



面向深度学习模型的对抗攻击与防御方法综述

姜妍, 张立国

(哈尔滨工程大学 计算机科学与技术学院, 哈尔滨 150001)

摘要: 深度学习作为人工智能技术的重要组成部分,被广泛应用于计算机视觉和自然语言处理等领域。尽管深度学习在图像分类和目标检测等任务中取得了较好性能,但是对抗攻击的存在对深度学习模型的安全应用构成了潜在威胁,进而影响了模型的安全性。在简述对抗样本的概念及其产生原因的基础上,分析对抗攻击的主要攻击方式及目标,研究具有代表性的经典对抗样本生成方法。描述对抗样本的检测与防御方法,并阐述对抗样本在不同领域的应用实例。通过对对抗样本攻击与防御方法的分析与总结,展望对抗攻击与防御领域未来的研究方向。

关键词: 人工智能;深度学习;对抗攻击;安全防御;对抗样本

开放科学(资源服务)标志码(OSID):



中文引用格式: 姜妍,张立国.面向深度学习模型的对抗攻击与防御方法综述[J].计算机工程,2021,47(1):1-11.

英文引用格式: JIANG Yan, ZHANG Ligu. Survey of adversarial attacks and defense methods for deep learning model[J]. Computer Engineering, 2021, 47(1): 1-11.

Survey of Adversarial Attacks and Defense Methods for Deep Learning Model

JIANG Yan, ZHANG Ligu

(College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China)

[Abstract] As an important part of artificial intelligence technology, deep learning is widely used in computer vision, natural language processing and other fields. Although deep learning performs well in tasks such as image classification and target detection, its application security is potentially threatened by adversarial attacks, which further affects the security of the deep learning model itself. To address the problem, this paper briefly introduces the concept of adversarial sample and its causes, and on this basis analyzes the main attack modes and the targets of adversarial attacks. Then the paper studies the typical generation methods of adversarial samples, describes the detection methods and defense methods for adversarial samples, and represents the application cases of adversarial samples in the different fields. Finally, based on the analysis and summary of the attack modes and defense methods of adversarial attacks, the paper discusses the future directions of adversarial attack and defense research.

[Key words] artificial intelligence; deep learning; adversarial attacks; security defense; adversarial samples

DOI: 10.19678/j.issn.1000-3428.0059156

0 概述

随着人工智能技术的快速发展,深度学习已广泛应用于图像分类^[1-3]、目标检测^[4-6]和语音识别^[7-9]等领域,但由于其自身存在若干技术性不足,导致深度学习在给人们生活带来极大便利的同时也面临着较多挑战,模型算法的安全隐患更是加剧了深度学习技术被对抗样本欺骗以及隐私泄露等安全风险,因此深度学习的安全问题^[10]引起了研究人员的广泛关注。在早期研究中,针对深度学习算法潜在攻击以及相应防御方法的研究,主要关注模型的攻击成功

率以及是否能够成功规避某种攻击方法。以传统的分类模型为例,其存在判断准确度越高则模型鲁棒性越低这一问题,因此,学者们开始关注模型鲁棒性和准确度的平衡问题。

现有综述性文献多数倾向于阐述传统的对抗攻击与防御方法。近年来,对抗样本的研究变得多样化,早期研究通常将对抗样本视为神经网络的一种威胁,近期学者们聚焦于如何在不同领域利用对抗样本的特性来更好地完成分类和识别等任务。

2013年, SZEGEDY 等人^[11]利用难以察觉的扰动来揭示深度神经网络的脆弱特性。2014年,

基金项目: 中央高校基本科研业务费专项资金(3072020CF0604, 3072020CFQ0602)。

作者简介: 姜妍(1991—),女,博士研究生,主研方向为深度学习、对抗学习;张立国(通信作者),副教授、博士。

收稿日期: 2020-08-03 **修回日期:** 2020-09-29 **E-mail:** zhangliguo@hrbeu.edu.cn

GOODFELLOW 等人^[12]提出对抗样本的概念。此后,越来越多的研究人员专注于该领域的研究。早期的研究工作致力于分析不同深度学习模型(如循环神经网络、卷积神经网络等)中的漏洞以及提高模型对对抗样本的鲁棒性。BARRENO 等人^[13]对深度学习的安全性进行了调研,并针对机器学习系统的攻击进行分类。PAPERNOT 等人^[14]总结了已有关于机器学习系统攻击和相应防御的研究成果,并系统分析机器学习的安全性和隐私性,提出机器学习的威胁模型。近年来,针对对抗样本的研究更加多样化。2019年,XIE 等人^[15]提出利用对抗样本改进图像识别模型精度的方法。2020年,DUAN 等人^[16]利用风格迁移技术^[17]使对抗样本在物理世界变得人眼不可察觉,以达到欺骗算法的目的。由于深度学习模型存在脆弱性,类似的对抗攻击同样会威胁深度学习在医疗安全、自动驾驶等方面的应用。

自从对抗攻击的概念被提出之后,研究人员不断提出新的攻击方法和防御手段。现有的对抗攻击方法研究主要针对对抗样本的生成方法以及如何提高对模型的攻击成功率,对抗防御研究主要关注基于对抗样本的检测与提高模型鲁棒性2个方面。本文介绍对抗样本的概念、产生的原因及对抗样本的可迁移性,分析现阶段经典的对抗样本生成方法以及检测手段,并归纳针对上述检测手段的防御策略,通过梳理分析较为先进的对抗样本应用方法以展望该领域未来的研究方向。

1 对抗攻击

1.1 深度学习

深度学习^[18]是一种深层模型,其利用多层非线性变换进行特征提取,由低层特征抽取出高层更抽象的表示。从广义上而言,深度学习所用到的神经网络主要分为循环神经网络^[19]、深度置信网络^[20]和卷积神经网络^[21]等。与所有连接主义模型固有的脆弱性问题相同,深度学习系统很容易受到对抗样本的攻击。

1.2 对抗攻击的概念

对抗样本指人为构造的样本。通过对正常样本 x 添加难以察觉的扰动 η ,使得分类模型 f 对新生成的样本 x' 产生错误的分类判断。新生成的对抗样本为 $x' = x + \eta$,即:

$$f(x') \neq f(x), \|x' - x\| \leq \varepsilon \quad (1)$$

目前,寻找扰动的主流方法包括快速梯度攻击(FGSM)^[12]、C&W 攻击^[22]、替代黑盒攻击^[23]、DeepFool 攻击^[24]、单像素攻击(One-Pixel Attack, OPA)^[25]、AdvGAN 攻击^[26]、通用对抗扰动^[27]和后向传递可微近似(Backward Pass Differentiable Approximation, BPDA)方法^[28]等。一些研究成功攻击了除卷积神经网络和深度神经网络之外的其他深度学习模型,甚至在现

实世界中产生对抗的实例,如对抗眼镜^[29]、对抗停止标志^[30]等,这些都对物理世界中的深度学习系统造成了干扰。

图1所示为通过FGSM方法生成的对抗样本,加入了扰动的对抗样本使左图的熊猫被错误分类为长臂猿。FGSM方法在各个维度上移动相同大小的一步距离,虽然一步很小,但每个维度上的效果相加,也足以对分类器的判别结果产生显著影响,因此,FGSM攻击方法可应用于任何可以计算 $\nabla_x L(x, y)$ 的深度学习模型。

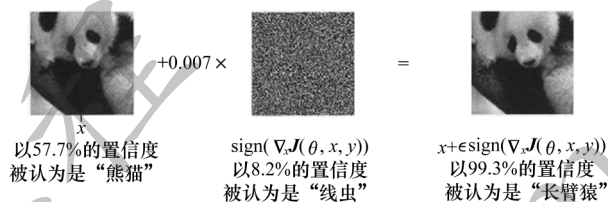


图1 FGSM方法生成的对抗样本

Fig.1 Adversarial samples generated by FGSM method

对抗样本可以轻易欺骗某种深度神经网络模型,且其具有可迁移性^[31],可用于欺骗其他模型。对抗样本的可迁移性分为以下3种类别:

- 1)在同一数据集训练的不同模型之间的可迁移性,如深层神经网络下的VGG16^[32]和ResNet^[2]之间。
- 2)在不同机器学习技术之间的可迁移性,如支持向量机^[33]和深度神经网络之间。
- 3)在执行不同任务的模型之间的可迁移性,如语义分割^[34]、图像分割和目标检测模型之间。

影响样本可迁移性的4个因素具体如下:

- 1)模型类型。PAPERNOT 等人^[35]研究发现,深度神经网络和k近邻算法对跨技术可迁移性更为稳健,但对技术内可迁移性较为脆弱,线性回归^[36]、支持向量机、决策树^[37]和集成方法对技术内可迁移性更为稳健,但对跨技术可迁移性较为脆弱。

- 2)对抗样本的攻击力。KURAKIN 等人^[38]研究发现,能够穿透坚固防御模型的更强的对抗样本不太可能迁移到其他模型,而生成攻击但并未成功攻击防御模型的对抗样本更容易迁移,为渗透特定防御方法而产生的对抗样本可能“过拟合”欺骗特定模式。

- 3)非目标攻击比目标攻击更容易迁移。LIU 等人^[39]通过研究ImageNet数据集的可迁移性,发现可迁移的非目标对抗样本比目标样本更多,且不同模型的决策边界一致。

- 4)数据的统计规律。JO 和BENGIO^[32]认为卷积神经网络倾向于学习数据中的统计规律而非抽象概念。由于对抗样本具有可迁移性,使其在同一数据集上训练的模型之间可迁移,这些模型可能学习相同的统计信息从而落入同样的“陷阱”。

1.3 对抗样本的产生原因

自从对抗样本被发现以来,其产生原因一直是学者们争议的热点。

2014年,SZEGEDY等人^[11]认为对抗样本位于数据流形的低概率区域,由于分类器在训练阶段只学习局部子区域,而对样本超出了学习的子集,导致神经网络分类错误。如图2所示,A类和B类分别表示不同的样本空间,模型训练所得的分类边界(曲线)与真实决策边界(直线)并不重合,在曲线与直线相交的区域出现样本会导致模型判断失误,曲线和直线包围的区域即为对抗区域。

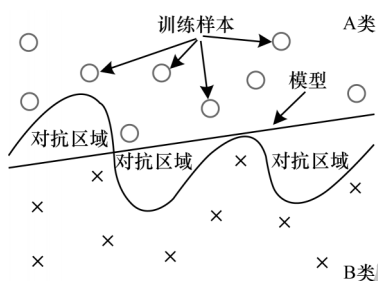


图2 对抗样本区域

Fig.2 The area of adversarial samples

2015年,GOODFELLOW等人^[12]反驳了SZEGEDY等人的观点,认为深度神经网络的脆弱性是由于模型的局部线型特性所导致,特别是模型使用如ReLU^[35]或Maxout^[40]等线性激活函数时,更容易受到对抗样本的攻击。虽然神经网络也使用非线性激活函数,但是为避免出现梯度消失等现象^[40-42],研究人员通常在激活函数的线性区域内训练网络。此外,GOODFELLOW等人认为快速梯度攻击是基于线性假设而设计的,能够有效欺骗深层神经网络,从而验证了神经网络行为类似于线性分类器的论点。

2017年,ARPIT等人^[31]通过分析神经网络对训练数据的记忆能力,发现记忆能力强的模型更容易受到对抗样本的影响。

2018年,GILMER等人^[43-44]认为数据流形的高维几何结构产生了对抗样本,他们在合成数据集的基础上对对抗样本与数据流形高维几何结构之间的关系进行了分析论证。

截至目前,深度学习模型易受对抗样本攻击的原因仍然是一个开放的研究课题,缺乏完备的理论体系,这也制约着深度学习系统的进一步发展。

2 对抗样本的攻击方式及目标

根据攻击者掌握的模型信息可将攻击分为白盒攻击与黑盒攻击2种,通过攻击者选择的攻击目标可将攻击分为目标攻击、无目标攻击和通用攻击3种。

2.1 攻击方式

白盒攻击与黑盒攻击具体如下:

1)白盒攻击:攻击者了解攻击模型的详细信息,

如数据预处理方法、模型结构和模型参数等,某些情况下攻击者还能掌握部分或全部的训练数据信息。在白盒攻击环境中,攻击者对可攻击的模型拥有控制能力,能够观测并设计相应的攻击策略并更改程序运行时的内部数据。

2)黑盒攻击:攻击者不了解攻击模型的关键细节,攻击者仅能够接触输入和输出环节,不能实质性地接触任何内部操作和数据。在黑盒攻击环境中,攻击者可以通过对模型输入样本并根据模型的输出信息来对模型的某些特性进行推理。

2.2 攻击目标

目标攻击、无目标攻击和通用攻击具体如下:

1)目标攻击:攻击者指定攻击范围和攻击效果,使被攻击模型不但样本分类错误并且将样本错误分类成指定的类别。

2)无目标(无差别)攻击:攻击者的攻击目标更为宽泛,攻击目的只是让被攻击模型对样本进行错误分类但并不指定分类成特定类别。

3)通用攻击:攻击者设计一个单一的转换,例如图像扰动,该转换是对所有或者多数输入值造成模型输出错误的攻击。

3 对抗样本的生成方法

现阶段较为经典的攻击方法是FGSM方法及其变体、C&W攻击、替代黑盒攻击、DeepFool攻击、单像素攻击、AdvGAN攻击、通用对抗扰动、后向传递可微近似方法,具体如下:

1)FGSM方法。FGSM方法最早由GOODFELLOW等人^[12]提出,其工作原理是计算输入的损失函数的梯度,并通过将一个选定的小常数乘以梯度的符号向量来产生一个小的扰动,如下:

$$x' = x + \epsilon \cdot \text{sign}(\nabla_x L(x, y)) \quad (2)$$

其中, ϵ 表示调节系数, $\nabla_x L(x, y)$ 是相对于输入 x 损失函数的一阶导数。FGSM是早期经典的攻击方法,此后衍生出许多以FGSM为基础的对攻击方法,如基本迭代方法(Basic Iterative Method, BIM)、动量迭代的FGSM方法和多样性的FGSM方法等。

(1)基本迭代方法。BIM是FGSM的一种拓展,由KURAKIN等人^[38]提出。BIM通过迭代的方式沿着梯度增加的方向进行多步小的扰动,并且在每一小步后重新计算梯度方向,迭代过程如下:

$$x'_{i+1} = \text{Clip}_{\epsilon} \{x'_i + \alpha \cdot \text{sign}(\nabla_x L(x'_i, y))\} \quad (3)$$

for $i = 0$ to $n, x'_0 = x$

其中, $\text{Clip}_{\epsilon}\{\cdot\}$ 约束坐标的每个输入特征,如像素,将其限制在输入 x 的扰动邻域以及可行的输入空间中, n 为迭代总数量, α 为步长。BIM相比FGSM能构造出更加精准的扰动,攻击效果更好,并在诸多对抗样本攻防比赛中得到了广泛应用,但是其不足之处是提高了计算量。

(2) 动量迭代的 FGSM 方法。2018 年, DONG 等人^[45]提出一种优化的基于动量迭代^[46]的 FGSM (Momentum Iterative FGSM, MI-FGSM) 方法。使用动量能够稳定扰动的更新方向, 也有助于逃离局部极大值, 从而提高样本的可迁移性并提升攻击的成功率。将动量融入到基本迭代的方法中从而产生扰动, 首先输入 x'_i 到分类器 f 以得到梯度 $\nabla_x J(x'_i, y)$, 通过式(4)累积梯度方向上的速度矢量从而更新 g_{i+1} , 然后应用式(5)中的符号梯度来更新 x'_{i+1} , 最后产生扰动。

$$g_{i+1} = \mu \cdot g_i + \frac{\nabla_x J(x'_i, y)}{\|\nabla_x J(x'_i, y)\|_1} \quad (4)$$

$$x'_{i+1} = x'_i + \alpha \cdot \text{sign}(g_{i+1}) \quad (5)$$

上述过程能够证明 BIM 生成的对抗样本比 FGSM 生成的对抗样本更不可迁移, 更强的样本通常更不可迁移, 与 FGSM 和 BIM 攻击相比, MI-FGSM 提高了对抗样本的可迁移性。

(3) 多样性的 FGSM 方法。XIE 等人^[47]针对现有多数对抗样本攻击在黑盒攻击下只能实现较低成功率的问题, 在 BIM 以及 MI-FGSM 等算法的基础上提出考虑多样性^[48]的攻击方法 DI-FGSM。在每次迭代时, 图像会沿着损失梯度 $\nabla_x L(X, y^{\text{true}}; \theta)$ 的方向产生扰动, 导致很容易陷入局部极大值并过度拟合特定的参数 θ 。为解决该问题, DI-FGSM 方法将图像变换应用于输入的图像, 通过每次迭代来缓解过拟合现象。DI-FGSM 方法过程为:

$$X_{n+1}^{\text{adv}} = \text{Clip}_X \left\{ X_n^{\text{adv}} + \alpha \cdot \text{sign} \left(\nabla_x L \left(T(X_n^{\text{adv}}; p), y^{\text{true}}; \theta \right) \right) \right\} \quad (6)$$

其中, $T(\cdot)$ 表示图像变换。DI-FGSM 方法可以和其他攻击方法相结合, 例如 PGD 和 C&W。实验结果表明, 加入多样性的 D-C&W 的攻击成功率明显高于原始的 C&W 攻击。使用 DI-FGSM 方法能够同时实现白盒攻击和黑盒攻击的高成功率, 并在此基础上提高对抗样本的可迁移性。DI-FGSM 方法的更新过程与基本迭代方法相似。

图 3 所示为 FGSM 方法及其变体的转换关系, 其中, N 表示可迁移性概率, μ 表示衰减因子, p 表示总的迭代数量。

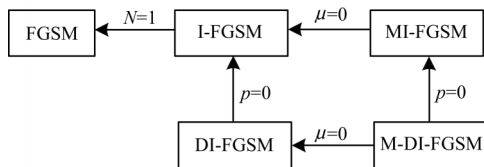


图 3 FGSM 方法及其变体之间的转换关系

Fig.3 Transformation relationship between FGSM method and its variants

2) C&W 攻击。CARLINI 和 WAGNER^[22]提出 3 种对抗攻击方法 (L_0 攻击、 L_2 攻击、 L_∞ 攻击), 用于寻找能够最小化各种相似性度量的扰动。通过限制 L_0 、 L_2 、 L_∞ 范数, 使得扰动近似于无法被察觉。实验结果表明, 这 3 种攻击以 100% 的成功率绕过了防御蒸馏, 同时在 L_0 、 L_2 和 L_∞ 范数下保持对抗样本与原始输入相似, 具有很强的可迁移性。在 MNIST、CIFAR10 和 ImageNet 上进行评估时, C&W 方法优于同一时期较先进的攻击方法, 如 BIM 攻击, 即使在现阶段, C&W 攻击依旧被广泛应用。

3) 替代黑盒攻击。PAPERNOT 等人^[23]提出一种早期的黑盒攻击方法, 即替代黑盒攻击 (Substitute Blackbox Attack, SBA)。SBA 攻击训练一个模仿黑盒模型的替代模型, 在该替代模型上使用白盒攻击。具体而言, 攻击者首先从目标模型收集一个合成数据集, 训练一个替代模型来模拟目标模型的预测。在训练替代模型后, 可以使用任何白盒攻击生成对抗样本, 原因是替代模型的细节已知。SBA 可用于攻击其他机器学习模型, 如逻辑回归和支持向量机等模型。通过在 MNIST 数据集上定位来自亚马逊、谷歌和 MetaMind 的真实世界图像识别系统来评估 SBA, 结果表明, SBA 能够以高精度 (>80%) 欺骗所有目标, 且其可以规避依赖梯度掩蔽的防御方法。

4) DeepFool 攻击。MOOSAVI 等人^[24]提出一种不对原始样本扰动范围进行限制的新方法 DeepFool, 其作为一种早期的对抗样本生成方法, 可以生成比快速梯度攻击更小的扰动。DeepFool 首先初始化原始图像并假定分类器的决策边界限制图像分类的结果, 然后通过每一次迭代, 沿着决策边界方向进行扰动, 逐步地将分类结果向决策边界另一侧移动, 使得分类器分类错误。和 FGSM 相比, DeepFool 计算速度更快, 可以生成更精确的扰动。

5) 单像素攻击。OPA^[25]是一种“半黑盒”攻击方法, 其使用差分进化算法策略来寻找对抗性扰动。OPA 的目的是通过只修改给定图像 x 的一个特征来欺骗目标模型。通过对原有数据修改 3 个或 5 个像素来产生对抗样本, 在多种模型下都可达到误分类的效果, 好的情况下修改 1 个像素即可产生误分类结果。

6) AdvGAN 攻击。XIAO 等人^[26]提出一种基于生成对抗网络 (GAN) 框架的对抗生成方法 AdvGAN, AdvGAN 主要由 3 个部分组成, 分别为生成器 G 、判别器 D 和目标神经网络 C 。该方法将原始样本 x 通过 GAN 生成器 G 映射成对抗扰动 $G(x)$, 然后将扰动输入原始样本 x 中, 一旦经过训练, 网络可以迅速产生新的对抗样本 $x + G(x)$, 判

别器 D 判别输入的样本是否为对抗样本,同时用生成的对抗样本欺骗目标神经网络 C 。AdvGAN 在对抗性训练中的表现优于快速梯度攻击和 C&W,其产生的对抗样本在视觉上与真实样本难以区分。但是,AdvGAN 方法存在一个潜在问题,尽管其被证明能够绕过阻碍快速梯度攻击和 C&W 方法的防御方法,但与其他基准对抗性攻击和防御方法相比,AdvGAN 在对抗性训练设置中较为单一,可能不会被广泛应用。

7) 通用对抗扰动。通用对抗扰动 (Universal Adversarial Perturbation, UAP)^[27] 是一种适用于不同网络模型的通用扰动计算方法,其工作原理是累积单个输入的扰动,以这种方式产生的扰动 v 可以添加到每个数据样本中,以便将它们推向目标的决策边界附近,重复此过程,直至样本被错误分类。实验结果表明,各种模型均存在通用扰动,通用扰动在这些模型之间表现出较高的可转换性。UAP 攻击的一个潜在缺陷是其不能保证每一个更新的通用扰动 v 对更新前出现的数据点仍然具有对抗性。

8) 后向传递可微近似方法。ATHALYE 等人^[28] 针对现有多数防御方法依赖于模糊模型梯度的问

题,提出利用防御模型的可微近似来获得有意义的对抗梯度估计从而修改对抗攻击的方法,该方法称为 BPDA 方法。BPDA 方法结合期望大于转化攻击 (Expectation over Transformation, EoT) 方法^[40],可以攻破混淆梯度防御。BPDA 方法给定输入样本数据 x ,假设神经网络可写为函数 $f^{1,2,\dots,j}(x)$,在计算对抗样本梯度时,攻击者可以用另一个函数 $g(x)$ 来进行计算,在前向传播验证对抗样本是否成功时仍然使用 $f^{1,2,\dots,j}(x)$ 进行判断。BPDA 方法成功攻破了 7 种基于混淆梯度的防御方法。

本文总结以上 8 种比较经典的对抗攻击方法,包括攻击类型、目标、学习方式、攻击强度及算法优势和劣势。学习方式可分为单次迭代和多次迭代,单次迭代方法可以快速生成对抗样本,并用于对抗训练从而提高模型的鲁棒性;多次迭代方法则需要更多的计算时间来生成对抗样本,但其攻击效果强且难以防范。以上经典对抗攻击方法的对比分析结果如表 1 所示,其中,单步表示单次迭代,迭代表示多次迭代,W 表示白盒攻击,B 表示黑盒攻击,T 表示有目标攻击,NT 表示无目标攻击,* 的数量代表攻击强度。

表 1 攻击方法性能对比结果
Table 1 Performance comparison results of attack methods

攻击方法	攻击类型	目标	学习方式	攻击强度	优势	劣势
FGSM	W	T, NT	单步	***	生成效率高,具有可转移性	单步攻击,白盒攻击成功率低
BIM	W	T, NT	迭代	****	迭代攻击,白盒攻击成功率更高	可转移性较差,容易过拟合
MI-FGSM	W	T, NT	迭代	****	提高了迭代攻击的可转移性,能够稳定更新	容易受到黑盒攻击的攻击
D-MI-FGSM	W	T, NT	迭代	*****	可以与其他攻击相结合,提高白盒与黑盒攻击成功率,提升可转移性	计算效率较低
C&W	W	T, NT	迭代	*****	生成的扰动较小,可以破解很多的防御方法,具有可移植性	生成时间较长
SBA	B	T, NT	迭代	*****	可用于攻击其他机器学习模型,有效规避依赖梯度掩蔽的防御方法	在实际中应用较少,攻击者几乎不可能得到模型的详细信息
DeepFool	W	NT	迭代	****	计算的扰动相比 FGSM 更小,提高了计算速度	生成时间是 FGSM 的 5 倍
OPA	B	T, NT	迭代	**	在多种模型下具有高误分类率	寻找可攻击像素的时间较长
AdvGAN	W	T, NT	迭代	****	在对抗性训练中有更好的表现,产生的对抗样本在视觉上与真实图像难以区分	在对抗性训练设置中较为单一,可能不会推广到更广泛的设置
UAP	W	NT	迭代	*****	证明通用扰动的存在,具有很高的可移植性	不能保证每一个更新的通用对抗扰动对更新前出现的数据点仍然具有对抗性
BPDA	W	T, NT	迭代	*****	可有效规避依赖混淆梯度的防御方法	只针对混淆梯度的防御

4 对抗防御

对抗样本的存在促使学者开始思考如何成功防御对抗攻击,从而避免模型识别错误。对抗防御主要分为对抗攻击检测和提高模型鲁棒性 2 种方式,检测方法独立于防御方法,可以单独用来检测样本的对抗性,也可以与防御方法结合使用。

4.1 对抗攻击的检测

对抗样本产生原因的复杂性使得对于对抗样本的通用化检测变得十分困难。对抗攻击检测通过检测样本的对抗性来判断其是否为对抗样本。对抗攻击检测主要包括如下方法:

1) H&G 检测方法。HENDRYCKS 等人^[49]提出

3种对抗性检测方法,统称为H&G检测方法。从广义上而言,H&G检测方法利用了正常样本和扰动问题之间的经验差异来区分正常样本和对抗样本。3种对抗性检测方法具体如下:

(1)通过对对抗样本的主成分分析白化输入系数的方差从而检测样本的对抗性。当攻击者不知道防御措施是否到位时,该方法可用于检测FGSM和BIM攻击。

(2)正常输入和对抗输入之间的Softmax分布不同,H&G检测方法利用该分布差异执行对抗检测,测量均匀分布和Softmax分布之间的Kullback-Leibler散度,然后对其进行基于阈值的检测。研究发现,正常样本的Softmax分布通常比对抗样本的均匀分布离散,原因是模型倾向于以高置信度预测输入。

(3)在以逻辑为输入的分类器模型中加入一个辅助译码器重构图像从而检测对抗样本,解码器和分类器只在正常样本上联合训练,检测通过创建一个检测器网络来完成,该网络以重建逻辑和置信度得分为输入,输出一个输入具有对抗性的概率,其中,探测器网络在正常样本和对抗样本上都受过训练。该方法能够检测FGSM和BIM产生的对抗样本。

2) 对抗性检测网络。METZEN等人^[50]提出对抗性检测网络(Adversary Detector Network, ADN),其为一种用二元检测器网络扩充预训练神经网络的检测方法,检测器网络被训练以区分正常样本和对抗样本。ADN方法能有效检测FGSM、DeepFool和BIM攻击,但CARLINI等人^[51]发现该方法对C&W等强攻击具有较高的假阳性,并可以通过SBA攻击来规避。GONG等人^[52]对ADN方法进行改进,改进方法中的二进制分类器是一个与主分类器完全分离的网络,其不是针对检测器生成对抗样本,而是为预训练分类器生成对抗样本,并将这些对抗样本添加到原始训练数据中以训练二进制分类器。但CARLINI等人^[51]指出,该改进方法在CIFAR10模型

上测试时具有较高的假阳性,并且容易受到C&W攻击。

3) 核密度法和贝叶斯不确定性估计法。FEINMAN等人^[53]假设对抗样本不在非对抗性数据流形中,在此情况下提出核密度法和贝叶斯不确定性估计(Bayesian Uncertainty Estimates, BUE)2种对抗性检测方法。使用核密度估计(Kernel Density Estimates, KDE)的目的是确定一个数据点是否远离类流形,而BUE可以用来检测靠近KDE无效的低置信区域的数据点。BUE是较难欺骗的检测方法,作为现有网络的附加组件,其实现也相对简单。

4) 特征压缩。XU等人^[54]认为输入特征的维度通常过大,导致出现一个大的攻击面。根据该原理,他们提出基于特征压缩的检测方法(FS),用以比较压缩和非压缩输入之间的预测结果。特征压缩的目的是从输入中去除不必要的特征,以区分正常样本与对抗样本。如果模型对压缩和非压缩输入的预测结果之间的 L_1 范数差大于某个阈值 T ,则该输入被标记为对抗性输入。FS方法独立于防御模型,因此,其可以与其他防御技术结合使用。特征压缩被证明能够在攻击者不了解所使用的防御策略的情况下检测由FGSM、BIM、DeepFool、JSMA^[55]和C&W攻击生成的对抗样本。

5) 逆交叉熵检测。2017年,PANG等人^[56]提出利用新的目标函数进行反向检测的逆交叉熵(Reverse Cross-Entropy, RCE)方法,该方法训练一个神经网络以区分对抗样本和正常样本。在FGSM、BIM/ILLCM、C&W、MNIST和CIFAR10数据集上进行评估,结果表明RCE具有有效性。与使用标准交叉熵作为目标函数的方法相比,RCE不仅允许用户进行对抗性检测,而且在总体上提高了模型的鲁棒性。

本节总结现阶段主要的对抗攻击检测方法的性能,结果如表2所示。

表2 对抗攻击检测方法性能对比结果

Table 2 Performance comparison results of adversarial attacks detection methods

检测方法	防御类型	优势	劣势
H&G 检测方法	对抗检测	包括不同检测方法的组合	已经被成功规避
对抗性检测网络	对抗检测	该方法对FGSM、DeepFool和BIM有效	对C&W等强攻击具有较高的假阳性,可以通过SBA规避
KDE&BUE	对抗检测	能够处理JSMA、BIM-A和C&W攻击	在CIFAR10上仅使用KDE方法时不起作用
特征压缩	对抗检测	可以检测由FGSM、BIM、DeepFool、JSMA和C&W生成的攻击	颜色深度减少压缩对 L_0 攻击的效果较差
逆交叉熵检测	对抗检测	针对FGSM、BIM/ILLCM、JSMA和C&W有效	可以通过强攻击规避

4.2 对抗攻击的防御

为了使模型对抗抗性攻击更加具有鲁棒性,研究人员提出不同的防御方法,这些方法建立在对抗性和正常输入下同样具有良好表现的模型上,使模

型对输入的不相关变化不太敏感,从而有效地正则化模型以减少攻击面,并限制对非流形扰动的响应。目前,针对对抗攻击的防御方式主要分为以下4类:

1) 数据扩充,该方法通过在训练集中加入对抗

样本进行再训练,从而提高模型的鲁棒性。

2) 预处理方法,该方法通过对原有数据进行处理从而降低对抗样本的有效性。

3) 正则化方法,该方法使用防御蒸馏方法降低网络梯度的大小,提高发现小幅度扰动对抗样本的能力。

4) 数据随机化处理,该方法通过对输入进行随机调整来消除扰动。

4.2.1 数据扩充

具有代表性的数据扩充方法如下:

1) 对抗训练。为提高神经网络模型在对抗攻击环境下的鲁棒性,很多学者对对抗样本进行代入训练^[12]。在每次迭代训练中,通过在训练集中注入对抗样本来对模型进行再训练。由于单步对抗训练的鲁棒性主要由梯度掩蔽引起,因此该模型可以被其他类型的攻击所规避。此外,单步对抗训练可能会出现标签泄漏问题,容易导致模型过度拟合。

2) 映射梯度下降对抗训练。2018年,MADRY等人^[57]改进了对抗训练,提出一种映射梯度下降对抗训练(Projected Gradient Descent, PGD)。标准对抗训练方法是在正常样本和对抗样本上训练模型,而在PDG框架中,模型只在对抗样本上训练。PGD方法在白盒和黑盒2种设置下对各种类型的攻击都保持一致的鲁棒性,但其模型可能无法达到最优的性能。由于PGD的计算代价随迭代次数的增加而提高,因此该方法的计算代价通常高于标准对抗训练。

3) 综合性对抗训练。2018年,针对传统对抗训练容易出现过拟合的问题,TRAMER等人^[58]提出综合性对抗训练,其为对抗性训练的另一种变体。在综合性对抗训练中,模型根据生成的对抗样本进行再训练,以攻击其他各种预先训练的模型。这种目标模型和对抗训练实例的分离能够有效克服传统对抗训练的过拟合问题。

4) 逻辑配对防御机制。KANNAN等人^[59]提出逻辑配对防御(ALP)机制,其鼓励输入对(即对抗性和非对抗性输入对)的逻辑相似,并设计对抗性逻辑配对和正常逻辑配对(CLP)2种不同的逻辑配对策略。ALP在原始输入及其对抗输入之间强制执行逻辑不变性,而CLP在任何一对输入之间强制执行逻辑不变性。KANNAN等人发现PGD攻击的对抗性训练与ALP相结合,在ImageNet模型上对白盒攻击与黑盒攻击都具有较优的鲁棒性。

4.2.2 预处理方法

通过对数据进行预处理能够降低对抗样本的有效性,现有的预处理方法主要包括:

1) 通过学习非对抗性数据集的分布,将对抗性

输入投射到学习的非对抗性流形中。

2) 通过对对抗样本的过滤或去噪将其转化为纯净样本。

3) 对输入进行变换处理,使攻击者难以计算模型的梯度,从而达到防御对抗攻击的目的。

4) 对输入数据进行量化和离散化处理,有效消除对抗性扰动的影响。

具有代表性的预处理方法具体如下:

1) 去噪特征映射方法。XIE等人^[60]研究发现,与原始输入相比,对抗性扰动导致模型生成的特征图所发生的变化较大,基于此,他们提出一种去噪特征映射(FDB)方法。实验结果表明,去噪块不会大幅降低非对抗性输入的性能,当与PGD对抗训练相结合时,无论是在黑盒还是白盒模式下,FDB防御都能达到当时较优的对抗鲁棒性。

2) 综合分析方法。一些基于生成对抗网络的防御机制相继被提出,如基于生成模型的GAN防御方法,该方法学习非对抗性数据集的分布,以将对抗性输入投射到学习的非对抗性流形中。SCHOTT等人^[61]提出了综合分析(ABS)防御方法,该方法并非学习整个数据集的输入分布,而是学习每个类的输入分布。在MNIST数据集上,ABS在对抗 L_0 和 L_2 对抗样本时表现出比PGD对抗性训练更优、更健壮的效果,但针对 L_∞ 对抗样本时ABS的鲁棒性较低。

3) ME-Net方法。YANG等人^[62]提出基于预处理的防御方法ME-Net,其对输入进行预处理,以破坏对抗性噪声的结构。ME-Net方法的工作原理是根据一定的概率 r 随机丢弃输入图像中的像素点,假设该概率 r 可以破坏对抗干扰,使用矩阵估计算法重建图像。ME-Net方法是从噪声观测中恢复矩阵数据的方法,在CIFAR-10、MNIST、SVHN和小型ImageNet数据集上的黑盒和白盒模式中,ME-Net测试各种 L_∞ 攻击时都表现出了很强的健壮性。

4) 总方差最小化和图像拼接方法。在分类之前,可以对输入图像应用各种图像变换方法,在这些图像变换方法中,GUO等人^[63]研究发现总方差最小化和图像拼接最有效,特别是当模型在转换后的图像上训练时,总方差最小化和图像拼接都引入了随机性,并且都是不可微的操作,使得攻击者很难计算模型的梯度。该防御是模型不可知的,意味着模型无需再训练或微调,且这种防御方法可以与其他防御方法结合使用。

5) 温度计编码防御方法。BUCKMAN等人^[64]提出神经网络的线性使其易受攻击的假设,并根据该假设提出温度计编码防御(TE)方法。TE防御对输入数据进行量化和离散化处理,有效消除了通常

由对抗性攻击引起的对抗扰动影响。TE防御和对抗训练相结合后具有很高的对抗稳健性,可以超过PGD对抗训练,但是,TE防御依赖梯度掩蔽,可以使用BPDA攻击绕过。

4.2.3 正则化方法

正则化方法包括深度压缩网络^[65]和防御蒸馏^[66]等。防御蒸馏是早期较为经典的一种方法,“蒸馏”一词由HINTON等人^[67]引入,是一种将深层神经网络集合中的知识压缩为单一神经网络的方法。防御蒸馏由原始网络和蒸馏网络2个网络组成,原始网络也叫教师网络,一般为参数多且结构复杂的网络,蒸馏网络也叫学生网络,一般为参数少且结构简单的网络。蒸馏方法可以将教师网络的知识有效地迁移到学生网络,从而起到压缩网络的作用。防御蒸馏对由早期攻击方法生成的对抗样本具有健壮性,

但是,这种防御易受到C&W与SBA变体的攻击。

4.2.4 数据随机化处理

数据随机化处理包括随机调整大小、填充、随机激活剪枝^[68]等。XIE等人^[69]提出基于随机调整大小和填充(RRP)的防御机制,其通过输入变换消除扰动,并在推理过程中引入随机性,使得相对于输入的损失梯度更难计算。该机制不需要对防御模型进行微调就能保证精确性,并且可以与如对抗性训练等其他防御方法相结合,对FGSM、BIM、DeepFool和C&W等白盒攻击都表现出良好的性能。

现阶段主要的4类防御方法总结对比如表3所示。在保证计算成本的情况下,目前较常用的防御方法是数据扩充方法。随着攻击手段的提高,未来可能会以多种方法相结合的方式来提高模型的鲁棒性,并且使得模型的鲁棒性与准确率达到平衡。

表3 各种防御方法总结对比结果

Table 3 Summary and comparison results of various defense methods

对抗防御方法	防御类型	防御特点
对抗训练	提升鲁棒性	对抗性示例训练
PGD对抗训练	提升鲁棒性	只在PGD对手上训练
综合性对抗训练	提升鲁棒性	与标准对抗训练相比,更能抵抗黑盒攻击
逻辑配对防御	提升鲁棒性	被多次迭代的PGD绕过
去噪特征映射	提升鲁棒性	基于可微消噪运算的隐表示消噪
综合分析	提升鲁棒性	基于VAE的各类输入模型分布
ME-Net	提升鲁棒性	基于矩阵估计算法的防御
总方差最小化和图像拼接	提升鲁棒性	BPDA与EOT联合应用可绕过该防御
温度计编码防御	提升鲁棒性	可以被BPDA绕过
防御蒸馏	提升鲁棒性	可以被C&W、SBA和JSMA变体绕过
随机调整大小	提升鲁棒性	在白盒攻击下表现良好

5 对抗样本应用实例

随着对抗样本研究的多样化发展,学者们开始从不同角度探索对抗样本的特性,发现除对抗样本对神经网络模型构成威胁之外,还可以利用对抗样本的特性提高模型性能,具体如下:

1)利用对抗样本提高图像识别准确率。XIE等人^[15]研究发现已有方法可以共同训练原始图像和对抗样本,但此类方法往往会导致最终图像识别准确率下降,即使从不同的分布中提取图像,也会导致同样的结果。由此他们假设原始图像与对抗样本之间分布不匹配是导致此类方法性能下降的关键因素,基于该假设,XIE等人提出一种新的训练方法——AdvProp方法,其通过一种简单且高效的两批次标准方法来解决分布不匹配的问题。使用2个批处理规范统计信息,一个用于原始样本,另一个用于对抗样本,2个批处理规范在归一化层正确分散了2个分布,以进行准确的统计估计。实验结果表明,AdvProp

大幅提高了卷积网络的模型识别准确率。

2)利用对抗性特征解决超分辨率问题。感知损失函数在解决图像超分辨率问题时虽然取得了较好效果,但也会在超分辨率输出中产生不期望的图案伪像。TEJ等人^[70]针对图像超分辨率不确定的问题,提出利用内容损失函数增强现有感知损失函数的方法,该函数使用鉴别器网络的潜在特征来过滤多个对抗相似性级别上的不需要的伪像。实验结果表明,上述损失函数具有互补的优势,相结合后可以有效提高超分辨率重建的保真度。

3)利用对抗扰动检测木马。ZHANG等人^[71]针对深度神经网络木马中毒的问题,提出一种验证预训练模型是否被特洛伊木马攻击的方法。该方法利用从网络梯度中学到的对抗性扰动的形式捕获神经网络指纹,在系统后门插入神经网络会更改其决策边界,这些系统后门可以在其对抗性干扰中有效地进行编码,从其全局(L_u 和 L_v 有界)扰动以及每个扰动内的高能量局部区域训练2个流网络来检测木

马。前者对网络的决策边界进行编码,后者对未知的触发形状进行编码,并设计一种不会改变触发类型、触发大小、训练数据和网络架构的异常检测方法来识别木马网络中的目标类。实验结果表明,该方法能够取得92%以上的检测精度。

6 未来研究展望

深度学习技术的迅速发展,使得其在图像分类、目标检测等领域取得重大进展的同时也暴露了数据、模型等安全隐患。针对在深度学习系统中出现的安全问题,研究人员开展了一系列攻击防御方法研究,但是,对于深度学习系统的安全性能而言,未来还有很多问题等待解决。本文总结以下3个未来的研究方向:

1)应用对抗样本技术作为数据增强的手段。对抗样本可用于提升模型的泛化性,起到数据增强的作用,目前通常在图像分类中提高分类准确率,也可以在恶意软件检测中提升对恶意软件的检测率。相较于普通的数据增强,对抗样本的优势是可以根据模型自身去调整正则化的强度,从而更好地优化模型。

2)改进对抗训练。对抗训练是目前较优的提高模型鲁棒性的方法,但其存在速度慢、在小数据集上训练会过拟合等问题。后续将在兼顾计算效率与效果的情况下,结合不同的损失函数或者改进应用的网络结构。

3)研究除范数约束和对抗训练之外的攻击防御方法。现有的攻击防御大多是基于范数约束和对抗训练,而这些方法不是唯一有效的攻击防御手段,例如,通过风格迁移技术可以生成对抗样本、利用3D打印技术能够实现攻击等。因此,在物理场景中应用并开展对抗样本防御的研究,从不同角度探索其他的攻击防御方式也具有实际意义。

7 结束语

针对深度学习技术的安全问题,本文介绍对抗样本和对抗攻击的概念,对比分析目前比较经典的对抗攻击方法,在此基础上,总结现阶段相应的防御方法和对抗攻击检测方法的性能。深度学习模型的安全领域未来仍有许多问题需要解决,对抗样本防御技术将与统计学习等方法相结合,为同时提升模型的泛化性和鲁棒性提供新思路,加快推进深度学习模型的安全建设,保护人们的信息隐私安全。

参考文献

- [1] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.
- [2] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2016: 770-778.
- [3] HUANG G, LIU Z, VAN DER MAATEN L, et al.ensely connected convolutional networks[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2017: 4700-4708.
- [4] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2014: 580-587.
- [5] GIRSHICK R. Fast R-CNN[C]//Proceedings of IEEE International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2015: 1440-1448.
- [6] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]//Proceedings of Advances in Neural Information Processing Systems. Washington D. C., USA: IEEE Press, 2015: 91-99.
- [7] GRAVES A, MOHAMED A, HINTON G. Speech recognition with deep recurrent neural networks[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. Washington D. C., USA: IEEE Press, 2013: 6645-6649.
- [8] GRAVES A, JAITLY N. Towards end-to-end speech recognition with recurrent neural networks[C]//Proceedings of International Conference on Machine Learning. Washington D. C., USA: IEEE Press, 2014: 1764-1772.
- [9] ZHANG Y, PEZESHKI M, BRAKEL P, et al. Towards end-to-end speech recognition with deep convolutional neural networks[EB/OL]. [2020-07-20]. <https://arxiv.org/abs/1701.02720>.
- [10] CHEN Yufei, SHEN Chao, WANG Qian, et al. Artificial intelligence system security and privacy risks[J]. Computer Research and Development, 2019, 56(10): 2135-2150. (in Chinese)
陈宇飞, 沈超, 王骞, 等. 人工智能系统安全与隐私风险[J]. 计算机研究与发展, 2019, 56(10): 2135-2150.
- [11] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[EB/OL]. [2020-07-20]. <https://arxiv.org/abs/1312.6199>.
- [12] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[EB/OL]. [2020-07-20]. <http://de.arxiv.org/pdf/1412.6572>.
- [13] BARRENO M, NELSON B, JOSEPH A, et al. The security of machine learning[J]. Machine Learning, 2010, 81(2): 121-148.
- [14] PAPERNOT N, MCDANIEL P, SINHA A, et al. Towards the science of security and privacy in machine learning[EB/OL]. [2020-07-20]. <https://arxiv.org/abs/1611.03814>.
- [15] XIE C H, TAN M, GONG B, et al. Adversarial examples improve image recognition[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2020: 819-828.
- [16] DUAN R J, MA X J, WANG Y S, et al. Adversarial camouflage: hiding physical-world attacks with natural styles[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2020: 1000-1008.

- [17] WANG Lu, ZENG Guohui, HUANG Bo. Implementation of style transfer algorithm based on deep learning [J]. Intelligent Computer and Application, 2020, 10(2): 57-60, 65. (in Chinese)
王鹿, 曾国辉, 黄勃. 基于深度学习的风格迁移算法的研究与实现[J]. 智能计算机与应用, 2020, 10(2): 57-60, 65.
- [18] GOODFELLOW I, BENGIO Y, COURVILLE A. Deep learning[M]. Cambridge, USA: MIT Press, 2016.
- [19] SU Jiongming, LIU Hongfu, XIANG Fengtao, et al. Survey of interpretation methods for deep neural networks[J]. Computer Engineering, 2020, 46(9): 1-15. (in Chinese)
苏炯铭, 刘鸿福, 项凤涛, 等. 深度神经网络解释方法综述[J]. 计算机工程, 2020, 46(9): 1-15.
- [20] ZHAO Guosheng, CHAO Mianxing, XIE Baowen, et al. Application of deep belief network in cloud security situation prediction [J]. Journal of Chinese Computer Systems, 2020, 41(6): 1195-1202. (in Chinese)
赵国生, 晁绵星, 谢宝文, 等. 深度信念网络在云安全态势预测中的应用[J]. 小型微型计算机系统, 2020, 41(6): 1195-1202.
- [21] HONG Qifeng, SHI Weibin, WU Di, et al. Overview of the development of deep convolutional neural network models [J]. Software Guide, 2020, 19(4): 84-88. (in Chinese)
洪奇峰, 施伟斌, 吴迪, 等. 深度卷积神经网络模型发展综述[J]. 软件导刊, 2020, 19(4): 84-88.
- [22] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//Proceedings of 2017 IEEE Symposium on Security and Privacy. Washington D. C., USA: IEEE Press, 2017: 39-57.
- [23] PAPERNOT N, MCDANIEL P, GOODFELLOW I, et al. Practical black-box attacks against deep learning systems using adversarial examples[EB/OL]. [2020-07-20]. <https://arxiv.org/abs/1602.02697>.
- [24] MOOSAVI-DEZFOOLI S M, FAWZI A, FROSSARD P. DeepFool: a simple and accurate method to fool deep neural networks[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2016: 2574-2582.
- [25] 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.
- [26] XIAO C W, LI B, ZHU J Y, et al. Generating adversarial examples with adversarial networks [EB/OL]. [2020-07-20]. <https://arxiv.org/abs/1801.02610>.
- [27] MOOSAVI-DEZFOOLI S M, FAWZI A, FAWZI O, et al. Universal adversarial perturbations [C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2017: 1765-1773.
- [28] ATHALYE A, CARLINI N, WAGNER D. Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples [EB/OL]. [2020-07-20]. <https://arxiv.org/abs/1802.00420>.
- [29] SHARIF M, BHAGAVATULA S, BAUER L, et al. Accessorize to a crime [C]//Proceedings of 2016 ACM SIGSAC Conference on Computer and Communications Security. New York, USA: ACM Press, 2016: 1528-1540.
- [30] EVTIMOV I, EYKHOLT K, FERNANDES E, et al. Robust physical-world attacks on machine learning models [EB/OL]. [2020-07-20]. <https://arxiv.org/abs/1707.08945>.
- [31] ARPIT D, JASTRZEBSKI S, BALLAS N, et al. A closer look at memorization in deep networks [EB/OL]. [2020-07-20]. <https://arxiv.org/abs/1706.05394>.
- [32] JO J, BENGIO Y. Measuring the tendency of CNNs to learn surface statistical regularities [EB/OL]. [2020-07-20]. <https://arxiv.org/abs/1711.11561>.
- [33] LIU Fangyuan, WANG Shuihua, ZHANG Yudong. Survey of support vector machine models and applications [J]. Computer System Application, 2018, 27(4): 1-9. (in Chinese)
刘方园, 王水花, 张煜东. 支持向量机模型与应用综述[J]. 计算机系统应用, 2018, 27(4): 1-9.
- [34] JING Zhuangwei, GUAN Haiyan, PENG Daifeng, et al. Survey of research in image semantic segmentation based on deep neural network [J]. Computer Engineering, 2020, 46(10): 1-17. (in Chinese)
景庄伟, 管海燕, 彭代峰, 等. 基于深度神经网络的图像语义分割研究综述[J]. 计算机工程, 2020, 46(10): 1-17.
- [35] PAPERNOT N, MCDANIEL P, GOODFELLOW I. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples [EB/OL]. [2020-07-20]. <https://arxiv.org/abs/1605.07277>.
- [36] QIN Yan. Comparison of neural network model and multiple linear regression in predicting CT value of kidney stones [J]. Imaging Research and Medical Applications, 2020, 4(6): 26-28. (in Chinese)
覃延. 神经网络模型和多元线性回归预测肾结石 CT 值的比较[J]. 影像研究与医学应用, 2020, 4(6): 26-28.
- [37] YIN Ru. Research on model decision tree method [D]. Taiyuan: Shanxi University, 2019. (in Chinese)
尹儒. 模型决策树方法研究 [D]. 太原: 山西大学, 2019.
- [38] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial machine learning at scale [EB/OL]. [2020-07-20]. <https://arxiv.org/pdf/1611.01236.pdf>.
- [39] LIU Yanpei, CHEN Xinyun, LIU Chang, et al. Delving into transferable adversarial examples and black-box attacks [EB/OL]. [2020-07-20]. <https://arxiv.org/pdf/1611.02770.pdf>.
- [40] ZHANG X H, TRMAL J, POVEY D, et al. Improving deep neural network acoustic models using generalized maxout networks [C]//Proceedings of 2014 IEEE International Conference on Acoustics, Speech and Signal Processing. Washington D. C., USA: IEEE Press, 2014: 215-219.
- [41] HOCHREITER S. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies [EB/OL]. [2020-07-20]. <http://www.bioinf.at/publications/older/ch7.pdf>.
- [42] PASCANU R, MIKOLOV T, BENGIO Y. On the difficulty of training recurrent neural networks [C]//

- Proceedings of International Conference on Machine Learning. Washington D. C., USA; IEEE Press, 2013; 1310-1318.
- [43] GILMER J, METZ L K, FAGHRI F, et al. Adversarial spheres [EB/OL]. [2020-07-20]. <https://arxiv.org/abs/1801.02774>.
- [44] GILMER J, METZ L, FAGHRI F, et al. The relationship between high-dimensional geometry and adversarial examples [EB/OL]. [2020-07-20]. <https://arxiv.org/pdf/1801.02774v3.pdf>.
- [45] DONG Yinpeng, LIAO Fangzhou, PANG Tainyu, et al. Boosting adversarial attacks with momentum [C]//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA; IEEE Press, 2018; 9185-9193.
- [46] POLYAK B T. Some methods of speeding up the convergence of iteration methods [J]. USSR Computational Mathematics and Mathematical Physics, 1964, 4(5): 1-17.
- [47] XIE Cihang, ZHANG Zhishuai, ZHOU Yuyin, et al. Improving transferability of adversarial examples with input diversity [C]//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA; IEEE Press, 2019; 2730-2739.
- [48] XIE Cihang, WANG Jianyu, ZHANG Zhishuai, et al. Mitigating adversarial effects through randomization [EB/OL]. [2020-07-20]. <https://arxiv.org/abs/1711.01991>.
- [49] HENDRYCKS D, GIMPEL K. Visible progress on adversarial images and a new saliency map [EB/OL]. [2020-07-20]. <https://arxiv.org/abs/1608.00530>.
- [50] METZEN J H, GENEWEIN T, FISCHER V, et al. On detecting adversarial perturbations [EB/OL]. [2020-07-20]. <https://arxiv.org/abs/1702.04267>.
- [51] CARLINI N, WAGNER D. Adversarial examples are not easily detected: Bypassing ten detection methods [C]//Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. New York, USA; ACM Press, 2017; 3-14.
- [52] GONG Z T, WANG W L, KU W. Adversarial and clean data are not twins [EB/OL]. [2020-07-20]. <https://arxiv.org/abs/1704.04960>.
- [53] FEINMAN R, CURTIN R, SHINTRE S, et al. Detecting adversarial samples from artifacts [EB/OL]. [2020-07-20]. <https://arxiv.org/abs/1703.00410>.
- [54] XU W L, EVANS D, QI Y J. Feature squeezing: detecting adversarial examples in deep neural networks [C]//Proceedings of 2018 Network and Distributed System Security Symposium. Washington D. C., USA; IEEE Press, 2018; 15-26.
- [55] PAPERNOT N, MCDANIEL P, JHA S, et al. The limitations of deep learning in adversarial settings [C]//Proceedings of 2016 IEEE European Symposium on Security and Privacy. Washington D. C., USA; IEEE Press, 2016; 372-387.
- [56] PANG T Y, DU C, ZHU J. Robust deep learning via reverse cross-entropy training and thresholding test [EB/OL]. [2020-07-20]. <https://arxiv.org/abs/1706.00633>.
- [57] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks [EB/OL]. [2020-07-20]. <https://arxiv.org/abs/1706.06083>.
- [58] TRAMER F, KURAKIN A, PAPERNOT N, et al. Ensemble adversarial training: attacks and defenses [EB/OL]. [2020-07-20]. <https://arxiv.org/abs/1705.07204>.
- [59] KANNAN H, KURAKIN A, GOODFELLOW I. Adversarial logit pairing [EB/OL]. [2020-07-20]. <https://arxiv.org/abs/1803.06373>.
- [60] XIE C H, WU Y, MAATEN L, et al. Feature denoising for improving adversarial robustness [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA; IEEE Press, 2019; 501-509.
- [61] SCHOTT L, RAUBER J, BETHGE M, et al. Towards the first adversarially robust neural network model on MNIST [EB/OL]. [2020-07-20]. <https://arxiv.org/abs/1805.09190>.
- [62] YANG Y Z, ZHANG G, KATABI D, et al. ME-Net: towards effective adversarial robustness with matrix estimation [EB/OL]. [2020-07-20]. <https://arxiv.org/abs/1905.11971>.
- [63] GUO C, RANA M, CISSE M, et al. Countering adversarial images using input transformations [EB/OL]. [2020-07-20]. <https://arxiv.org/abs/1711.00117>.
- [64] BUCKMAN J, ROY A, RAFFEL C, et al. Thermometer encoding: one hot way to resist adversarial examples [EB/OL]. [2020-07-20]. https://machine-learning-and-security.github.io/papers/mlsec17_paper_26.pdf.
- [65] GU S, RIGAZIO L. Towards deep neural network architectures robust to adversarial examples [EB/OL]. [2020-07-20]. <https://arxiv.org/pdf/1412.5068v1.pdf>.
- [66] PAPERNOT N, MCDANIEL P, WU X, et al. Distillation as a defense to adversarial perturbations against deep neural networks [C]//Proceedings of IEEE Symposium on Security and Privacy. Washington D. C., USA; IEEE Press, 2016; 582-597.
- [67] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network [EB/OL]. [2020-07-20]. <https://arxiv.org/abs/1503.02531>.
- [68] DHILLON G, AZIZZADENESHELI K, LIPTON Z, et al. Stochastic activation pruning for robust adversarial defense [EB/OL]. [2020-07-20]. <https://arxiv.org/abs/1803.01442>.
- [69] XIE C H, WANG J Y, ZHANG Z S, et al. Mitigating adversarial effects through randomization [EB/OL]. [2020-07-20]. <https://arxiv.org/abs/1711.01991>.
- [70] TEJ A R, SUKANTA HALDER S, SHANDEELYA A P, et al. Enhancing perceptual loss with adversarial feature matching for super-resolution [C]//Proceedings of 2020 International Joint Conference on Neural Networks. Washington D. C., USA; IEEE Press, 2020; 168-198.
- [71] ZHANG X Y, MIAN A, GUPTA R, et al. Cassandra: detecting trojaned networks from adversarial perturbations [EB/OL]. [2020-07-20]. <https://arxiv.org/abs/2007.14433>.