抗样本实现误分，这意味着对抗样本成为了训练算法的一个盲点. Biggio 等人阐述了对抗样本的相关概念[8]，提出了对抗模型(Adversary Model)、对抗目标(Adversary's goal)、对抗攻击所需知识(Adversary's knowledge)和对抗能力(Adversary's capability)等概念.

为了激发对抗样本最优的攻击策略，我们有必要知道对抗攻击所需知识及对手操控数据的能力，最后利用对抗模型修改潜在的数据分布，利用对抗攻击能力产生最优的攻击策略. 为此，我们利用对手的一般模型来阐明对抗样本的某些概念.

定义 2. 对抗目标. 对抗的目标是根据需要最大或最小化损失函数. 在攻击设置中，攻击目标就是操控某一样本使它被系统错误地分类. 严格来说，对抗样本 C 应该满足 $g(C)<-\varepsilon$(其中 g 为判别函数，$-\varepsilon$ 是阈值，满足 $\varepsilon>0$)，即对抗样本仅需越过决策边界即可被错误分类. 但是，一个好的攻击策略是通过创建一个样本使系统以高错误率误判为错误分类，即在满足某些约束的情况下，最小化判别函数 $g(C)$.

定义 3. 对抗攻击所需知识. 对抗样本与正常样本关于目标学习系统的认知可能会有很大不同. 对抗样本的系统知识可能包含：

(1) 全部或部分训练集；

(2) 每个样例的特征表示，真实的物体如电子邮件、网络数据包是如何被映射到分类器的特征空间中的；

(3) 学习算法的类型及其决策函数的形式；

(4) 被训练的分类器所使用的模型(例如线性分类器的权重)；

(5) 或分类器的输出(对抗样本能选择样例的分类标签等).

定义 4. 对抗能力. 在攻击中，对抗样本的能力仅限于测试数据的修改，即不允许改变训练数据. 但是，在这个限制下，攻击者的能力变化包括：

(1) 修改输入样本；

(2) 修改特征向量；

(3) 对某些特定特征进行独立修改(输入数据的语义可能决定某些特征是相互依赖的)等.

2.2 出现对抗样本的原因

神经网络很容易受到“欺骗”，当对抗样本出现时，神经网络会将一个无法识别的图像看成为确定已知类型的图像，且对抗样本能通过不同的生成策

略生成对抗样本来愚弄神经网络[9-10]. 这暴露了机器学习算法的盲区，表明了深度神经网络通过反向传播学习的过程存在隐含特征及盲区，神经网络的输出结果以一种不明显的方式与数据分布相关联. 对抗样本的起因是个谜，人们曾猜测对抗样本出现的原因是深度神经网络的高度非线性特性，以及监督学习中模型平均不充分和正则化不充分. 后来 Goodfellow 等人提出高维空间的线性特性而非非线性特性才是对抗样本存在的真正原因[7]. Szegedy 等人提出是高层神经网络空间而非单个神经元包含神经网络更高层的抽象语义信息. 同时，Szegedy 等人发现深度学习网络的输入输出映射是不连续的，神经网络将对抗样本错误分类的原因可能是神经网络将对抗样本正确分类的概率极低，故对抗样本在测试集中很难观察到[5]. 因此，在未来的研究中还需进一步探索如何提高神经网络对对抗样本正确分类的概率.

目前，针对对抗样本出现的原因主要有三种观点：流形中的低概率区域解释，线性解释，此外，还有一种观点认为线性解释存在局限性，如图 1 所示.

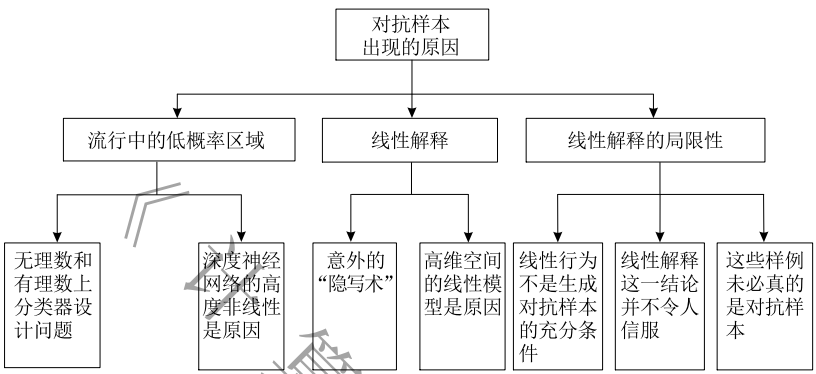


图 1　出现对抗样本的可能原因及其局限性

2.2.1　流形中的低概率区域解释

为什么神经网络模型不能正确分类对抗样本呢？一种解释是由于这些对抗样本是从总体样本所在概率空间的某一子空间中抽样得到的，因此，对于这些在训练阶段只学习了一些局部子区域，且训练样本个数有限的分类算法而言，对抗样本已经超出了分类器所能学习的概率分布所在支集[5]. 据此可知，深度神经网络处理对抗样本能力有限的问题就无法避免，尤其对于这种利用目标函数对模型向量求梯度，使目标函数最优的模型向量，当输入数据异常时，我们很难从输出的结果中找到这一非正常的输入. 因此，这些被“动过手脚”的输入数据虽然不能“骗过”人眼，但却能轻易“骗过”算法.

文献[5]中对于对抗样本没有给出详细的解释，只给出一个简单的实数分类问题表示对抗样本的存在，即对抗负数集合出现的概率很低，因此在训练中从来没有(或很少)观察到，但它是稠密的(类似于有理数)，所以在每个测试集中都能找到它. 我们可以使用概率密度的数学概念来定义一个分类器. 通过在实数域上使用如下的判别准则设计分类器 C：

(1) 如果 x 是正无理数或负有理数则属于＋类；

(2) 如果 x 是负无理数或正有理数则属于－类.

实验中随机选取实数，分类器 C 完美地区分了正数和负数. 在实数集合中随机选择任意一个测试数 x，由于 x 为无理数的概率更大，则分类器 C 正确地将它分类. 但是，分类器 C 存在对抗样本的问题：由于有理数集在整个实数集合中是稠密的，因此 x 接近有理数而构成了对抗样本.

如上所述，无理数和有理数的分类器设计问题很有趣，但也影射出了一个重要问题，即深度神经网络定义的判别准则是什么，通过何种机制才能创造出低概率区域. 文献[5]中虽未详细提及，但阐明了深度网络的高度非线性是对抗样本生成的可能原因.

2.2.2　线性解释

Goodfellow 等人最早提出了线性解释的概念[7]. 在许多问题中，单个输入特征的表示精度是有限的. 例如，数字图像通常只使用 8 位二进制表示像素强度值，则有限精度集合表示为$\{1,2,3,\cdots,255\}$. 由于特征表示的精度有限，那么分类器对输入 x 及加了干扰的输入 $\tilde{x}=x+\eta$(干扰元素 η 比输入 x 的精度小)的响应不同是不合理的，通常对于好的分类器设计，只要 $\|\eta\|_\infty<\varepsilon$($\varepsilon$ 足够小)，分类器会将 x 及 $\tilde{x}$ 归为相同的类别.

对于给定的输入 x 及对抗样本 $\tilde{x}=x+\eta$(其中 η 满足约束 $\|\eta\|_\infty<\varepsilon$)，考虑权重向量 w 和对抗样本

$\tilde{x}$ 的内积：

$$\boldsymbol{w}^{\mathrm{T}}\cdot\tilde{x}=\boldsymbol{w}^{\mathrm{T}}\cdot x+\boldsymbol{w}^{\mathrm{T}}\cdot\eta.$$

假定激活函数为线性函数，对抗干扰使激活函数 $f(\boldsymbol{w}^{\mathrm{T}}\cdot x)$ 增加了 $f(\boldsymbol{w}^{\mathrm{T}}\cdot\eta)$，只要 $f(\boldsymbol{w}^{\mathrm{T}}\cdot\eta)$ 中 η 满足最大范数约束 $\eta=\varepsilon\mathrm{sign}(\boldsymbol{w})$，我们就可以最大限度地增加对抗干扰. 若 $\boldsymbol{w}$ 的维数是 n，权重向量每一个分量的平均值是 m，则激活函数值就会增长 εmn. 由于无穷大范数 $\|\eta\|_{\infty}$ 不会随着维数的增加而增加，但干扰 η 能引起激活函数值线性增加，因此，对于高维问题，输入的微小变化能够引起输出做出很大的改变.

文献[7]用下面的例子来解释这一现象：将上述问题看作是一种“意外的隐写术”，在这种情况下，即使存在多个信号或其它大振幅信号，线性模型也只与其权值最接近的信号相关联. 当输入为图像时，沿着梯度方向很小的线性运动就可以导致深度神经网络改变预测结果. Goodfellow 等人得出结论：“若一个线性模型的输入有足够的维数，那么这个线性模型就会存在对抗样本问题.”因此，我们可知，高维空间的线性模型是生成对抗样本的原因.

2.2.3 解释的局限性

(1) 线性解释并不令人信服

“意外的隐写术”似乎很好地说明了对抗样本出现的原因，但这一说法并不能令人信服. 因为当扰动与激活函数本身相关时，就不会引起激活函数线性增长. 重新考虑权重向量 $\boldsymbol{w}$ 和对抗样本 $\tilde{x}$ 的内积，正如我们之前所看到的，激活函数的改变量 $\boldsymbol{w}^{\mathrm{T}}\cdot\eta$ 随着 η 的改变呈线性增长，但激活函数的 $\boldsymbol{w}^{\mathrm{T}}\cdot x$ 相应部分也随之增长，实际上这两个量之间的比例仍保持不变.

为了解释上述问题，采用线性分类器对改良版的 MNIST 数据集“3”和“7”进行分类，将该数据集的图像像素大小增加到 200×200. 通过线性内插法、对原始数据添加噪声和修改数据集来增加样例的维数，两幅图像的大小仍非常相似，即使图像的维数增加了 50 以上，图像分辨率也不会影响对抗扰动的感知度.

因此，高维问题更容易生成对抗样本这一结论并不成立. 据此我们可知，线性行为是对抗样本存在的充分原因这一观点也不能令人信服.

(2) 线性行为不是生成对抗样本的充分条件

根据 Goodfellow 等人对对抗样本的线性解释可知线性行为是对抗样本存在的原因. 所有的线性分类问题都应遭受对抗样本的干扰. 然而，在线性分类问题上却可以找到一类根本不存在对抗样本的图像，这说明线性行为不是生成对抗样本的充分条件.

(3) 上述样例未必真的是对抗样本

赞成对抗样本线性解释的一个关键论点是 logistic 回归函数上也存在对抗样本现象. 然而，Tanay 等人指出 MNIST 数据集上的线性分类器与 ImageNet 数据集上深度神经网络是不同的[11]，主要区别如下：

首先，对抗干扰有更高的幅值，在 MNIST 数据集上通过线性分类器很容易被人们观察到干扰的存在. 值得注意的是，图像的分辨率不是引起这种差别的原因：增加 MINST 数据集上图像的大小不能影响对抗干扰的感知程度，这种线性的解释不仅不能可靠地预测对抗样本是否会在某个特定的数据集上发生，也无法预测对抗扰动的幅值是否会使分类器改变预测值.

其次，在对抗干扰的表现形式上也有所不同. 在数据集 ImageNet 上，GoogLeNet 的对抗干扰结构以高频为主，不能进行有意义的解释，而在 MNIST 数据集上，logistic 回归函数中的扰动则主要以低频为主. 虽然 Goodfellow 等人认为人们不容易察觉数据集 MNIST 上数字 3 和 7 之间的区别，但在 GoogLeNet 上，可以这样解释：通过 logistic 回归点在某一方向上找到的权向量在接近通过两个类的平均图像的方向上，因此定义一个判别边界类似于估计一个最近的质心分类器.

在数据集 MNIST 上由 SVM 和 logistic 回归定义的简单的线性模型可以受到来自视觉上可感知扰动的“欺骗”，这看起来类似于最近的质心分类器的权重向量. 这个结果不足为奇，并不能解释为什么更复杂的模型(如深度神经网络)能够被人眼所不能察觉的细小改变所“欺骗”，所以很明显线性解释仍然不充分[11].

2.3 对抗样本的攻击方式及原因

2.3.1 对抗样本的攻击方式

专注于互联网安全的极棒实验室总监王海兵表示：“用对抗样本攻击人工智能，其实就是从最核心的算法层面来攻击它.”对抗样本的攻击方式从应用场景上看主要分为两种，一种是白盒攻击，一种是黑盒攻击[2]. 所谓白盒攻击指的是攻击者对目标分类器充分了解，即攻击者知道分类器类标签所在的空间，分类器的类型以及训练模型，能够获知机器学习

所使用的算法，以及算法所使用的参数. 攻击者在产生对抗性攻击数据的过程中能够与机器学习的系统有所交互；黑盒攻击指攻击者知道目标的特征表示和分类器的类型，但不知道所要学习的分类器形式或是训练数据，但攻击者仍能与机器学习的系统有所交互，比如可以通过输入任意输入观察并判断分类器的输出.

2.3.2　对抗样本的可迁移性

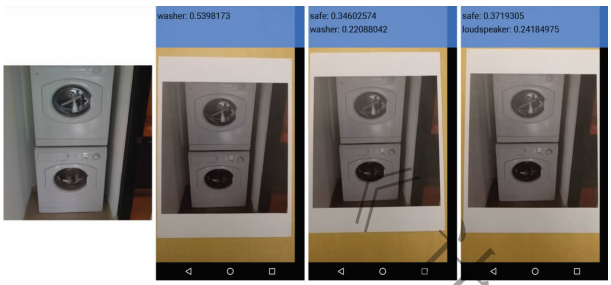
对抗样本具有可迁移性是 Szegedy 等人在文献[5]中首次提出的，对抗样本的可迁移性是指对抗样本被 M_1 模型错误分类，也同样可以被 M_2 模型错误分类. 对抗样本的迁移属性意味着攻击者可以不用直接接触基础模型，而选择攻击一个机器学习模型使样本被错误分类. Szegedy 等人在相同数据集上研究了不同模型的迁移性，此外，还在数据不相交子集上训练相同或不同模型并研究其间的迁移性问题；但不足的是，Szegedy 等人的实验成果都是在 MNIST 数据集上实现的. Goodfellow 等人提出对抗样本在不同模型间的泛化能力是由于对抗干扰与模型的向量高度一致所致，所以当训练相同的任务时，对手可以在不同的模型上学习相似的函数[7]. 这种泛化特性意味着，若敌人要对模型进行攻击，无需访问目标模型，仅将自身模型训练产生的对抗样本送至目标模型中即可实现.

Papernot 等人主要研究如何构造一个替代模型进行黑盒攻击[12]. 为了训练替代模型，他们开发了一种人工生成训练集，通过查询目标模型的标签来体现其可迁移性. 此外，他们还研究了深度神经网络与其它模型（如决策树）和 k 近邻之间的可迁移性.

Moosavi-Dezfooli 等人指出干扰是普遍存在的[13]，且可以在不同的图像之间迁移. 他们还表明，这些干扰生成的对抗图像可以在 ImageNet 上的不同模型之间迁移.

文献[12]中仅研究了模型和训练过程为黑盒，训练集和测试集均由攻击者控制的迁移问题；与文献[12]所做的工作不同，Liu 等人攻击 Clarifai 网站时[14]，其模型的训练数据和训练过程，乃至测试标签集都是攻击者未知的，这在更大的模型和数据集上验证了对抗样本的可迁移性. 其假定在攻击黑盒机器学习系统时，也不需要依赖查询系统来构造替代模型. 此外，Liu 等人还首次研究了无特定目标（non-targeted）情况下对抗样本的可迁移性问题. Fischer 等人展示了现有的对抗攻击是如何在任务间进行迁移的，并能够在几乎不改变神经网络原有预测结果的前提下，通过造成细微的对抗干扰使深度神经网络将所选类的所有像素错误分类[15]. 但这一研究仍存在许多不完善的地方，比如是否需要使用更高性能的网络；是否需要比较不同的网络结构对对抗样本问题的影响，并使用更复杂有效的方法来测试对抗样本的有效性；对抗样本是否能应用于实际物理世界等. Kurakin 等人指出人们通常不能发现微小的改动，但分类器能够捕捉到并对样例进行错误分类[7]. 对抗样本让人们对机器学习的安全问题产生担忧，因为对手即使不知道模型，也能够攻击机器学习系统. 当前的研究工作均为在假设对手将数据直接送入机器学习的分类器模型中这一条件下开展的，但在物理世界中情况并不总是这样，例如来自相机或其它传感器的信号输入. 现在，越来越多的研究转向对抗样本对物理世界的攻击问题，如 Carlini 等人通过创建一段模糊的音频输入，让手机识别出清晰的语音指令攻击[16]；基于图像的脸部识别，容易受到重播的攻击，攻击者能用用户曾经拍摄的图像而非真实的人脸呈现给相机来攻击机器学习系统[17]. 对抗样本原则上能够应用于这些物理领域，如在语音命令中包含一个记录对人类观察者来说是无用的（例如一首歌），但是这个语音指令却会被机器学习算法识别. 在人脸识别领域，可以对人脸作出细微的纹理改变，人类观察者能正确地进行识别，但是机器学习系统却将其识别成不同的人. Kurakin 等人展示了即使在物理世界中，机器学习系统也容易受到对抗样本的攻击. 通过将来自手机和照相机的对抗照片，送入到 ImageNet 感受器分类器中发现，即使通过相机接收的图片，大部分对抗样本也被错误分类了. 如图 2 所示，是 Kurakin 等人用物理对抗样本在手机应用程序上对图像分类的例子. 从数据集中得到一幅未增加对抗干扰的图片（a）并使用不同的干扰 ε 来生成不同的干扰图像，之后打印出未受干扰的图像，使用 TensorFlow 相机演示应用程序如何对这些图片分类. 图片（b）被正确地分类为洗衣机，而（c）和（d）则被错误地分类. Sharif 等人同 Kurakin 等人所做的工作相似，都在纸上打印出对抗样本并且成功地“愚弄”了识别系统[18]. Lu 等人在文献[19]中提出了不同的观点，尽管物理世界中对抗样本是存在的，但是这些实验忽略了物理世界一个至关重要的特性：相机能从不同

的距离及不同的角度来观察物体，且实验表明当前对抗样本的结构不能“愚弄”一个移动平台上的目标检测算法. 相反，一个经过训练的神经网络能将来自不同距离及角度的干扰图片正确地分类，Lu 等人认为这是由于干扰的对抗性对所观察干扰图片的比例比较敏感造成的，因此一辆无人驾驶汽车仅会在一小段距离内将汽车的停止标志识别错误. 由此可知，当对抗干扰的方法运用在停止标志检测中时，仅对于精心选择的情形起作用，故在很多实际情况中，我们不必过于担心对抗样本带给我们的影响.



(a) 数据集上的图片 (b) 未添加干扰的图片 (c) 干扰=4 (d) 干扰=8

图 2 取不同的干扰时生成的对抗样本

对抗样本是普遍存在的，能在图像和不同分类器之间迁移[20]. 对黑盒学习系统的攻击依赖于对抗样本的可迁移性，对抗样本通过一个替代的分类器来攻击机器学习算法[20-21]. 此外，一些学者提出了基于集成的方法来生成可迁移性对抗样本实例[14].

3 对抗样本的典型生成方式

对抗样本是评估和提高机器学习鲁棒性的关键步骤. 因此，研究对抗样本的生成方式是研究对抗样本的必要步骤. 对抗样本的生成方式有很多种，目前主要的生成方式有 L-BFGS 方法、FGS 方法、迭代法、迭代最小可能类方法等.

3.1 L-BFGS 方法

对抗样本的生成方式有很多种，包括在图像像素上计算梯度或是直接在图像像素上求解某个目标函数的优化问题，Szegedy 等人在文献[5]中提出 L-BFGS 方法，通过简单的最优化过程，对一个能够正确分类的输入图像作微小的扰动，使它不再被正确分类. 从某种意义上说，这种方法是通过优化遍历流形网络表示并在输入空间中发现对抗样本. 对抗样本存在流形空间中的低概率区域，因此很难通过对输入点附近简单的随机采样获得.

目前，各种最新的计算机视觉模型在训练过程中都采用输入变换来提高模型的鲁棒性及收敛速度. 然而，对于一个给定的样例来说，这些变换不影响统计结果，样例与样例变换之间是高度相关的，且是从整个模型训练数据的相同概率分布中变换出来的. 因此，Szegedy 等人提出 L-BFGS 方法是围绕训练数据，利用该模型和它的不足建立局部空间的.

令 $f:R^m\to\{1,\cdots,k\}$ 是将图片像素值向量映射成离散标签集的分类器，且假设 f 上的一个连续损失函数为 $loss_f:R^m\to\{1,\cdots,k\}\to R^+$. 对于给定的图像 $x\in R^m$ 及对应的类标签 $l\in\{1,\cdots,k\}$，$D(x,l)$ 是图像 x 到对应类标签 l 的映射函数，通过求解如下最优化问题来找出对抗样本：

$$\begin{aligned}&\text{minimize}\ \|r\|_2\\&\text{满足}\ f(x+r)=l,\ x+r\in[0,1]^m\end{aligned}\tag{1}$$

其中 r 值可能不是唯一的，f 函数把 $x+r$ 分成最接近的 x 的 l 类图像，很明显 $D(x,f(x))=f(x)$，所以这个任务只有在 $f(x)\neq l$ 时才有意义. 然而，计算 $D(x,l)$ 是一个很难的问题，所以 Szegedy 等人提出 L-BFGS 方法. 具体地，通过线性搜索找出满足 $c>0$ 且最小的 r 值满足 $f(x+r)=l$，从而找到 $D(x,l)$ 的估计值.

$$\begin{aligned}&\min\ c\|r\|_2+loss_f(x+r,l)\\&\text{满足}\ x+r\in[0,1]^m\end{aligned}\tag{2}$$

这个罚函数法在凸损失函数条件下将产生 $D(x,l)$ 的精确解，据此可找到最优的近似解. 然而，一般情况下，神经网络是非凸的，因此我们只能得到 $D(x,l)$ 的近似值.

这个方法虽然可靠且稳定有效，可以使用各种各样的模型对输入进行错误分类，但算法复杂性很高.

3.2 FGS 方法

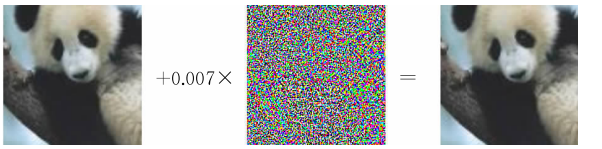
Goodfellow 等人在文献[7]中提出了快速梯度符号方法，这是生成对抗样本最简单的方法之一. 其核心是让输入图像 $x\in R^m$ 朝着类别置信度降低的方向移动，令 $x\in R^m$ 是输入图像，y 是输入 x 对应的类标签，θ 是模型参数，η 是步长，ε 是所选的超参数，$J(\theta,x,y)$ 是训练神经网络的损失函数，$\nabla_x J(\theta,x,y)$ 是损失函数的偏导数. 我们可以在 θ 当前值附近对损失函数线性化得到干扰的最大范数限制：

$$\eta=\varepsilon\,\text{sign}(\nabla_x J(\theta,x,y))\tag{3}$$

再通过求解下式求得对抗样本：

$$\tilde{x}=x+\eta\tag{4}$$

如图 3 所示，是用 FGS 方法生成的对抗样本，使左图的这只熊猫被认定为是一只长臂猿[3]. 且在各个维度上移动相同大小的一步，虽然一步很小，但每个维数的效果加在一起，通常也足以对分类器的判别结果产生很大的影响.



x
以57.7%的置信度被认为是“熊猫”

$\mathrm{sign}(\nabla_x J(\theta, x, y))$
以8.2%的置信度被认为是“线虫”

$x+\varepsilon\mathrm{sign}(\nabla_x J(\theta, x, y))$
以99.3%的置信度被认为是“长臂猴”

图 3　FGS 方法生成的对抗样本

叠加在典型图片输入上的对抗扰动会让分类器产生错觉，误将熊猫识别为长臂猿. 在计算上，这种方法有巨大优势，因为只需要一次前向和一次后向梯度计算就可以产生对抗样本了. FGS 法因为非常简单，用任何框架都很容易实现，所以通常用这种方式产生对抗样本.

3.3　基本迭代法

事实上，在多数情况下，单纯采用 FGS 方法产生的对抗样本是无效的，也许是因为两个类别类内差异过大，最极端的情况是某个类别可能处在整流激活函数(Rectified Linear Unit，ReLU)都小于 0 的“死区”内. 若考虑上述两种情况，则需要研究比 FGS 方法更好、更实用的方法. 若 FGS 方法采用朝着类别置信度降低的方向直接前进一大步可能是错的，那么借鉴梯度下降的思路，一步步迭代前进则更为合适. 虽然这种迭代过程在梯度方向上不能实现线性迭代，且需多次计算，但相较于 L-BFGS 方法而言还是要更简单，效果更好.

基于此，Kurakin 等人在研究物理世界中是否存在对抗样本问题时，对 FGS 方法进行改进并提出了一种更直接的方法：基本迭代方法(Basic Iterative Method，BIM)，这种方法运用梯度下降的思想，一步步迭代前进. 通过一个小的步长来多次使用快速梯度法，且剪裁每一步中间结果的像素值并确保它们在原始图像的某个邻域中[9]. 将这种方法进一步细分为两种：(1) 减小分类器预测样例曾属类别的置信度；(2) 增大样例曾被预测为最小可能类别的置信度：

$$X_0^{\mathrm{adv}} = X,$$

$$X_{N+1}^{\mathrm{adv}} = Clip_{X,\varepsilon}\{X_{N+1}^{\mathrm{adv}} + \alpha\,\mathrm{sign}(\nabla_X J(X_N^{\mathrm{adv}}, y_{\mathrm{true}}))\} \quad (5)$$

上式中，X 是 3-D 输入图像，假设每个像素强度值都是在[0，255]之间的整数值，y_{true} 是图像 X 的真实类标签，$J(X, y)$ 是神经网络在给定的图像 X 和标签 y 上的交叉熵损失函数，α 是步长，通常取 $\alpha=1$，选择迭代次数为 $\min(\varepsilon+4, 1.25\varepsilon)$. 对于神经网络输出层，将交叉熵损失函数应用于整数类标签等价于求解给定图像的真实类标签上的负对数条件概率：$J(X, y) = -\log p(y \mid X)$；$Clip_{X,\varepsilon}\{X'\}$ 是图像 X' 像素的剪裁函数，确保剪裁后的图像仍然在原始图像的 ε 邻域内.

然而上述方法仅仅试图增加正确分类的损失值，没有明确说明模型应该选择哪种错误的类标签，这些方法仅适于应用在类的种类很少且各类彼此互不相同的数据集(例如 MNIST 和 CIFAR-10)上.

3.4　迭代最小可能类方法

数据集 ImageNet 包含 3000 种不同的类，且各个类之间有不同程度的差异，上述的这些方法只适用于分类个数比较少的情况，如将一个品种的狗拉雪橇误分成另一个品种的狗拉雪橇. 为了产生更多的错误类，Kurakin 等人介绍了迭代最小可能类方法[7]，这种迭代方法试图让对抗样本被分成某个特定目标类，根据在图像 X 上训练深度神经网络得到预测的结果，选择最小可能类：

$$y_{LL} = \arg\min_y \{p(y \mid X)\} \quad (6)$$

对于一个训练好的分类器，最小可能类方法产生的对抗样本通常与真实类完全不同，所以这种攻击方法会导致有趣的错误分类，比如将狗误以为是飞机等.

迭代最小可能类方法是迭代法的改进，通过朝着方向 $\mathrm{sign}\{\nabla_X \log p(y_{LL} \mid X)\}$ 迭代使 $\log p(y_{LL} \mid X)$ 最大化，从而使对抗图像被分类成 y_{LL}. 对神经网络的交叉熵损失函数来说，最后一步的表达式等价于 $\mathrm{sign}\{-\nabla_X J(X, y_{LL})\}$. 因此，对抗样本的迭代产生过程有如下的表达式：

$$X_0^{\mathrm{adv}} = X,$$

$$X_{N+1}^{\mathrm{adv}} = Clip_{X,\varepsilon}\{X_{N+1}^{\mathrm{adv}} - \alpha\,\mathrm{sign}(\nabla_X J(X_N^{\mathrm{adv}}, y_{LL}))\} \quad (7)$$

对于上述迭代过程，使用了相同的 α 及同基本迭代法相同的迭代次数.

3.5　其它方法

同文献[22]提出的方法类似，Rozsa 等人在文献[23]中讨论对手样例多样性和出现难分类的正样例问题时，提出了图像逆转的方法，通过减小网络倒数第二层的一个特征向量独热来表示在给定类上的

损失进行重组图像.将倒数第二层的一个独热向量(一个向量只有一个非零元素)反转,最终会产生属于热类(独热表示所选中的类)的较低层特征及由零向量表示的倒数第二层(非热类)的其它特征.

Rozsa 等人将"冷"类加到这一概念中来进一步减少当前类的作用.具体来说,需要为倒数第二层(软最大函数前一层)设计特征值.在这一层中每一个值仍然与某个特定的输出类相关,所以可以在输入图像空间中定义使输出远离原始类(冷类)而朝着目标类前进(热类).

为了建立热/冷模型,令 $h(x)\in R^n$ 为输入图像 x 的神经网络倒数第二层的特征值,y 是 x 所对应的类标签,基于 $h(x)$ 建立一个热/冷特征向量 $\boldsymbol{w}_{hc}$:首先定义一个目标热类 $\tilde{y}\neq y$,通过定义某一类上的 $|h_{\tilde{y}}(x)|$ 函数值来增加 x 属于 $\tilde{y}$ 类的可能性.其次,定义 y 为冷类且其值为 $-h_y(x)$,通过从输入图像中移除冷类的独占特征,从而远离冷类.因此,基于 $h(x)$ 重构的倒数第二层的特征向量为

$$\boldsymbol{w}_{hc}(x)=\begin{cases}|h_j(x)|, & j=\tilde{y}\\ -h_j(x), & j=y\\ 0, & \text{否则}\end{cases} \tag{8}$$

使用标量 $|h_j(x)|$ 表示热类,进而通过计算干扰 $\eta_{hc}=B_l(\boldsymbol{w}_{hc})$ 来重构图像 $\tilde{x}$,其中 $\tilde{x}=x+\eta_{hc}$,算子 $B_l(\cdot)$ 是图像反向传播到输入层的偏导数的估计值.

Rozsa 等人提出了一个简单且高效地通过深度卷积神经网络(Deep Convolutional Neural Networks,DCNNs)自动地对每一个面部特征属性进行提取的方法[23].为了测试神经网络的稳定性,通过快速翻转属性(Fast Flipping Attribute,FFA)技术生成对抗图片,通过 FFA 技术可知其相较于其它相关算法能产生更多的对抗样本.此外,Rozsa 等人首次提出了自然对抗样本的概念,自然对抗样本是指那些被误分的图片通过小的扰动很容易被拉回到正确分类的样本中去,这一现象的发现是令人兴奋的,因为在训练数据的指导下纠正一个错误的分类结果,仅仅需要改变 DCNNs 的一个参数.所以,为了进一步提高网络的整体性能,未来的工作还应进一步考虑用对抗样本、难分类的正确样例及纠正的自然对抗样本,甚至是用含有较低扰动的样本来提高对对抗样本生成模式的鲁棒性.

对对抗样本的研究目前主要集中在图像识别领域,而 Hu 等人指出在对抗攻击下序列机器模型(the sequential machine models)也同样不安全[24],且提出了一种新的生成对抗样本序列算法,用来攻击基于恶意软件检测的递归神经网络(Recurrent Neural Networks,RNNs)系统.Hu 等人还提出了从原始恶意软件输入序列生成对抗样本输出序列的生成式 RNNs 模型.实验结果表明基于恶意软件检测的 RNNs 检测算法不能检测大多数生成的恶意对抗样本,这意味着提出的模型能够绕过检测算法.该模型已经成功地使产生的大部分对抗样本能够绕过几个结构不同的黑盒受害者 RNNs 的检测.但在这篇文献中作者仅仅对 RNNs 系统作了研究,在未来的研究中,可以将提出的模型推广到基于卷积神经网络(Convolutional Neural Network,CNN)的恶意软件检测中.

Rozsa 等人在文献[23]中提出了新的、更自然的生成对抗样本图像的方式(以前的生成方式类似于添加随机噪声),这种方式有很强的结构效应,其通过扩大对抗样本图像产生额外的难分类的正样例(不存在目标集中只存在于测试集中的样例)来扩大训练数据集并进一步提高单独训练对抗样本的准确性.Baluja 等人在文献[25]中通过自我监督的方式训练前馈神经网络并提出了对抗迁移网络(Adversarial Transformation Networks,ATNs)的概念以及通过 ATNs 生成对抗样本的两种方式.不同于探索对抗样本在分类任务中的应用,Kos 等人在文献[26]中探索了对于深度生成式模型(例如变分自编码器以及变分自编码器生成式网络)的对抗样本的生成方法.Xie 等人在文献[27]中通过在像素集上最大化损失函数来生成对抗干扰,并提出稠密对抗生成(Dense Adversary Generation)算法,其能够生成大规模的对抗样本.理想情况下,必须通过改变任务的损失函数来生成对抗样本,例如,对抗样本要攻击一个语音识别系统,那么必须要最大限度地提高目标系统的识字错误率.现有的生成对抗样本的方法是在正确样本的某个邻域内,向可微的损失函数梯度方向搜索找到对抗样本.然而,一些对偏好进行分类的结构不可分解的预测问题,是不能基于梯度的方法来生成对抗样本的.例如,度量评估人体姿态估计的正确率,自动语音识别系统使用音素进行错误率评估.所有这些度量标准都是不可微的评价标准.Cisse 等人考虑任务的目标函数提出了新的生成对抗样本的方法,并将它命名为胡迪尼法(Houdini)[28],不管损失函数是组合优化问题还是不可微的函数,这种方法可以"欺骗"任何基于梯度的机器学习方法,其目前已成功地运用到语音识别,姿态估计和语

义分割中.

4　几种主要生成对抗样本方法的比较

如图4所示，图中比较了基于已训练好的Inception V3下深度神经网络模型的不同的生成对抗样本的方式及不同的 ε 值生成的对抗样本的准确率(分别用 Top-1、Top-5 两种图片检测准确率的标准).

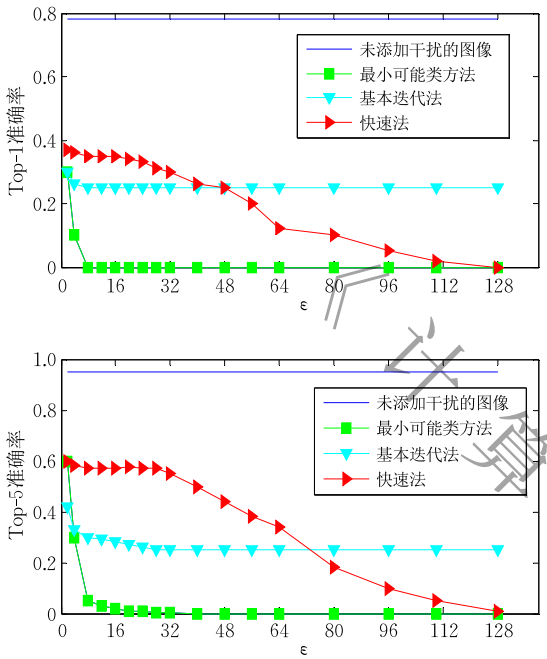


图 4　几种主要生成对抗样本的方式的比较

上述实验是在 ImageNet 数据集(Russakovsky et al., 2014)上的 50000 个样本上进行的，从图 4 中我们可以看出，每一种对抗样本的生成方式都不能保证每一个对抗样本图像会被误分，有时攻击者获胜，有时机器学习模型获胜.

对于每个验证图像，我们使用不同的方法和不同的干扰值生成对抗样本. 对于每种方法及 ε 值，我们在 50000 个图像上计算它的准确率，同样也在所有未加扰动的干净的图片上计算准确率作为基准值. 从图中可以看出，用 Top-1 标准检测时，当 ε 值增加，FGS 法的准确率以两倍的速度下降，而当 Top-5 检测时，即使最小的 ε 值准确率也只有 40%. 当增加 ε 值时，由 FGS 法生成的对抗样本准确率保持不变直到 $\varepsilon=32$，然后当 $\varepsilon=128$ 时准确率降到 0. 这是因为 FGS 法为每一个图像增加了干扰 ε，当 ε 值大于一定值时，图像基本被破坏了，即使人类也很难辨识出，所以准确率几乎为 0. 当 $\varepsilon<48$ 时 BIM 法能够产生更好的对抗样本，然而，当我们增加 ε 时，BIM 法不能提高错误率. 当 ε 很小时，最小可能类方法使得大多数图像被错误分类.

如图 5 所示是不同对抗方法产生对抗样本的例子. 图中选取 $\varepsilon=32$，从图 5 中可以看出与 FGS 法相比，基本迭代法产生的扰动更精细. 此外，基本迭代法不是只选择 ε 邻域边界上的一个点作为对抗图像，因此，即使存在较大扰动，图像也不会被破坏且分类器仍具有较高的混淆率.

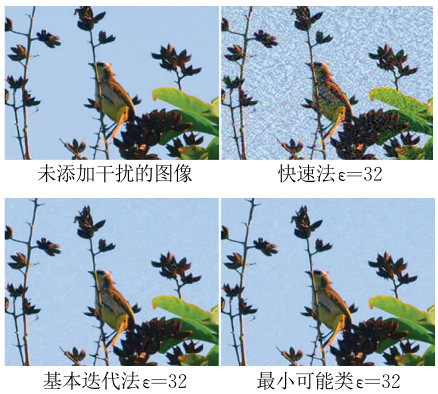


图 5　取相同的干扰值时不同的方法生成的对抗样本

为讨论物理世界中对抗样本的可能性及在不同方法下的区别，在包含对抗样本的打印照片上进行实验并计算正常样本及对抗样本在照片转换前后的精度及对抗样本经过照片转换后的破坏率. 定义照片的破坏率如下：

$$d=\frac{\sum_{k=1}^{n}C(X^k,y_{\text{true}}^k)\overline{C(X_{\text{adv}}^k,y_{\text{true}}^k)}C(T(X_{\text{adv}}^k),y_{\text{true}}^k)}{\sum_{k=1}^{n}C(X^k,y_{\text{true}}^k)\overline{C(X_{\text{adv}}^k,y_{\text{true}}^k)}},$$

其中，n 是图片个数，X^k 是数据集中的图像，y_{true}^k 是图像正确分类，x_{adv}^k 是相应的对抗样本，函数 $T(\cdot)$ 是对图像的任意转换，其形式包括打印和对结果拍照. 函数 $C(X,y)$ 是结果函数表示照片是否被正确分类：

$$C(X,y)=\begin{cases}1, & \text{如果 } X \text{ 被分成 } y\\ 0, & \text{若不是}\end{cases}.$$

照片的转换结果总结见表 1～表 3.

从表中我们可以看出快速法比迭代法生成的对抗样本图片对图像变换的鲁棒性更强，这是因为迭代方法使用不同种类的干扰，因此这些干扰更容易被图像变换所破坏. 而在某些情况下，对于已过滤的样本，对抗样本的破坏率比正常样本要高，由迭代方法生成的对抗样本，对抗样本被错误分类的概率在正常样本上竟整体比经滤波的样本低，这表明迭代方法生成的对抗样本做出的改变无法适应图像变换.

表 1 对抗样本在随机选择的图片上的各种方法的准确率

对抗样本的生成方法	照片				源图片			
	正常图片		对抗图片		正常图片		对抗图片	
	Top-1/%	Top-5/%	Top-1/%	Top-5/%	Top-1/%	Top-5/%	Top-1/%	Top-5/%
快速法 $\varepsilon=16$	79.8	91.9	36.4	67.6	85.3	94.1	36.3	58.8
快速法 $\varepsilon=8$	70.6	93.1	49.0	73.5	77.5	97.1	30.4	57.8
快速法 $\varepsilon=4$	72.5	90.2	52.9	79.4	77.5	94.1	33.3	51.0
快速法 $\varepsilon=2$	65.7	85.9	54.5	78.8	71.6	93.1	35.3	53.9
基本迭代法 $\varepsilon=16$	72.9	89.6	49.0	75.0	81.4	95.1	28.4	31.4
基本迭代法 $\varepsilon=8$	72.5	93.1	51.0	87.3	73.5	93.1	26.5	31.4
基本迭代法 $\varepsilon=4$	63.7	87.3	48.0	80.4	74.5	92.2	12.7	24.5
基本迭代法 $\varepsilon=2$	70.7	87.9	62.6	86.9	74.5	96.1	28.4	41.2
最小可能类法 $\varepsilon=16$	71.1	90.0	60.0	83.3	79.4	96.1	1.0	1.0
最小可能类法 $\varepsilon=8$	76.5	94.1	69.6	92.2	78.4	98.0	0	6.9
最小可能类法 $\varepsilon=4$	76.8	86.9	75.8	85.9	80.4	90.2	9.8	24.5
最小可能类法 $\varepsilon=2$	71.6	87.3	68.6	89.2	75.5	92.2	20.6	44.1

表 2 在滤波情况下的各种方法生成对抗图像的准确率(干净的图像正确分类,以数字或图像形式打印出的对抗样本图像被错误地分类)

对抗样本的生成方法	照片				源图片			
	正常图片		对抗图片		正常图片		对抗图片	
	Top-1/%	Top-5/%	Top-1/%	Top-5/%	Top-1/%	Top-5/%	Top-1/%	Top-5/%
快速法 $\varepsilon=16$	81.8	97.0	5.1	39.4	100	100	0	0
快速法 $\varepsilon=8$	77.1	95.8	14.6	70.8	100	100	0	0
快速法 $\varepsilon=4$	81.4	100.0	32.4	91.2	100	100	0	0
快速法 $\varepsilon=2$	88.9	99.0	49.5	91.9	100	100	0	0
基本迭代法 $\varepsilon=16$	93.3	97.8	60.0	87.8	100	100	0	0
基本迭代法 $\varepsilon=8$	89.2	98.0	64.7	91.2	100	100	0	0
基本迭代法 $\varepsilon=4$	92.2	97.1	77.5	94.1	100	100	0	0
基本迭代法 $\varepsilon=2$	93.9	97.0	80.8	97.0	100	100	0	0
最小可能类法 $\varepsilon=16$	95.8	100.0	87.5	97.9	100	100	0	0
最小可能类法 $\varepsilon=8$	96.0	100.0	88.9	97.0	100	100	0	0
最小可能类法 $\varepsilon=4$	93.9	100.0	91.9	98.0	100	100	0	0
最小可能类法 $\varepsilon=2$	92.2	99.0	93.1	98.0	100	100	0	0

表 3 用不同方法生成对抗样本与原始图片的破坏率

对抗样本的生成方法	正常情况		过滤情况	
	Top-1/%	Top-5/%	Top-1/%	Top-5/%
快速法 $\varepsilon=16$	12.5	40.0	5.1	39.4
快速法 $\varepsilon=8$	33.3	40.0	14.6	70.8
快速法 $\varepsilon=4$	46.7	65.9	32.4	91.2
快速法 $\varepsilon=2$	61.1	63.2	49.5	91.9
基本迭代法 $\varepsilon=16$	40.4	69.4	60.0	87.8
基本迭代法 $\varepsilon=8$	52.1	90.5	64.7	91.2
基本迭代法 $\varepsilon=4$	52.4	82.6	77.5	94.1
基本迭代法 $\varepsilon=2$	71.7	81.5	80.8	96.9
最小可能类法 $\varepsilon=16$	72.2	85.1	87.5	97.9
最小可能类法 $\varepsilon=8$	86.3	94.6	88.9	97.0
最小可能类法 $\varepsilon=4$	90.3	93.9	91.9	98.0
最小可能类法 $\varepsilon=2$	82.1	93.9	93.1	98.0

5 对抗样本的防御技术

让机器学习模型更加稳健的传统技术,比如权重衰减、随机丢弃神经元或神经元连接边,通常无法切实防范对抗样本. Szegedy 等人在文献[5]上通过实验结果观察到对抗样本是普遍存在的,并不是某个模型的过拟合或某个模型在数据集上的特定选择等效于对抗样本问题. 此外,文献[5]中提到了对抗样本可迁移性的特征,即相同的干扰能够使不同的神经网络对相同的输入产生错误分类. 机器学习模型经常受到对抗样本的干扰,对抗样本通过恶意的干扰输入以便在测试时误导模型[5,8,21,29]. 对抗样本是对机器学习系统实际部署的安全威胁. 值得注意的是,这些输入在模型之间变换,从而对部署的模型实施黑盒攻击[5,13,22].

针对对抗样本的防御技术主要有:基于正则化方法、对抗性的预处理训练方法、蒸馏方法,拒绝分类方法等,如图 6 所示.

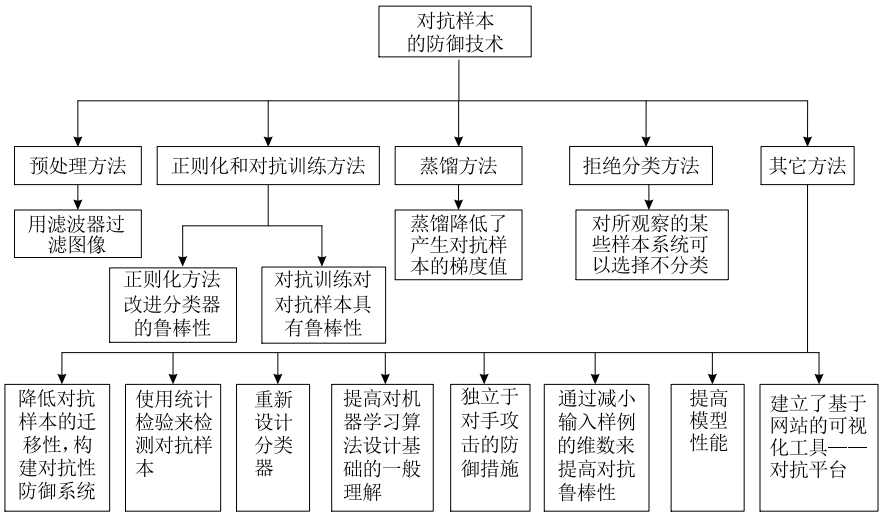


图 6　对抗样本的防御技术

5.1　预处理方法

自然图像具有特殊的属性，如相邻像素之间的高相关性、高频域中低能量性[30]. 假设这种对抗性的改变并不存在于与自然图像相同的空间，一些工作[31-35]考虑通过用滤波器滤波图像来去除对抗样本的干扰. 虽然预处理输入使攻击更具挑战性，但它并没有消除被攻击的可能性. 此外，这些过滤器通常会降低未受对抗干扰的数据的分类精度.

5.2　正则化和对抗训练方法

一些研究提出通过正则化方法[35-37]和对抗性训练方法[20,38-44]改进分类器的鲁棒性. 既能持续不断地更新对抗样本来攻击当前的模型，也可以使用这种带有对抗样本的模型来训练神经网络使其降低错误率，也就是说这种模型能在某种程度上对抗对抗样本. 所以通过对抗训练能够在某种程度上提高深度学习对对抗样本的抗干扰能力. 对抗样本本来是个系统漏洞，但我们可以利用它使之成为人类对抗对抗样本的手段. 对抗样本并不限于一个具体的神经网络，因此制造对抗样本也不需要获得该模型的源代码. 只要模型是被训练来执行相同的任务的，它们就会被同样的对抗样本“欺骗”，即使这些模型有不同的结构或使用了不同的训练样本. 因此，人们只要设计一个模型，产生出相应的对抗样本，就能用这些样本攻击那些相似任务的人工智能算法. 对抗样本问题很难用常规的办法解决，有研究组尝试了多种传统手段，包括多个模型取平均值、同一图像多次判断取平均值、带噪音训练和构造生成模型等等，都不能解决对抗样本问题. 有针对性地专门训练可以让模型的抵抗力更强，但也无法真正消灭盲区. 虽然人们很难被这些样本欺骗，但有时候也会在意想不到的地方跌倒，心理学已经提供了浩如烟海的关于视觉错觉的例子，这些错觉可以认为是人类特有的“对抗样本”问题，但是面对神经网络的对抗样本和面对人类的对抗样本二者并不完全重合. 人和机器都会犯错，但犯的错误不一样.

Kurakin 等人对对抗学习做了进一步探索[45]，表明对抗训练对对抗样本具有鲁棒性，且对大型模型及数据集进行对抗训练提出了建议. 另外，他提出在进行对抗样本训练时，提高模型的性能可以增加对抗样本的鲁棒性. 不同的对抗样本干扰程度会有不同的对抗力量来抵制攻击，基于这种现象，Song 等人将对抗训练样本和不同的对抗力量相结合提出多力量对抗训练方法(Multi-strength Adversarial Training Method，MAT)来减轻对抗样本的攻击[46]. Tramèr 等人在文献[47]中指出经过对抗训练的模型仍然容易受到攻击，因为该模型的判别超平面在数据点附近明显变化，从而阻碍了基于模型损失的一阶近似攻击，但其无法拒绝来自对抗样本迁移的黑盒攻击. Song 等人对对抗训练进一步推广，提出集成对抗训练方法(Ensemble Adversarial Training)，该方法通过将某些固定预训练模型的干扰输入增加到训练数据中来提高训练效果.

然而实验表明，基于对抗样本训练的分类错误率比仅在未受干扰数据上分类错误率低 1%[27]，这意味着将这些学习系统应用于实际时，不能完全依赖防御机制. 文献中的实验表明，在模型未知时，我们也可以成功地对鲁棒学习系统进行攻击，因此，这些方法并不能有效地抵御对抗样本的攻击.

5.3　蒸馏方法

蒸馏方法是 Hinton 等人提出来的一种用小模

型模仿大模型的方法[48]，基本思路为：训练分类模型输出的独热向量，称为硬标签，用硬标签对一个模型进行训练后，不仅保留软最大函数后最大概率的维度，也将整个概率向量作为标签（称软标签），这样来说，每个输入样本不仅只有一个信息量较小的独热向量，还有一个对每个类别都有一定概率的向量.这样训练网络就会得到一些附加信息，若一张图片在某两类之间难以分别，它们就会有较高的概率，这样的标签实际上附带了大模型训练得到的信息，因此可以提高小模型的效果.

Papernot 等人针对深度神经网络对对抗样本提出了蒸馏防御机制，并在两类深度神经网络上验证了其防御机制的有效性，这是由于蒸馏降低了产生对抗样本时的梯度值，同时提高了产生对抗样本所需修改特征样例的最小平均值[21]. 而 Carlini 等人指出对手能通过访问分类器参数恢复蒸馏的效果[49]，且研究了蒸馏法在黑箱和未知模型中的应用，验证了该方法不能提高分类器的鲁棒性并指出蒸馏法并不十分有效[50]. 通过在标准的攻击上稍加修改蒸馏机制，就不足以对抗对抗样本攻击，原因是对于小的输入变化，虽然蒸馏法能显著地降低损失函数梯度值，但在黑盒攻击和未知模型函数情况下，特征值的改变并不能有效地抵抗对手攻击.

5.4 拒绝分类方法

一些学者研究了拒绝分类方法[51-55]，即对所观察的某些样本系统可以选择不分类. 例如，当类条件后验概率接近时，就选择拒绝选项. 在文献[56]中，作者指出正确分类的样本往往比错误分类和没有在概率分布中的样本有更大的最大类条件后验概率，因此，通过检查相应的类条件后验概率来拒绝输入样本，对于检测对抗样本是无效的，类似的方法是将无效数据归类为“垃圾类”[57-59]. 在这些方法中，通过对分类器进行训练，将无效数据均匀分布，然后在测试时，以低置信度丢弃样本. 虽然这种方法提高了对抗样本分类器的鲁棒性，但它未考虑检测样本丢弃的概率. 在文献[60]中，作者提出一个检测器子网络（二值分类器）来增广分类器，并用其将对抗样本和干净样本区分开，然而作者没有给出该方法的错误率，即将未受干扰的数据看成对抗样本的错误概率，此外，值得注意的是，对抗样本不仅能“欺骗”分类器同样也能“欺骗”检测器，因此这一方法也不十分有效.

5.5 其它方法

(1) 降低对抗样本的迁移性，构建对抗性防御系统

由于针对对抗样本的很多防御措施都失败了，Carlini 等人从另一个角度出发，研究如何检测对抗样本并提出了十种检测方案[61]. 通过研究这些防御措施可以发现，如果能够找到一种降低迁移性的方法，就可以通过基于随机化的防御措施检测到对抗样本. 在所有的这些防御措施中，使用另一个神经网络找出对抗样本的效果最差，虽然这些防御措施在 MNIST 数据集上对于较弱的攻击表现出很强的鲁棒性，但当受到较强的攻击时，这些防御措施都失败了. 这种结果并不令人吃惊：使用神经网络能够从输入数据中提取层次多且抽象有意义的特征，但当操作对象是原始图像时，一个简单的线性分类器并不起作用，故当检测对抗样本时也同样不起作用. 因此，对上述防御措施，对手即使不知道模型参数也能通过迁移性攻破防御. Carlini 等人指出这些方法都不能抵御白盒攻击，甚至在黑盒攻击中也都失败了，其主要原因是现有的防御措施缺乏足够的安全评估，据此，Carlini 等人对开发新的防御手段提出了几点建议：

① 使用更强的攻击进行测试；
② 验证白盒攻击是否失败；
③ 验证黑盒攻击是否失败；
④ 计算对抗样本的平均失真值；
⑤ 计算估计的正确率及错误率；
⑥ 在 CIFAR 数据集上进行测试；
⑦ 开放源代码以便大家参与进来进行验证比较.

开发出好的防御措施以对抗对抗样本，远比我们之前想象的要难. 现有的防御缺乏全面的安全评估，因此在深度神经网络应用于潜在的关键安全领域时，构建对抗性防御系统是必须的.

Xie 等人指出对抗干扰能在基于不同结构、不同识别任务上的不同训练数据的网络间迁移[27]，这种迁移性表明深度神经网络虽然初始值不同，训练方式不同，但却有一些共性，如局部线性，这使得深度神经网络对相似的干扰源很敏感，总结各种干扰能让迁移性更好，这提供了一种对抗样本黑盒攻击更有效的干扰实现机制.

Goodfellow 等人提出解决对抗样本可迁移性是问题的关键，并在文献[62]中对对抗样本空间做了探索，其估计了对抗样本输入空间的维数，发现了对抗样本张成了大维度空间中的一个连续子空间，提出了寻找对抗样本多个独立攻击的方法，以及解释了对抗样本可迁移性产生的原因，即不同的模型共享了该空间中的重要部分. 此外，该研究还发现不同模型学习到的判别边界非常接近，并总结了限制对抗样本迁移的条件：必须满足数据分布的充分条

件意味着简单模型类之间的可迁移性不成立；存在模型能抵御可迁移性的攻击.

Hosseini 等人指出保护黑盒系统免遭对抗样本迁移性的干扰，等价于阻止对抗样本的迁移[63]. 基于这种想法，他提出了模仿人类推理的训练方法，用一个附加的标签来增广分类器的输出类集，并适当地训练分类器以恰当的概率给新标签分配对抗样本. Hossein 等人提出了一种训练方法使分类器平滑地输出较低置信度的原始标签，而不是将该输入检测为拒绝样本. 实际上，作者用一个空标记来增加输出类集，并训练分类器，将对抗样本分类成空标签来拒绝对抗样本. 因此，可将该方法广泛应用在基于黑盒系统的对抗攻击上，使分类器能够有效抵抗对抗样本的干扰，同时也保持不破坏数据分类的准确性.

(2) 使用统计检验来检测对抗样本

作为研究防御模型的第一步，Grosse 等人指出，对抗样本和原始数据不是来自同一概率分布[64]，因此可以使用统计检验来检测对抗样本. 此外，Grosse 等人通过将机器学习模型增加一个额外类来训练分类模型对抗所有的对抗样本，提高样本上学习模型的鲁棒性.

(3) 重新设计分类器

不同于直接训练深度神经网络去测试对抗样本，Li 等人基于卷积层的输出分析提出了一个更简单的方案，并通过特殊的对抗生成机制设计了一个级联分类器[65]，该分类器能同时有效地检测对抗样本. 实验表明，对抗样本可以通过小的平均滤波在图像上重构. 这些发现也促使我们对深度卷积神经网络分类机制进行进一步思考.

Fawzi 等人侧重对分类器的鲁棒性进行探索[66]，其认为在分类器性能及鲁棒性之间应有一个权衡. Huang 等人对这种权衡在实际应用中做了进一步探索[42]，其把学习过程看成是最小-最大问题，并考虑在最坏情况下的深度神经网络学习问题，即允许对手在每一个数据点上作出不同的干扰，学习过程就是在预期的干扰上最小化损失误差，并把这种学习过程称为“对抗学习”.

(4) 提高对机器学习算法设计基础的一般理解

Wang 等人指出机器学习对对抗样本的攻击表现效果差是由于总体上缺乏对机器学习算法设计基础的一般理解[67]. 通过引入一种新的理论框架来分析算法脆弱性的原因，并将这种理论应用在最近邻算法中，结果表明其鲁棒性比标准的最近邻算法好. Gu 等人也认为神经网络易受对抗样本的干扰主要与训练过程和目标函数相关，而不是深度模型的拓扑结构[37]，其认为提出一个适当的训练过程和目标函数是提高神经网络鲁棒性的关键. Gu 等人提出了深层收缩网络(Deep Contractive Networks)来明确学习每层的不变特征，实验结果初步验证其有效性. Goodfellow 等人指出虽然现在还没有模型既能有一定的预测准确性，又能成功抵制对抗样本的攻击，但从长远看提高对机器学习算法设计基础的一般理解是抵御对抗样本攻击的根本方法.

(5) 独立于对手攻击的防御措施

Meng 等人主张对对抗样本的防御应是独立于对手攻击的，而不是寻找特定生成对抗样本的属性[68]. 通过在对抗样本的生成过程中找出内在的共同属性，使防御措施更具有可转移性，适用范围更宽. Meng 等人提出的防御方法是朝着这个目标迈出的第一步，他提出神经网络分类器对抗对抗样本的防御框架：MagNet. MagNet 不修改受保护的分类器也不知道对抗样本的生成过程，但其包含一个或多个分开的探测器网络及一个重构网络. MagNet 通过正常样本的流型估计来区分正常及对抗样本. 由于它不依赖于任何对抗样本的生成过程，所以它具有强大的泛化能力. 而且，MagNet 通过将对抗样本变换为流形空间来重构对抗样本，这对带有小扰动的对抗样本实现正确分类来说是一个有效的方法. 此外，这篇文献还讨论了防御白盒攻击的难度，提出了一个机制来抵御可疑攻击. 受密码学中随机性的启发，使用多样性来增强 MagNet，实验表明，MagNet 对样例黑箱和可疑的攻击是有效的，同时对正常样例的分类错误率也很低.

(6) 通过减小输入样例的维数来提高对抗鲁棒性

Fawzi 等人展示了深度神经网络能通过减小输入样例的维数来提高对抗鲁棒性[69]，对抗样本是通过对输入数据作出微小扰动生成的，这些改变虽然肉眼很难察觉，但却给神经网络带来很大灾难，这是高维空间中拟线性分类器的通用特性. 因此，通过降低输入的维数来减小对抗样本存在的“空间”在直觉上是可能的，Maharaj 在 CIFAR10 的 5 层神经网络数据集及 CaffeNet 和 GoogLeNet 在 Tiny-Imagenet-200 数据集上验证了其鲁棒性[32]. Bastani 等人给出了神经网络鲁棒性的度量指标[70]，并将基于编码鲁棒性的度量指标看成某个线性规划问题，通过实验展示了在数据集 MNIST 和 CIFAR-10 上，提出的度量指标是如何用来估计深度神经网络的鲁棒性

的.另外,Bastani 等人还展示了现有的提高鲁棒性的方法是通过"过度拟合"某个特定算法生成的对抗样本来实现的.最后,文献也说明了通过 Bastani 等人的方法不仅可以提高神经网络鲁棒性度量指标,也可以改善之前提出的鲁棒性度量指标.Nayebi 等人受神经回路中非线性树突计算的生物物理原理启发,提出了通过在饱和区寻找高度非线性网络来抵抗对抗样本攻击深层神经网络的方案[71].通过这种方案不仅提高了在由 FGS 法生成的对抗样本上作对抗训练网络的性能,且进一步通过信息几何方法确定了这些网络是如何实现鲁棒性的.Nayebi 等人发现这些网络对输入维数的高度压缩非常不敏感,此外,他们发现在网络中如同大脑一样采用峰度系数的权值分布是如何保护深度神经网络,甚至是线性分类器免遭对抗样本的攻击的.在未来的研究中,可以进一步将理论与实验相结合形成对抗鲁棒性的生物学可信机制.

(7)提高模型性能

Rozsa 等人指出对抗样本在相似的神经网络中最容易迁移,但不容易攻击较好的学习模型[72].对于给定的分类任务,好的模型能学习特征映射,使得各类之间更加分离,因此,好的模型不仅有更好的预测准确度也能提高系统鲁棒性,所以提高模型性能也是对抗对抗样本的重要手段.Lu 等人提出了新的网络——安全网(SafetyNet)来抵制对抗样本的攻击[73],并经过合理的分析及实验证明 SafetyNet 网络的鲁棒性.这个安全网的体系结构采用一个新的技术——场景证明(sceneproof),该技术可以可靠地检测图像是否是真实的图像场景.Cisse 等人也提出了帕塞(Parseval)网络来提高鲁棒性[74].

(8)建立基于网站的可视化工具—对抗平台

Norton 等人建立了一个基于网站的可视化工具—对抗平台[75-76],展示了常见的对抗卷积神经网络算法的有效性,为用户在探索生成对抗样本算法上提供了经验.

6 对抗样本的应用

截至目前,对抗样本的应用主要是用在对抗评估及对抗训练上.

(1)对抗评估

对抗评估有助于模型尽早地分析错误,并判断模型是否成功.Smith 等人提到对抗评估的中心思想是通过不同的学者研究不同的对抗角色,使评价不同角色的分工明确且发挥最大贡献[77].基于不同学者及其模型,在不同的对抗角色上,Smith 等人提出了新的自然语言评价模型.Jia 等人在文献[78]中提到对对抗评估来说,阅读理解是一个很有吸引力的平台,在计算机视觉中是通过对输入图像加入细微的对抗干扰,但这种细小的干扰不能改变一个图像的真实标签,然而,改变一段话中的某个单词却能够完全改变句子的意义,通过在输入图片中增加一段分散的句子,而非添加一个保留语义的干扰来生成对抗样本,这样生成的句子能够使图片"迷惑"模型,但其与正确的分类结果并不矛盾,且不能"迷惑"人类.Jia 等人提到尽管阅读理解系统在标准的评价系统中非常成功,但是在对抗评估中表现却很差.标准评估对依赖字面的提示模型过于宽容,相反,对抗评估揭示了现存模型对于改变语义的干扰过于稳定.Kannan 等人在文献[79]中指出递归神经网络编码器-解码器模型在数据驱动的对话系统中取得了显著的进展,但对话结果的评价仍是具有挑战性的问题.对抗损失是一种评价生成对话结果是否更像人类表达的直接方式,其将减少人类参与对话评估的需求,通过训练递归神经网络来区别对话模型样本和人类生成的样本,虽然可以证明这种方式是可行的,但在实际应用中还存在诸多问题.

(2)对抗训练

Nøkland 通过将对抗梯度增加到训练过程中来提高反向传播算法的性能[44].不同于改变输入来提高通用性,Park 等人通过最大化丢弃网络的输出与监督网络输出之间的差异得到一组最小的丢弃集合.这些确认的丢弃集合用于重新训练神经网络,实验表明在配置好的子网络上训练能提高在数据集 MNIST 和 CIFAR-10 上监督和半监督学习模型的泛化能力[80].

在实际的分类任务中,很难从所有可能的环境类别中收集训练样本.因此,当一个不可见类的样例出现时,一个好的分类器应该能够判断它是未知类,而不是把它分类为任何已知的类别.Yu 等人利用对抗训练的思想提出对抗样本生成(Adversarial Sample Generation,ASG)框架用于类数不确定的分类.ASG 通过对抗训练策略以无监督的方式产生已知类别的正负样例.对于生成的样例,ASG 通过有监督的方式分辨它们所属的类[81].

(3)其它方面

Lee 等人利用对抗样本提出流形正则化网络(Manifold Regularized networks,MRnet)来学习更

有意义的表示[82]，将 MRnet 最小化样本及对抗样本的多层嵌入结果之间的差作为训练目标函数. 实验结果表明，MRnet 网络更加适应对抗样本，且有利于学习流形上的嵌入表示.

7　对抗样本的未来发展方向

综上所述，研究对抗样本已经是机器学习安全领域的热点问题，而找出对抗样本的特性、对抗样本的生成机理及对抗样本的攻击方式是研究对抗样本的重点问题，探索不同对抗性样本攻击的防御算法是主要目标，将这两部分结合解决对抗样本的攻击是今后对抗样本的主要研究方向.

(1) 如何定义对抗样本的迁移性，如何度量迁移程度，如何确定迁移的上下界，以便利用迁移性实现有效地生成对抗样本、有效地探测对抗样本、防御对抗样本的攻击.

(2) 一般数据或特定数据上对抗样本产生的原因及原理，使得分类器无法正确识别对抗样本. 建立数学理论完备的、统一的对抗样本生成式理论，进而实现高概率担保的防御对抗样本攻击的深度神经网络实现理论和实现形式，为深度神经网络机器学习算法实际布署奠定理论基础.

(3) 构造普适的生成对抗样本的基准软件平台，使得目前对抗样本的研究能够在统一的标准数据集上，评测实验结果.

(4) 为了对抗对抗样本的攻击，深度神经网络学习算法首先需要分辨正常样本和对抗样本，需要检测甄别对抗样本，统计假设检验中序列假设测试和序列改变点测试，多比较过程、多假设检验理论可能会提供理论借鉴.

(5) 拓展对抗样本的应用领域. 对抗样本本身的研究可能带动机器学习算法本身的研究，进而影响到机器学习算法的应用.

8　结论与展望

对抗样本问题越来越引人关注，探讨对抗样本出现的原因及生成方式是研究对抗样本的关键问题. 本文首先总结了对抗样本的出现原因及最新的研究进展，并指出当前的猜想都不能令人信服，进一步研究对抗样本出现的原因是未来重要的研究内容；其次，对对抗样本的主要生成方式即 F-BFGS 方法、FGS 方法、迭代方法、迭代最小可能类方法及其它方法进行了详细的述评，指出了其优缺点和适用的场景；研究对抗样本出现的原因和生成方法的目的是使机器学习系统免受对抗样本的攻击，本文最后对目前流行的主要防御技术即基于正则化方法、对抗性的预处理训练方法，蒸馏方法、拒绝分类方法和其它方法进行了评述，指出了不同防御措施的适用场景与不足，阐释了上述防御措施均不能完全避免对抗样本的攻击. 综上所述，进一步研究对抗样本的特性，给出其具有实际应用前景的数学描述，探讨普适性的对手样本生成方法，彻底解决对抗攻击问题，仍有大量的理论和实践问题需要解决.

参　考　文　献

[1] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift// Proceedings of the 32nd International Conference on Machine Learning. Lille, France, 2015: 448-456

[2] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. Journal of Nature, 2015, 518(7540): 529-533

[3] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 2014, 15(1): 1929-1958

[4] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770-778

[5] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks. arXiv preprint arXiv: 1312.6199, 2013

[6] Li Pan, Zhao Wen-Tao, Liu Qiang, et al. Review of machine learning security and its defense technology. Computer Science and Exploration, 2018, 12(2): 171-184(in Chinese)
(李盼，赵文涛，刘强等. 机器学习安全性问题及其防御技术研究综述. 计算机科学与探索，2018，12(2)：171-184)

[7] Kurakin A, Goodfellow I, Bengio S. Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533, 2016

[8] Biggio B, Corona I, Maiorca D, et al. Evasion attacks against machine learning at test time. arXiv preprint arXiv: 1708.06131, 2017

[9] Nguyen A, Yosinski J, Clune J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2014: 427-436

[10] Wang Pei. Why learning programs will be fooled. Science and Technology Review, 2016, 34(7): 88-89(in Chinese)
(王培. 为何学习程序会被愚弄. 科技导报，2016，34(7)：88-89)

[11] Tanay T, Griffin L. A boundary tilting persepective on the phenomenon of adversarial examples. arXiv preprint arXiv: 1608.07690, 2016

[12] Papernot N, Mcdaniel P, Goodfellow I, et al. Practical black-box attacks against deep learning systems using adversarial examples. arXiv preprint arXiv:1602.02697, 2016

[13] Moosavi-Dezfooli S M, Fawzi A, Fawzi O, et al. Universal adversarial perturbations//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2016: 86-94

[14] Liu Y, Chen X, Liu C, et al. Delving into transferable adversarial examples and black-box attacks. arXiv preprint arXiv:1611.02770, 2017

[15] Fischer V, Kumar M C, Metzen J H, et al. Adversarial examples for semantic image segmentation. arXiv preprint arXiv:1703.01101, 2017

[16] Carlini N, Mishra P, Vaidya T, et al. Hidden voice commands //Proceedings of the 25th USENIX Security Symposium. Austin, USA, 2016: 513-530

[17] Smith D F, Wiliem A, Lovell B C. Face recognition on consumer devices: Reflections on replay attacks. Journal of the IEEE Transactions on Information Forensics and Security, 2015, 10(4): 736-745

[18] Sharif M, Bhagavatula S, Bauer L, et al. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition//Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. Vienna, Austria, 2016: 1528-1540

[19] Lu J, Issaranon T, Forsyth D. SafetyNet: Detecting and rejecting adversarial examples robustly//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 446-454

[20] Papernot N, Mcdaniel P, Goodfellow I, et al. Practical black-box attacks against machine learning//Proceedings of the IEEE European Symposium on Security and Privacy. Saarbrücken, Germany, 2016: 506-519

[21] Papernot N, Mcdaniel P, Wu X, et al. Distillation as a defense to adversarial perturbations against deep neural networks//Proceedings of the IEEE Symposium on Security and Privacy. San Jose, USA, 2016: 582-597

[22] Mahendran A, Vedaldi A. Understanding deep image representations by inverting them//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 5188-5196

[23] Rozsa A, Rudd E M, Boult T E. Adversarial diversity and hard positive generation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Las Vegas, USA, 2016: 410-417

[24] Hu W, Tan Y. Black-box attacks against RNN based malware detection algorithms. arXiv preprint arXiv: 1705.08131, 2017

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

[26] Kos J, Fischer I, Song D. Adversarial examples for generative models. arXiv preprint arXiv:1702.06832, 2017

[27] Xie C, Wang J, Zhang Z, et al. Adversarial examples for semantic segmentation and object detection//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 1378-1387

[28] Cisse M, Adi Y, Neverova N, et al. Houdini: Fooling deep structured prediction models. arXiv preprint arXiv: 1707.05373, 2017

[29] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014

[30] Papernot N, Mcdaniel P, Jha S, et al. The limitations of deep learning in adversarial settings//Proceedings of the IEEE European Symposium on Security and Privacy. Saarbrücken, Germany, 2016: 372-387

[31] Bovik A C. Handbook of Image and Video Processing. Berlin, Germany: Springer, 2014

[32] Ross A S, Doshivelez F. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients//Proceedings of the 32nd Conference on Artificial Intelligence, New Orleans, USA, 2018: 1660-1669

[33] Dziugaite G K, Ghahramani Z, Roy D M. A study of the effect of JPG compression on adversarial images. arXiv preprint arXiv:1608.00853, 2016

[34] Wang Q, Guo W, Zhang K, et al. Random feature nullification for adversary resistant deep architecture. arXiv preprint arXiv:1610.01239, 2016

[35] Wang Q, Guo W, Ii A G O, et al. Using non-invertible data transformations to build adversarial-robust neural networks. arXiv preprint arXiv:1610.01934, 2016

[36] Graese A, Rozsa A, Boult T E. Assessing threat of adversarial examples on deep neural networks//Proceedings of the 15th IEEE International Conference on Machine Learning and Applications. Anaheim, USA, 2017: 69-74

[37] Gu S, Rigazio L. Towards deep neural network architectures robust to adversarial examples. arXiv preprint arXiv:1412.5068, 2014

[38] Zhao Q, Griffin L D. Suppressing the unusual: Towards robust CNNs using symmetric activation functions. arXiv preprint arXiv:1603.05145, 2016

[39] Rozsa A, Gunther M, Boult T E. Towards robust deep neural networks with BANG. arXiv preprint arXiv: 1612.00138, 2016

[40] Shaham U, Yamada Y, Negahban S. Understanding adversarial training: Increasing local stability of neural nets through robust optimization. arXiv preprint arXiv:1511.05432, 2016

[41] Lyu C, Huang K, Liang H N. A unified gradient regularization family for adversarial examples//Proceedings of the 2015 IEEE International Conference on Data Mining. Atlantic City, USA, 2016: 301-309

[42] Huang R, Xu B, Schuurmans D, et al. Learning with a strong adversary. arXiv preprint arXiv:1511.03034, 2015

[43] Miyato T, Maeda S, Koyama M, et al. Distributional smoothing with virtual adversarial training. arXiv: 1507.00677, 2015

[44] Nøkland A. Improving back-propagation by adding an adversarial gradient. arXiv preprint arXiv:1510.04189, 2015

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

[46] Song C, Cheng H P, Wu C, et al. A multi-strength adversarial training method to mitigate adversarial attacks. arXiv preprint arXiv:1705.09764, 2017

[47] Tramèr F, Kurakin A, Papernot N, et al. Ensemble adversarial training: Attacks and defenses. arXiv preprint arXiv:1705.07204, 2017

[48] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network, in Deep Learning and Representation Learning Workshop at NIPS 2014. arXiv preprint arXiv:1503.02531, 2014

[49] Carlini N, Wagner D. Towards evaluating the robustness of neural networks//Proceedings of the 2017 IEEE Symposium on Security and Privacy. San Jose, USA, 2017: 39-57

[50] Carlini N, Wagner D. Defensive distillation is not robust to adversarial examples. arXiv preprint arXiv: 1607.04311, 2016

[51] Elkan C. The foundations of cost-sensitive learning//Proceedings of the 17th International Joint Conference on Artificial Intelligence. Seattle, USA, 2001: 973-978

[52] Fumera G, Roli F, Giacinto G, et al. Reject option with multiple thresholds. Pattern Recognition, 2000, 33(12): 2099-2101

[53] Herbei R, Wegkamp M H. Classification with reject option. Canadian Journal of Statistics, 2010, 34(4): 709-721

[54] Bartlett P L, Wegkamp M H. Classification with a reject option using a hinge loss. Journal of Machine Learning Research, 2008, 9: 1823-1840

[55] Cortes C, Desalvo G, Mohri M. Learning with rejection//Proceedings of the 27th International Conference. Bari, Italy, 2016: 67-82

[56] Hendrycks D, Gimpel K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv preprint arXiv:1610.02136, 2016

[57] Bromley J, Denker J. Improving rejection performance on handwritten digits by training with "Rubbish". Journal of Neural Computation, 1993, 5(3): 367-370

[58] Lécun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998, 86(11): 2278-2324

[59] Yadav B, Devi V S. Novelty detection applied to the classification problem using probabilistic neural network//Proceedings of the 2014 IEEE Symposium on Computational Intelligence and Data Mining. Orlando, USA, 2014: 265-272

[60] Gong Z, Wang W, Ku W S. Adversarial and clean data are not twins. arXiv preprint arXiv:1704.04960, 2017

[61] Carlini N, Katz G, Barrett C, et al. Ground-truth adversarial examples. arXiv preprint arXiv:1709.10207, 2017

[62] Tramèr F, Papernot N, Goodfellow I, et al. The space of transferable adversarial examples. arXiv preprint arXiv: 1704.03453, 2017

[63] Hosseini H, Chen Y, Kannan S, et al. Blocking transferability of adversarial examples in black-box learning systems. arXiv preprint arXiv:1703.04318, 2017

[64] Grosse K, Manoharan P, Papernot N, et al. On the (statistical) detection of adversarial examples. arXiv preprint arXiv: 1702.06280, 2017

[65] Li X, Li F. Adversarial examples detection in deep networks with convolutional filter statistics//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 5775-5783

[66] Fawzi A, Fawzi O, Frossard P. Analysis of classifiers' robustness to adversarial perturbations. arXiv preprint arXiv:1502.02590, 2017

[67] Wang Y, Jha S, Chaudhuri K. Analyzing the robustness of nearest neighbors to adversarial examples. arXiv preprint arXiv:1706.03922, 2017

[68] Meng D, Chen H. MagNet: A two-pronged defense against adversarial examples//Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. Dallas, USA, 2017: 135-147

[69] Alhussein F, Seyed-Mohsen M-D, Pascal F. Robustness of classifiers: From adversarial to random noise//Proceedings of the 30th Conference on Neural Information Processing Systems. Barcelona, Spain, 2016: 1624-1632

[70] Bastani O, Ioannou Y, Lampropoulos L, et al. Measuring neural net robustness with constraints//Proceedings of the Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016. Barcelona, Spain, 2016: 2613-2621

[71] Nayebi A, Ganguli S. Biologically inspired protection of deep networks from adversarial attacks. arXiv preprint arXiv: 1704.01547, 2017

[72] Rozsa A, Günther M, Rudd E M, et al. Are facial attributes adversarially robust?//Proceedings of the 23rd International Conference on Pattern Recognition. Cancún, Mexico, 2016: 3121-3127

[73] Lu J, Sibai H, Fabry E, et al. No need to worry about adversarial examples in object detection in autonomous vehicles. arXiv preprint arXiv:1707.03501, 2017

[74] Moustapha C, Piotr B, Edouard G, et al. Parseval networks: Improving robustness to adversarial examples//Proceedings of the 34th International Conference on Machine Learning. Sydney, NSW, Australia, 2017: 854-863

[75] Norton A, Qi Y. Adversarial-playground: A visualization suite for adversarial sample generation. arXiv preprint arXiv: 1706.01763, 2017

[76] Norton A P, Qi Y. Adversarial-playground: A visualization suite showing how adversarial examples fool deep learning// Proceedings of the IEEE Symposium on Visualization for Cyber Security. Phoenix, USA, 2017: 1-4

[77] Smith N A. Adversarial evaluation for models of natural arXiv preprint arXiv:1207.0245, 2012

[78] Jia R, Liang P. Adversarial examples for evaluating reading comprehension systems//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: Association for Computational Linguistics, 2017: 2021-2031

[79] Kannan A, Vinyals O. Adversarial evaluation of dialogue models. arXiv preprint arXiv:1701.08198, 2017

[80] Park S, Park J K, Shin S J, et al. Adversarial dropout for supervised and semi-supervised learning. arXiv preprint arXiv:1707.03631, 2017

[81] Yu Y, Qu W Y, Li N, et al. Open-category classification by adversarial sample generation//Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne, Australia, 2017: 3357-3363

[82] Lee T, Choi M, Yoon S. Manifold regularized deep neural networks using adversarial examples. Computer Science, arXiv preprint arXiv:1511.06381, 2015

ZHANG Si-Si, Ph. D. candidate. Her main research interest is machine learning.

ZUO Xin, professor. His research interests are process control and real-time optimization, and reliability analysis.

LIU Jian-Wei, Ph. D., associate professor. His main research interests include machine learning intelligent information processing, analysis, prediction, controlling of complicated nonlinear system, and analysis of the algorithm and the designing.

Background

The development of science and technology makes artificial intelligence more and more close to human life. Adversarial examples are a hot issue in the field of machine learning security. More and more attention has been paid to the problem of adversarial examples. The reasons for the emergence of the adversarial examples and the way of generation are the key problems in the study of the adversarial examples.

This article summarizes the characteristics, generation, attack methods and the application of the adversarial examples. At the end of this paper, the future research direction of the adversarial examples is prospected.