中图法分类号:TP309　　**文献标识码**:A　　**文章编号**:1006-8961(2023)12-3629-22

论文引用格式:Sui C H, Wang A, Zhou S W, Zang A K, Pan Y H, Liu H and Wang H P. 2023. A survey on adversarial training for robust learning. Journal of Image and Graphics, 28(12):3629-3650(隋晨红, 王奥, 周圣文, 臧安康, 潘云豪, 刘颢, 王海鹏. 2023. 面向鲁棒学习的对抗训练技术综述. 中国图象图形学报, 28(12):3629-3650)[DOI:10.11834/jig.220953]

面向鲁棒学习的对抗训练技术综述

隋晨红[1],王奥[1],周圣文[1],臧安康[1],潘云豪[1],刘颢[2,3],王海鹏[4*]

1. 烟台大学物理与电子信息学院,烟台　264002; 2. 上海交通大学电子信息与电气工程学院,上海　200240;
3. 武汉数字工程研究所,武汉　430205; 4. 中国人民解放军海军航空大学信息融合研究所,烟台　264001

摘　要:深度学习在众多领域取得了巨大成功。然而,其强大的数据拟合能力隐藏着不可解释的"捷径学习"现象,从而引发深度模型脆弱、易受攻击的安全隐患。众多研究表明,攻击者向正常数据中添加人类无法察觉的微小扰动,便可能造成模型产生灾难性的错误输出,这严重限制了深度学习在安全敏感领域的应用。对此,研究者提出了各种对抗性防御方法。其中,对抗训练是典型的启发式防御方法。它将对抗攻击与对抗防御注入一个框架,一方面通过攻击已有模型学习生成对抗样本,另一方面利用对抗样本进一步开展模型训练,从而提升模型的鲁棒性。为此,本文围绕对抗训练,首先,阐述了对抗训练的基本框架;其次,对对抗训练框架下的对抗样本生成、对抗模型防御性训练等方法与关键技术进行分类梳理;然后,对评估对抗训练鲁棒性的数据集及攻击方式进行总结;最后,通过对当前对抗训练所面临挑战的分析,本文给出了其未来的几个发展方向。

关键词:深度学习;对抗攻击;对抗防御;对抗训练;鲁棒性

A survey on adversarial training for robust learning

Sui Chenhong[1], Wang Ao[1], Zhou Shengwen[1], Zang Ankang[1], Pan Yunhao[1], Liu Hao[2,3], Wang Haipeng[4*]

1. School of Physics and Electronic Information, Yantai University, Yantai　264002, China;
2. School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai　200240, China;
3. Wuhan Digital Engineering Institute, Wuhan　430205, China; 4. Information Fusion Research Institute, Naval Aviation University of the Chinese People's Liberation Army, Yantai　264001, China

Abstract: Deep learning has achieved great success in many fields. However, the solid data-fitting ability of deep learning hides the unexplained phenomenon of "shortcut learning", which leads to the vulnerability of the deep model. Many studies have shown that if an attacker adds slight perturbations to normal data that human beings cannot perceive, the model may produce catastrophic wrong output, which severely limits the application of deep learning in security-sensitive fields. Therefore, to deal with the threat of malicious attacks, an antagonistic defense should be set up, and the robustness of the model should be improved. In this regard, researchers have proposed a variety of adversarial defense methods. The existing defense methods for deep neural networks can be divided into three categories, namely, modifying-input-data-based

收稿日期:2022-10-27;**修回日期**:2023-03-15;**预印本日期**:2023-03-22

*通信作者:王海鹏　whp5691@163.com

基金项目:国家自然科学基金项目(61601397,62076249);国防科技基础加强技术领域基金项目(2022-JCJQ-JJ-0287);山东省自然科学基金创新发展联合基金项目(ZR202209130044)

Supported by: National Natural Science Foundation of China (61601397, 62076249); National Defense Basic Scientific Research Funds (2022-JCJQ-JJ-0287); Shandong Natural Science Foundation Innovation and Development Joint Fund Project (ZR202209130044)

methods, directly-enhancing-network-based methods, and adversarial-training-based methods. Modifying-input-data-based defense methods aim to alter the input in advance and reduce the attack intensity at the input end via denoising or image transformation, among others. Despite showing a certain anti-attack ability, this method is not only limited by the attack intensity but also faces the problem of over-correcting the normal input data. The former limitation hinders this method from dealing with slight disturbances that human beings cannot perceive, while the latter problem exposes this method to the risk of making wrong judgments on normal data, thereby reducing its classification accuracy. Directly-enhancing-network-based methods directly improve the anti-attack capability of the network by adding subnetworks and by changing the loss function, activation function, batch normalization layer, or network training process. Adversarial-training-based methods are typical heuristic defense methods compared with the other two. These methods inject the adversary attack and adversary defense into a framework, wherein adversarial examples are initially generated by attacking the existing models. Afterward, these adversarial examples are used to train the target model to produce an accurate output for these examples and enhance its robustness. Therefore, this paper primarily focuses on adversarial training. Apart from showing certain ability to defend against attacks, adversarial training also improves the robustness of the model at the cost of reducing its classification or recognition accuracy for normal data. Many researchers find that the more robust the model is, the lower is its classification or recognition accuracy for normal examples. In addition, the defense effect of the current adversarial training remains unsatisfactory for strong adversarial attacks with diversified attack modes. To address this issue, recent studies have improved the standard adversarial training from different perspectives. For instance, some studies have generated adversarial examples with high diversity or portability in the attack stage. To enhance model robustness, many scholars have combined adversarial training with network enhancement to resist an adversarial attack. This process involves network structure modification, model parameters adjustment, and adversarial training acceleration, which help the model resist different types of attacks. The standard adversarial training only considers the classification of adversarial examples in the defense stage and ignores the classification accuracy for the original examples. In this connection, many works not only introduce the spatial or semantic consistency constraints between the original and adversarial examples but also require the model to produce an accurate output with respect to the latter, thus ensuring that the model considers both robustness and accuracy. To enhance the transferability of the model, curriculum learning, reinforcement learning, metric learning, and domain adaptation technologies are integrated into adversarial training. This paper then comprehensively reviews adversarial training technologies. First, the basic framework of adversarial training is elaborated. Second, typical methods and critical technologies for the generation of adversarial samples are reviewed. We summarize the adversarial examples generation methods based on image space, feature space, and physical space attacks. To improve the diversity of adversarial examples, we also introduce interpolation- and reinforcement-learning-related adversarial example generation strategies. Given that standard adversarial training is extremely time consuming, we briefly describe optimization strategies based on temporal, spatial, and spatiotemporal mixed momentum, which are conducive to improving training efficiency. Defense is the fundamental problem of adversarial training that is devoted to absorbing the generated adversarial examples for training via loss minimization. Therefore, we briefly review the technologies typically used in the defensive training stage, including the loss regularization term, model enhancement mechanism, parameter adaptation, early stop, and semi-supervised or unsupervised expansion strategies. To evaluate the robustness of the model, we summarize the popular datasets and typical attack methods. After sorting out relevant adversarial training technologies, we still face challenges in dealing with multi-disturbance integrated attacks and the low efficiency of the model. We put forward these problems as directions for future research on adversarial training.

Key words: deep learning; adversarial attack; adversarial defense; adversarial training; robustness

0 引 言

深度神经网络(deep neural network, DNN)在计算机视觉领域取得了巨大成功,如面部识别、语音识别、医学成像、自动驾驶汽车、机器翻译和恶意软件检测等,这些实际应用的飞速发展使深度学习的安全可靠性问题日益凸显。研究发现,模型参数量的

庞大和对训练样本的极度依赖，导致DNN具有严重的对抗脆弱性。具体而言，攻击者向原始数据添加人类视觉系统不可察觉的微小扰动，生成对抗样本，该对抗样本能让DNN产生错误的输出结果，达到攻击DNN的目的（Goodfellow等，2015）。DNN的这种对抗脆弱性诱发了恶意攻击的大量涌入，给深度学习在各领域的应用带来了严峻的安全性挑战（Sharif等，2016；Wu等，2020b；余正飞等，2022；吴翼腾 等，2022）。为应对恶意攻击的威胁，开展对抗性防御进而提升模型鲁棒性变得尤为重要。

针对DNN的防御方法主要包括3类：基于修正输入数据的方法、直接增强网络的方法以及对抗训练（Madry等，2017）的方法。基于修正输入数据的防御方法旨在对输入DNN的数据提前进行修正，在输入端减弱攻击强度，如去噪（Xie等，2019）、图像变换（Xie等，2018）等。该类方法具有一定的抵御攻击能力，但其不仅受攻击强度限制，而且存在对正常输入数据过度修正的问题。前者导致其难以应对人类不可感知的微小扰动，而后者极易造成其对正常数据亦给出错误判断，从而降低模型的分类准确率。直接增强网络的方法主要是通过添加子网络（Bai等，2022）、更改损失函数、激活函数（Singla等，2021）、BN（batch normalization）层（Xie等，2020）或网络训练过程（Hendrycks 等，2019；Xiao 和 Zheng，2020；Chen等，2020；Cui等，2021）等方式，直接增强网络的抗攻击能力。

相较前两类方法，对抗训练是典型的启发式防御方法（孔锐 等，2022；李前 等，2022；Liu等，2020）。其核心是将对抗攻击与对抗防御注入一个框架，一方面利用对抗攻击已有模型，生成强干扰的对抗样本；另一方面利用对抗样本进一步训练目标模型，使其能够对对抗样本有准确的输出，进而提升模型抵御对抗攻击的鲁棒性。

尽管对抗训练具有一定的防御攻击能力，但其对模型鲁棒性的提升是以降低对正常数据的分类或识别精度为代价的。实践表明，模型越鲁棒，其对正常样本的分类或识别精度越低（Zhang等，2019b）。此外，针对攻击方式多样化的强对抗攻击，目前的对抗训练防御效果依然不尽人意，并且标准对抗训练存在耗时长的问题。

研究工作从不同角度改进了标准的对抗训练，如在攻击阶段，使生成的对抗样本具有多样性或可迁移性，能够使模型抵挡其他更多种类的攻击；在防御阶段，标准对抗训练仅考虑了对抗样本的分类情况，忽略了原始样本的分类精度要求。为此，许多工作（Kannan 等，2018；Zhang 等，2019b；Wang 等，2020）不仅引入对抗样本与原始样本的空间或语义一致性约束，而且要求模型对原始样本及对抗样本均有准确的输出，从而保证模型同时兼具高鲁棒性与高准确性。

为进一步提升模型的鲁棒性，研究者提出将对抗训练与直接增强网络的方法进行结合。例如，许多工作将修改模型网络结构（Mustafa等，2019；Xie等，2020；Bai等，2022）、自适应调整模型参数（Chen等，2020；Xiong和Hsieh，2020；Liu等，2020；Ye等，2021）、加速对抗训练（Zhang等，2019a；Zhang等，2019b；Shafahi 等，2019；Zheng 等，2020；Wong 等，2020；Kim等，2021）等嵌入对抗训练框架，增强模型抵御对抗攻击的能力。此外，为提升模型的迁移性，研究者将课程学习、强化学习、度量学习以及领域自适应等技术引入对抗训练（Cai等，2018；Sitawarin等，2021；Bashivan等，2022；Jia等，2022；钱申诚 等，2022）。例如，CAT（curriculum adversarial training）（Cai等，2018）通过引入课程学习的思想，使得对抗训练过程的扰动范围由小到大变化，增强了对抗攻击强度的多样性，使模型的迁移性能得到显著提升。

为此，本文围绕对抗训练，从对抗训练的基本框架入手，分别就对抗样本生成、对抗模型防御训练等相关技术进行详细分类与梳理，并对对抗训练性能评估涉及的数据集和对抗攻击方法展开介绍，最后通过对当前对抗训练所面临挑战的分析，对对抗训练未来的发展趋势进行了展望。

1 对抗训练的基本框架

深度学习模型的强大拟合能力，使其能够对数据产生丰富的表征。因此，当攻击者对原数据$\boldsymbol{x}$施加轻微扰动$\boldsymbol{\delta}$，便可引起特征表达的显著变化，进而导致错误的分类或检测。这种人为添加扰动而生成的样本称为对抗样本（adversarial example）$\boldsymbol{x}_{\text{adv}}$，其与扰动$\boldsymbol{\delta}$的关系可简单描述为

$$\begin{aligned} &\boldsymbol{x}_{\text{adv}} = \boldsymbol{x} + \boldsymbol{\delta} \\ &\text{s.t.} \quad \|\boldsymbol{\delta}\|_p \leqslant \varepsilon \end{aligned} \tag{1}$$

式中，ε为攻击或扰动的幅值，$\|\cdot\|_p$代表p范数，p可取0, 1, 2, ∞。

为提高模型对抗攻击的稳健性，对抗训练的核心思想是利用具有攻击性的对抗样本进行模型训练，使学习到的模型能抵御攻击。这表明包含微小扰动$\boldsymbol{\delta}$的对抗样本生成是对抗训练的关键。其中，扰动$\boldsymbol{\delta}$一方面要保证对抗样本的强攻击性，另一方面要满足低幅性，具体为

$$\boldsymbol{\delta} = \underset{\|\boldsymbol{\delta}\|_p \leqslant \varepsilon}{\operatorname{argmax}} \mathcal{L}\left(\boldsymbol{x} + \boldsymbol{\delta}, y; \boldsymbol{\theta}\right) \tag{2}$$

式中，y为原数据$\boldsymbol{x}$的真实标签，$\boldsymbol{\theta}$为模型参数，$\mathcal{L}$为度量模型预测结果与真实标签y之间一致性的损失函数。

为防御攻击，对抗训练在依据式(2)与式(1)获取对抗样本$\boldsymbol{x}_{\text{adv}}$后，将$\boldsymbol{x}_{\text{adv}}$作为训练样本输入模型并进行训练，则模型参数$\boldsymbol{\theta}$可通过最小化平均损失函数$\mathcal{L}(\boldsymbol{x}_{\text{adv}}, y;\ \boldsymbol{\theta})$进行更新，即

$$\boldsymbol{\theta} = \arg\min_{\boldsymbol{\theta}} E_{(\boldsymbol{x},y)\sim \boldsymbol{D}}\left[\mathcal{L}\left(\boldsymbol{x} + \boldsymbol{\delta}, y; \boldsymbol{\theta}\right)\right] \tag{3}$$

式中，$(\boldsymbol{x}, y)\sim \boldsymbol{D}$表示从分布$\boldsymbol{D}$中采样的数据及标签对，$E[\cdot]$表示期望。

基于这一思想，Madry等人(2017)从鲁棒优化的角度，使用自然鞍点(最大—最小)公式将对抗攻击和防御训练同时纳入对抗训练，构建了对抗训练的基本框架，具体为

$$\min_{\boldsymbol{\theta}} E_{(\boldsymbol{x},y)\sim \boldsymbol{D}}\left[\max_{\boldsymbol{\delta}\in \boldsymbol{B}(\boldsymbol{x},\varepsilon)} \mathcal{L}\left(\boldsymbol{x} + \delta, y; \boldsymbol{\theta}\right)\right] \tag{4}$$

式中，$\boldsymbol{B}(\boldsymbol{x}, \varepsilon)$表示满足$\|\boldsymbol{\delta}\|_p \leqslant \varepsilon$的扰动集合。

由式(4)不难看出，对抗训练是集攻击与防御于一体的启发式方法。一方面，其通过内层损失最大化来生成具有攻击性的对抗样本；另一方面，其利用外层损失最小化来更新模型，进而增强模型的抗攻击能力。

为此，本文将分别就对抗样本生成与模型防御训练两方面展开介绍。

2 对抗样本生成

在对抗训练的攻击阶段，依据攻击的对象不同，可将现有的对抗样本生成技术分为3类：基于图像空间攻击的对抗样本生成、基于特征空间攻击的对抗样本生成以及基于物理空间攻击的对抗样本生成。其中，基于图像空间的攻击是目前对抗训练采用的主流对抗样本生成方式(钱申诚 等，2022；赵宏等，2022)。其通过限制扰动幅度，能够生成人类感知不到扰动的对抗样本，具有较好的攻击隐藏性。因此，本文将重点对基于图像空间攻击的对抗样本生成方式展开分析。

2.1 基于图像空间攻击的对抗样本生成

基于图像空间的攻击方法首先基于损失函数最大化，获得满足幅值限制的扰动$\boldsymbol{\delta}$，然后将$\boldsymbol{\delta}$添加到原始图像$\boldsymbol{x}$生成对抗样本，如式(4)所示。典型方法如FGSM(fast gradient sign method)(Goodfellow 等，2015)、FGSM的迭代变体I-FGSM(iterative FGSM)(Kurakin等，2017)、基于动量的迭代变体MI-FGSM(momentum-based iterative FGSM)(Dong等，2018)以及PGD方法(projected gradient descent method)(Madry等，2017)等。这些方法均通过损失函数的梯度进行图像空间攻击，进而生成对抗样本。

表1列出了一系列基于梯度的对抗样本生成方法，总结了它们的应用范围、算法思想和优点。图1简要给出了基于损失函数梯度攻击原始图像，然后生成对抗样本的演进过程。在图1实线框内的方法是常用的基于梯度的攻击方法，而虚线框内的方法能够进一步增强对抗样本的可迁移性。

由图1不难看出，FGSM作为最基本的单步迭代方法，是其他基于梯度攻击图像的方法的基础。FGSM是通过对原始数据添加与损失函数的梯度方向一致而幅度较小的微小扰动得到对抗样本，具体为

$$\boldsymbol{x} + \delta\cdot\operatorname{sign}\left(\nabla_{\boldsymbol{x}}\boldsymbol{J}\left(\boldsymbol{x}, y; \boldsymbol{\theta}\right)\right) \tag{5}$$

式中，$\boldsymbol{J}(\boldsymbol{x}, y;\ \boldsymbol{\theta})$表示损失函数，而$\nabla_{\boldsymbol{x}}\boldsymbol{J}(\boldsymbol{x}, y)$是损失函数关于$\boldsymbol{x}$的梯度，sign(·)为符号函数，用于提取梯度的方向。特别地，为限制扰动的强度，$\boldsymbol{\delta}$需满足$\|\boldsymbol{\delta}\|_\infty < \varepsilon$。

显然，FGSM通过单步计算损失函数的梯度完成对抗样本的生成，无需迭代，具有效率高的优势。然而，这种单步生成对抗样本的方式导致FGSM缺乏对对抗样本多样性和合理性的探索，不利于提升模型的抗攻击能力。

为此，Kurakin等人(2017)在FGSM的基础上提

表 1　基于梯度的对抗样本生成方法对比

Table 1　Comparison of gradient-based attack methods

攻击方法	攻击方式	算法思想	优点
FGSM(Goodfellow 等,2015)	单步白盒攻击	梯度上升使损失最大化得到对抗扰动	计算效率高
I-FGSM(Kurakin 等,2017)	多步白盒攻击	迭代的 FGSM	梯度方向更精确,攻击成功率高
PGD(Madry 等,2017)	多步白盒攻击	一般添加初始化扰动	攻击强度大
MI-FGSM(Dong 等,2018)	多步白盒攻击	增加动量项	梯度更新方向更稳定
NI-FGSM(Lin 等,2020)	多步白盒攻击	使用 Nesterov 加速梯度	改进动量项,提高了对抗样本的可迁移性
VM(N)I-FGSM(Wang 和 He,2021)	多步白盒攻击	利用方差计算梯度	有效避免搜索梯度更新方向时出现局部最优
SMI-FGSM(Wang 等,2022a)	多步白盒攻击	引入空间域计算梯度	从时间和空间域同时稳定梯度的更新方向

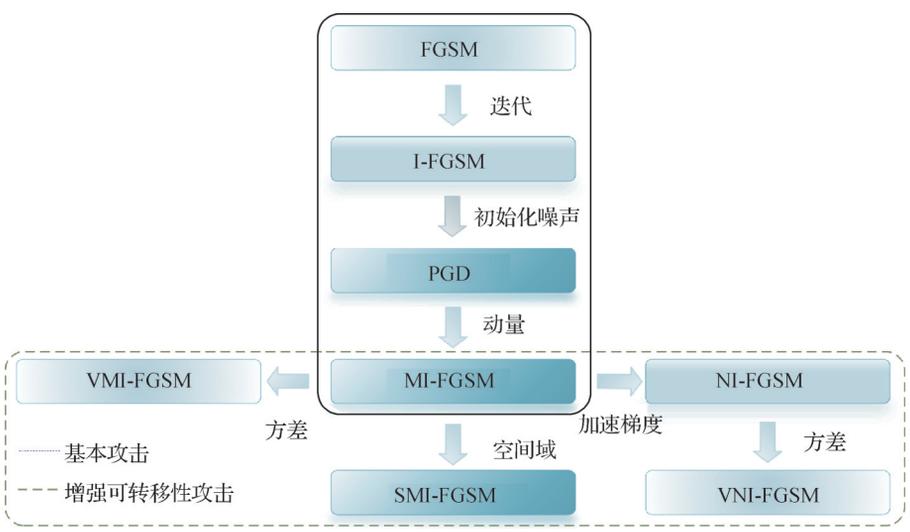


图 1　基于梯度的攻击方法演变过程

Fig. 1　Evolution of gradient-based attack methods

出了迭代版 FGSM,即 I-FGSM。I-FGSM 沿着梯度增加的方向进行多步微小扰动,并在每次扰动后,重新计算梯度方向。若设初始的对抗样本为

$$\boldsymbol{x}_0^{\mathrm{adv}} = \boldsymbol{x} \tag{6}$$

则第 $t+1$ 步更新后的对抗样本为

$$\boldsymbol{x}_{t+1}^{\mathrm{adv}} = Clip_{\boldsymbol{x},\varepsilon}\left\{\boldsymbol{x}_t^{\mathrm{adv}} + \alpha\cdot\mathrm{sign}\left(\nabla_{\boldsymbol{x}_t^{\mathrm{adv}}}\boldsymbol{J}\left(\boldsymbol{x}_t^{\mathrm{adv}}, y; \boldsymbol{\theta}\right)\right)\right\} \tag{7}$$

式中,$Clip_{\boldsymbol{x},\varepsilon}(\boldsymbol{A})$ 表示将输入向量 $\boldsymbol{A}_{i,j}$ 中的每个元素裁剪到 $\left[\boldsymbol{x}_{i,j}-\varepsilon, \boldsymbol{x}_{i,j}+\varepsilon\right]$ 之间的操作,用于控制扰动的强度。

为提高对抗样本的可迁移性,MI-FGSM(Dong 等,2018)将动量整合到 I-FGSM 中,实现了对抗样本更高的迁移性,具体为

$$\boldsymbol{g}_{t+1} = \mu\times\boldsymbol{g}_t + \frac{\nabla_{\boldsymbol{x}_t^{\mathrm{adv}}}\boldsymbol{J}\left(\boldsymbol{x}_t^{\mathrm{adv}}, y; \boldsymbol{\theta}\right)}{\left\|\nabla_{\boldsymbol{x}_t^{\mathrm{adv}}}\boldsymbol{J}\left(\boldsymbol{x}_t^{\mathrm{adv}}, y; \boldsymbol{\theta}\right)\right\|_1} \tag{8}$$

$$\boldsymbol{x}_{t+1}^{\mathrm{adv}} = \boldsymbol{x}_t^{\mathrm{adv}} + \alpha\cdot\mathrm{sign}\left(\boldsymbol{g}_{t+1}\right) \tag{9}$$

式中,$\boldsymbol{g}_t$ 为前 t 次迭代后的累积梯度,μ 为衰减因子。

此外,与 I-FGSM 不同的是,目前最流行的 PGD(Madry 等,2017)攻击是对原始样本添加其邻域范围内的随机扰动 S 作为初始对抗样本,具体为

$$\boldsymbol{x}_0^{\mathrm{adv}} = \boldsymbol{x} + S \tag{10}$$

PGD 第 $t+1$ 步迭代生成的对抗样本为

$$\boldsymbol{x}_{t+1}^{\mathrm{adv}} = \prod_{\boldsymbol{x}+S}\left\{\boldsymbol{x}_t^{\mathrm{adv}} + \alpha\cdot\mathrm{sign}\left(\nabla_{\boldsymbol{x}_t^{\mathrm{adv}}}\boldsymbol{J}\left(\boldsymbol{x}_t^{\mathrm{adv}}, y; \boldsymbol{\theta}\right)\right)\right\} \tag{11}$$

式中,$\prod_{\boldsymbol{x}+S}\{\cdot\}$ 代表将每次更新迭代后的对抗样本投影到规定的阈值范围内。

上述 I-FGSM、MI-FGSM 以及 PGD 等方法均利用损失函数的梯度,将损失函数增加的方向定义为扰动更新的最优方向。显然,这有利于增加对抗样本的攻击性。然而,受扰动幅度限制,模型在该类方法

所生成的对抗样本附近，具有高度非线性的损失表面，导致我们需要通过多次迭代更新扰动方向，进而找到合适的对抗扰动。这极大地降低了对抗训练的收敛速度。

为降低对抗样本的更新步数，Qin等人（2019）将正则项引入目标函数，旨在通过提升局部线性，减少生成对抗样本所需的迭代步数。Sriramanan等人（2020）通过引入正则化约束来寻找最优梯度方向，从而为对抗样本的生成提供可靠的扰动，使得单步迭代能产生更强的攻击。

另外，为加强对抗样本的多样性，进而增加模型的泛化性，集成对抗训练、多样性对抗训练等方法（Tramèr和Boneh，2019；Jang等，2019；Kariyappa和Qureshi，2019；Addepalli等，2022）相继提出。例如，集成对抗训练（Tramèr等，2020）旨在利用不同模型生成的对抗样本进行目标模型训练。Pang等人（2019）提出利用不同模型间的相互作用来提高集成模型的鲁棒性。Kariyappa和Qureshz（2019）提出多样性训练，旨在训练基于输入梯度间余弦距离损失的模型集合。

为进一步提升对抗样本的多样性，插值对抗训练（Zhang等，2021）引入插值的思想。具体来说，首先对输入图像与标签同时插值，然后最小化插值后的图像与原始图像的距离（此时，插值后的图像相当于对抗样本），同时，最大化插值后的标签与原始标签的距离。例如，双边对抗训练（Wang和Zhang，2019）对输入与标签同时添加扰动。Lee等人（2020）针对对抗特征过度拟合（adversarial feature overfitting，AFO）问题，提出了对抗顶点混合（adversarial vertex mixup，AVmixup）方法，即一种软标记数据增强方法。由于大多数插值方法仅在训练阶段进行，Pang等人（2019）发现由于用于训练的插值对抗样本与测试数据间的差异，导致分类器训练所得最佳决策边界并不适用测试数据，并引发泛化误差增加的风险。这说明仅对训练数据进行插值不利于保证模型的泛化性能。为此，Pang等人（2019）提出在训练与测试阶段均引入插值操作，使模型鲁棒性进一步得到提升。

表2进一步总结了上述方法对输入样本和标签的具体插值公式。若C为类别总数，则对抗插值训练（Zhang等，2020）进行标签平滑后的新标签为$\bar{y}'=\frac{1-y'}{C-1}$。在双边对抗训练（Wang和Zhang，2019）中，超参数β用于更新标签，而对抗顶点混合（Lee等，2020）方法使用标签平滑函数ϕ更新标签。

上述对抗样本生成方法均依赖手工设计的规则，因此，一定程度上限制了对抗样本生成。

为避免这一问题，有研究旨在通过网络直接生成对抗样本。Jang等人（2019）用递归生成网络（采用基于U-Net架构的卷积编码器—解码器）代替上述攻击方式生成对抗样本，损失函数由目标分类器的标准损失与多样性损失组成，可以产生更强和更多样的对抗样本；Xiong和Hsieh（2020）同样提出了基于L2L（learning-to-learn）（Chen等，2021）的递归神经网络（recurrent neural network，RNN）对抗训练框架，如图2所示，其框架可以与AT（adversarial training）或TRADES（trade-off between robustness and accuracy）相结合。为保证稳定训练，该框架还去除了标准RNN的偏差项。Chan等人（2020）提出

表2　插值对抗训练方法对比

Table 2　Comparison of interpolation methods

插值方法	样本插值	标签平滑
对抗插值训练（Zhang等，2020）	$\tilde{\boldsymbol{x}}=\boldsymbol{x}-\epsilon\frac{\partial\mathcal{D}(\boldsymbol{x},\boldsymbol{x}')}{\partial\boldsymbol{x}}$	$\tilde{y}=y-\frac{\epsilon_y}{2}\times\frac{\partial\left\|y-\bar{y}'\right\|_2^2}{\partial y}=\left(1-\epsilon_y\right)y+\epsilon_y\bar{y}'$
双边对抗训练（Wang和Zhang，2019）		$\tilde{y}=\frac{\epsilon_y}{C-1},\ \epsilon_y\in\left(\frac{1}{1+\beta},\frac{1}{1+\frac{\beta}{C-1}}\right)$
对抗顶点混合（Lee等，2020）	$\begin{cases}\boldsymbol{x}_{\text{av}}=\boldsymbol{x}+\gamma\boldsymbol{\delta}_{\boldsymbol{x}}\\\tilde{\boldsymbol{x}}=\alpha\boldsymbol{x}+\left(1-\alpha\boldsymbol{x}_{\text{av}}\right)\end{cases}$	$\tilde{y}=\alpha\phi\left(y,\lambda_1\right)+(1-\alpha)\phi\left(y,\lambda_2\right)$
测试阶段插值（Pang等，2020a）	$\tilde{\boldsymbol{x}}=\lambda\boldsymbol{x}+(1-\lambda)\boldsymbol{x}_s$	

利用雅可比矩阵替代对抗样本的对抗正则化网络（Jacobian adversarially regularized networks，JARN）。JARN将目标分类器作为生成对抗网络（generative adversarial network，GAN）（Goodfellow 等，2014）中的生成器模型，通过优化目标分类器使其产生的显著雅可比矩阵能够欺骗鉴别器网络，并将其认定为输入图像，以此来提升目标分类器的鲁棒性。

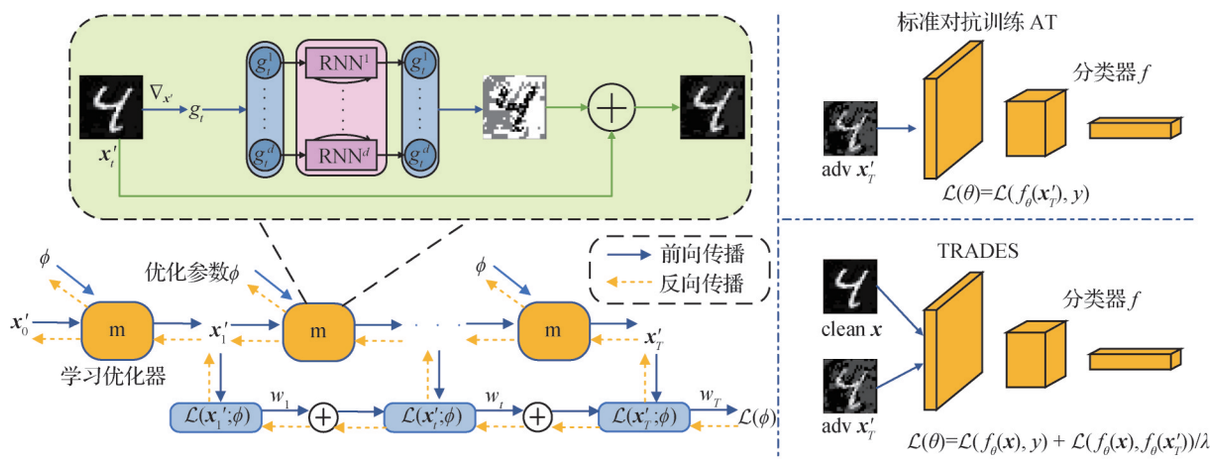


图2　基于L2L的RNN对抗训练框架（Xiong和Hsieh，2020）

Fig. 2　L2L-based RNN adversarial training framework（Xiong and Hsieh，2020）

此外，由于相关研究表明对抗样本具有可迁移性（Liu 等，2017；Papernot 等，2017），有研究工作旨在通过对FGSM进行优化来提升对抗样本的可迁移性（Lin等，2020；Wang和He，2021；Wang等，2022a）。如式（8）（9）所示，Dong等人（2018）提出MI-FGSM，将动量项整合到迭代攻击中，以此稳定更新的梯度方向，进而提高对抗样本的可迁移性。而Lin等人（2020）在此基础上提出NI-FGMS（nesterov iterative fast gradient sign method），通过引入NAG（nesterov accelerated gradient）$\alpha\times\mu\times g_t$来校正累积梯度$g_{t+1}$，加速跳出局部最优解的同时，使对抗样本获得更好的迁移性，具体为

$$g_{t+1}=\mu\times g_t+\frac{\nabla_{x_t^{\text{adv}}}\boldsymbol{J}\left(\boldsymbol{x}_t^{\text{adv}}+\alpha\times\mu\times\boldsymbol{g}_t,y;\boldsymbol{\theta}\right)}{\left\|\nabla_{x_t^{\text{adv}}}\boldsymbol{J}\left(\boldsymbol{x}_t^{\text{adv}}+\alpha\times\mu\times\boldsymbol{g}_t,y;\boldsymbol{\theta}\right)\right\|_1}\tag{12}$$

$$\boldsymbol{x}_{t+1}^{\text{nes}}=\boldsymbol{x}_t^{\text{adv}}+\alpha\times\mu\times\boldsymbol{g}_t\tag{13}$$

与NI-FGSM和MI-FGSM等基于梯度进行动量累积不同的是，Wang和He（2021）提出了一种称为方差调整的新方法。该方法通过在每次迭代中减少梯度的方差来提高对抗样本可迁移性，具体为

$$\boldsymbol{V}(\boldsymbol{x})=\frac{1}{N}\sum_{i=1}^{N}\nabla_{x^i}\boldsymbol{J}\left(\boldsymbol{x}^i,y;\boldsymbol{\theta}\right)-\nabla_x\boldsymbol{J}\left(\boldsymbol{x},y;\boldsymbol{\theta}\right)\tag{14}$$

$$\boldsymbol{g}_{t+1}=\mu\times\boldsymbol{g}_t+\frac{\hat{\boldsymbol{g}}_{t+1}+\boldsymbol{v}_t}{\left\|\hat{\boldsymbol{g}}_{t+1}+\boldsymbol{v}_t\right\|_1}\tag{15}$$

式中，$\boldsymbol{\theta}$为模型参数，$\boldsymbol{J}(\boldsymbol{x},y;\boldsymbol{\theta})$为损失函数，$\boldsymbol{g}_t$为前$t$次迭代后的累积梯度，$\mu$为衰减因子，$\boldsymbol{v}_t$为方差。可以看出，利用方差的攻击可以在MI-FGSM或NI-FGSM的基础上更进一步提升对抗样本的可迁移性。

除了在时间动量上进行优化，Wang等人（2022a）还提出了考虑空间动量的SMI-FGSM（spatial momentum iterative FGSM attack）。SMI-FGSM将动量累积机制从时间域引入到空间域，使用来自不同区域的信息来生成稳定的梯度，最后在时间和空间域同时稳定梯度的更新方向，具体为

$$\boldsymbol{g}_{t+1}^{s}=\sum_{i=1}^{n}\lambda_i\nabla_x\boldsymbol{J}\left(H_i\left(\boldsymbol{x}_t^{\text{adv}}\right),y;\boldsymbol{\theta}\right)\tag{16}$$

$$\boldsymbol{x}_{t+1}^{\text{adv}}=\boldsymbol{x}_t^{\text{adv}}+\alpha\cdot\text{sign}\left(\boldsymbol{g}_{t+1}^{s}\right)\tag{17}$$

式中，$H_i(\cdot)$表示对输入进行随机变换的函数，n表示空间域中进行随机变换的次数，λ_i是梯度的权重，$\sum\lambda_i=1$，$\lambda_i=1/n$。

显然，不同于NI-FGSM和MI-FGSM等方法在时间域累积动量，SMI-FGSM通过计算n种随机变换图像的平均梯度，得到空间域上的动量累积。

此外，Zheng等人（2020）根据替代模型训练的思想，基于相邻训练epoch的模型之间有高度可迁移性的特性，提出对抗训练可以由逐epoch累积的对抗扰动来增强训练模型的鲁棒性，以更少的迭代

生成相似(甚至更强)的对抗样本;还有研究提出对抗性变换网络代替传统的图像变换(Wu等,2021),以此增强对抗样本的可迁移性。

2.2 基于特征空间攻击的对抗样本生成

除了针对图像空间进行攻击生成对抗样本之外,还可以基于特征空间的攻击生成对抗样本,即在网络层添加噪声。如图3所示,输入图像进入某一DNN网络,可以在中间隐藏层的某几层或全部层添加噪声,将添加了噪声的特征图依次经过全连接层与分类层输出,得到最终的对抗样本。

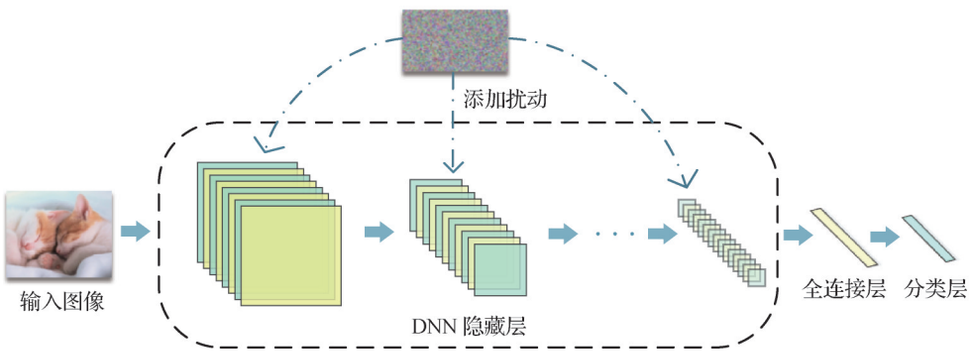


图3 特征空间添加扰动方法

Fig. 3 Method of adding perturbation to feature space

参数噪声注入(He等,2019)旨在在网络每一层的激活或权重上注入可训练的高斯噪声,并嵌入对抗训练;在此基础上,Jeddi等人(2020)提出了基于Learn2Perturb学习框架的扰动注入模块,如图4所示,并把模型训练分为两个部分:1)扰动注入网络训练:对特征注入扰动的情况下继续更新网络参数,以提高模型的对抗鲁棒性;2)扰动注入模块训练:更新扰动注入模块的参数,以增强针对改进网络的扰动能力。

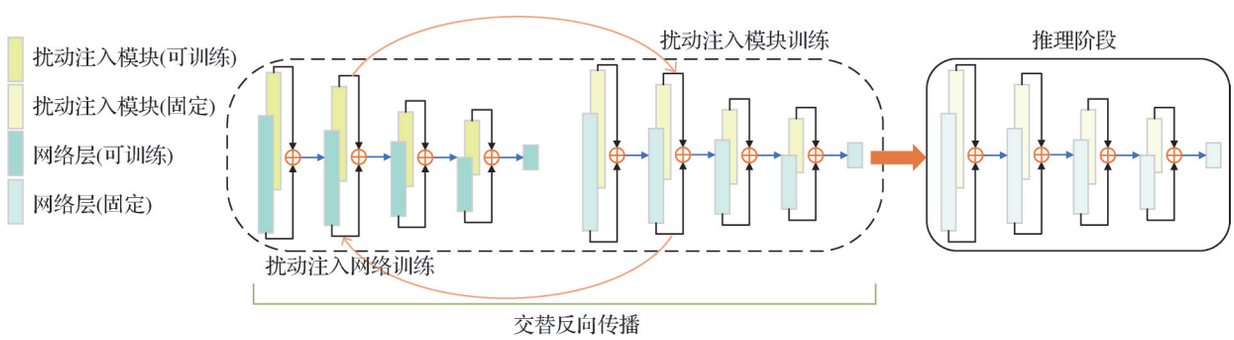


图4 扰动注入模块下的模型训练策略(Jeddi等,2020)

Fig. 4 Model training strategy under the perturbation injection module(Jeddi et al. ,2020)

Yu等人(2021)试图研究网络中间层的潜在特征,并研究了选择哪些中间层最适合攻击,使网络原始输出与特征空间的输出共同决定网络最终的输出预测,并使用替代损失函数来更新对抗样本。

逐层对抗训练(Chen等,2020)得出结论:1)将扰动注入到网络的每一层(包括输入层)会得到最高的模型鲁棒性;2)接受对抗性训练的层次越多,防御性能越强;3)对抗性训练层越靠近网络输出层,防御性能越强。

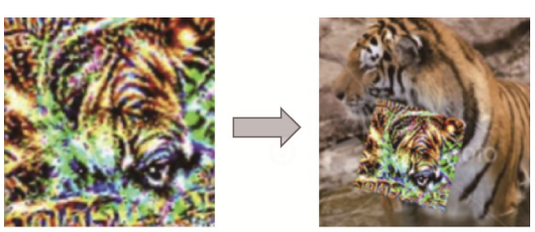
此外,由于对抗训练的模型更偏向于全局特征,Song等人(2020)试图研究对抗性训练的泛化和鲁棒的局部特征之间的关系。通过对对抗样本随机区块洗牌(random block shuffle,RBS)变换获得局部特征,然后将其迁移到正常对抗样本的训练中;特征级可迁移攻击(Zhang等,2022)旨在生成更多可迁移的对抗样本,通过对神经元重要性进行估计,破坏正面特征或放大负面特征来生成对抗样本。

2.3 基于物理空间攻击的对抗样本生成

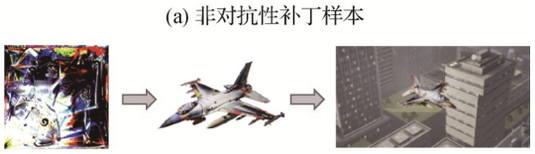
2.1和2.2节分别介绍了如何在图像或特征空间生成人类视觉系统察觉不到的微小扰动。但在实际情况中,DNN还表现出对日常环境中常见自然破

坏(如雪、雨、亮度等)的弱鲁棒性，这些人眼可识别的物理扰动的存在，对人们的日常生活造成了极大的干扰，不仅如此，Brown等人(2018)在原始图像上添加看得见的物理补丁(Patch)作为扰动，同样能产生相同的后果。如在实际应用中，攻击者把这些恶意扰动应用在自动汽车驾驶上，可能会造成严重的交通事故。所以现实生活中更需要采取一系列有效防御措施，以在物理世界中建立同样鲁棒的深度学习模型(袁珑 等，2022)。

在这一领域，Salman等人(2021)采用相反思路构建非对抗性扰动，如图5所示，有两种方式生成非对抗性扰动：1)设计非对抗性补丁；2)设计非对抗性纹理。通过改变输入来强化正确的行为，而不是优化输入来误导模型。也就是说，当得到非对抗性样本后，与传统对抗训练内部最大化目标函数相反，本方法要使目标分类器的损失最小；类似地，Wang等人(2022b)提出了防御性补丁生成框架，旨在通过利用局部特征与全局特征加强模型的泛化性与不同模



(a) 非对抗性补丁样本



(b) 非对抗纹理样本

图5　非对抗性扰动(Salman等，2021)

Fig. 5　Unadversarial perturbation(Salman et al.，2021)
((a) an example unadversarial patch;
(b) an example unadversarial texture)

型之间的可迁移性，与Salman等人(2021)不同的是，其通过掩码的方式设计的补丁粘贴在目标图像周围，不会对其造成干扰，并且该框架可以与对抗训练结合，得到更高的鲁棒精确度。

3　防御性训练

对抗训练的防御阶段，其基本问题是利用已生成的对抗样本进行训练，并通过分类或检测损失最小化来更新模型，具体如式(4)的外部最小化学习过程。为进一步提升模型的鲁棒性，防御训练阶段还会结合以下技术：1)引入损失正则项。将正常样本与对抗样本的类别一致性、语义一致性或分布一致性等正则项引入损失函数，提升对抗训练的鲁棒性。2)增强模型。设计优化模块插入到网络模型结构中，或修改网络的某些部件，如BN层或激活函数，亦或是使用预训练模型训练。3)参数自适应。不再是手工设置某些具体的参数，而是根据训练情况使其参数自适应更改，如内部最大化时使用的步长、扰动约束，外部最小化使用的平衡参数等。4)早期停止。一般来说，模型训练收敛后的鲁棒性不是最佳鲁棒性(Rice等，2020)，所以需要采取早期停止策略得到最佳鲁棒性。5)半监督或无监督扩展。对抗训练提升模型鲁棒性的核心思想是利用大量原始数据生成对抗样本，进而开展对抗训练，使得模型对添加一定扰动的输入数据仍能产生与原来一致的输出。因此，为增加模型鲁棒性，许多研究工作指出可利用无标记样本进行半监督或无监督扩充，然后用于对抗训练(Carmon等，2019；Najafi等，2019；Zhai等，2019)。Carmon等人(2019)发现依靠有标签数据训练出的模型，能够获得无标签数据的伪标签，若利用这些伪标签数据进行对抗训练，依然能够提升模型的鲁棒性。6)加速对抗训练。对抗训练最棘手的缺点之一是训练时间过长。针对该问题，研究者们提出了一系列加速对抗训练的方法。下面将分别就上述技术展开介绍。

3.1　引入损失正则项

对抗训练防御阶段的核心是通过已生成的对抗样本来训练，基于损失最小化来更新优化模型。由于标准对抗训练的损失函数仅由生成的对抗样本与原始标签决定，没有考虑到原始样本和相应的对抗样本之间的关系。所以Kannan等人(2018)通过增加目标函数正则项，鼓励两对样本的输出逻辑(logits)相似，进一步增强了二者的相关性。

Tsipras等人(2019)发现模型鲁棒性和泛化性在本质上是矛盾的，存在权衡。为找到二者之间的平衡点，缩小鲁棒泛化差距，即在提升模型鲁棒性的同时，不会大幅度损害模型的泛化性，Zhang等人(2019b)把标准对抗训练的鲁棒误差分为自然误差和边界误差，并设计了一个新的优化目标，即目标分类器原始样本的目标函数与使用了KL(Kullback-

Leibler)散度测量对抗样本与原始样本之间距离的组合;与上文类似,在单步训练的基础上,Sriramanan等人(2020)在标准对抗训练中引入了松弛项,旨在找到最优的梯度方向。

Wan等人(2020)考虑了对抗样本与原始样本之间特征分布的差异性,旨在通过训练来学习数据的特征分布。对抗样本应该遵循不同于干净数据的分布。为了缩小不同类别之间的鲁棒精确度差异,Xu等人(2021)提出了公平性约束,但在差异减小的同时,一定程度上降低了对抗训练的最高准确度。表3列出了不同防御方法的损失函数,可以明显看出,标准对抗训练(PGD-K)仅考虑了对抗样本的分类损失,而后续基于梯度优化的方法增加了原始样本的分类损失以及对抗样本与原始样本的距离损失等正则项,或使用了替代损失函数;基于特征空间优化的方法还考虑了特征空间中类别之间的联系,即有效增加类间聚敛以及类内损失正则项,或通过添加辅助分类器得到了特征空间隐藏层的损失正则项。

此外,由于上述方法仅在正确分类的原始样本上生成对抗样本,不可避免地存在部分样本在训练中会被错误分类,Ding等人(2020)与Wang等人(2019)考虑了这种情况,其中,Ding等人(2020)从边际(输入到分类器的决策边界的距离)最大化的角度研究神经网络的对抗鲁棒性;而Wang等人(2019)提出了错误分类感知对抗训练,使用了替代损失函数进行训练。

表3　不同防御方法的损失函数

Table 3　Loss functions for different defense methods

防御方法	损失函数
PGD-K(Madry等,2017)	$CE\left(f_\theta(\hat{\boldsymbol{x}}),y\right)$
ALP(Kannan等,2018)	$CE\left(f_\theta(\hat{\boldsymbol{x}}),y\right)+\lambda\times\left\|f_\theta(\hat{\boldsymbol{x}})-f_\theta(\boldsymbol{x})\right\|_2^2$
TRADES(Zhang等,2019b)	$CE\left(f_\theta(\boldsymbol{x}),y\right)+\lambda\times KL(f_\theta(\hat{\boldsymbol{x}})\|f_\theta(\boldsymbol{x}))$
GAT(Sriramanan等,2020)	$CE\left(f_\theta(\boldsymbol{x}),y\right)+\lambda\times\left\|f_\theta(\hat{\boldsymbol{x}})-f_\theta(\boldsymbol{x})\right\|_2^2$
MMA(Ding等,2020)	$CE\left(f_\theta(\hat{\boldsymbol{x}}),y\right)\times 1\left(f_\theta(\boldsymbol{x})=y\right)+CE\left(f_\theta(\boldsymbol{x}),y\right)\times 1\left(f_\theta(\boldsymbol{x})\neq y\right)$
MART(Wang等,2019)	$BCE\left(f_\theta(\hat{\boldsymbol{x}}),y\right)+\lambda\times KL(f_\theta(\hat{\boldsymbol{x}})\|f_\theta(\boldsymbol{x}))\times\left(1-P_y(\boldsymbol{x},\boldsymbol{\theta})\right)$
PCL(Mustafa等,2019)	$CE\left(f_\theta(\boldsymbol{x}),y\right)+\sum_i\left\{\left\|f_i-w_{y_i}^c\right\|_2-\frac{1}{k-1}\sum_{j\neq y_i}\left(\left\|f_i-w_j^c\right\|_2+\left\|w_{y_i}^c-w_j^c\right\|_2\right)\right\}$
CAS-AT(Bai等,2022)	$CE\left(f_\theta(\hat{\boldsymbol{x}}),y\right)+\frac{\beta}{S}\times\sum_{s=1}^{S}CE(f_\theta^s(\hat{\boldsymbol{x}},M^s),y)$

3.2　增强模型

增强优化对抗训练过程的防御方法可以分为3个方面:

1)引入即插即用优化模块。不需改变网络原始结构,具有轻量级且有效的优点,对于提高模型鲁棒性有十分重要的意义。例如Pang等人(2020b)设计了超球嵌入(hypersphere embedding,HE)机制来扩充AT框架,具体来说,该HE模块包括3种典型的操作:特征归一化(feature normalization,FN)、权重归一化(weight normalization,WN)和角度余量(angular margins,AM),其中角度余量通过归一化网络逻辑(logits)输出层的特征和softmax层中的权重。

Mustafa等人(2019)通过在网络不同深度添加辅助分类器,同时添加了损失正则项,包含一个简单的最大分离约束,如图6所示,加强了模型的类间分离性,可以使攻击者的任务变得困难。

与之相似的是,Bai等人(2022)同样在网络中间层部分增加辅助分类器,其损失作为标准对抗训练目标函数的正则项,可以使网络自适应学习不同通道对类别预测的重要性,抑制不重要的通道,从而提高模型的鲁棒性。Xu等人(2020)提出了自适应网络,引入了以输入为条件的归一化模块,该模块允许网络针对不同的样本"调整"自身。此外,前文提到的对网络中间层添加辅助分类器(由全局平均池化层与全连接层组成)(Bai等,2022)也属于即插即用模块。

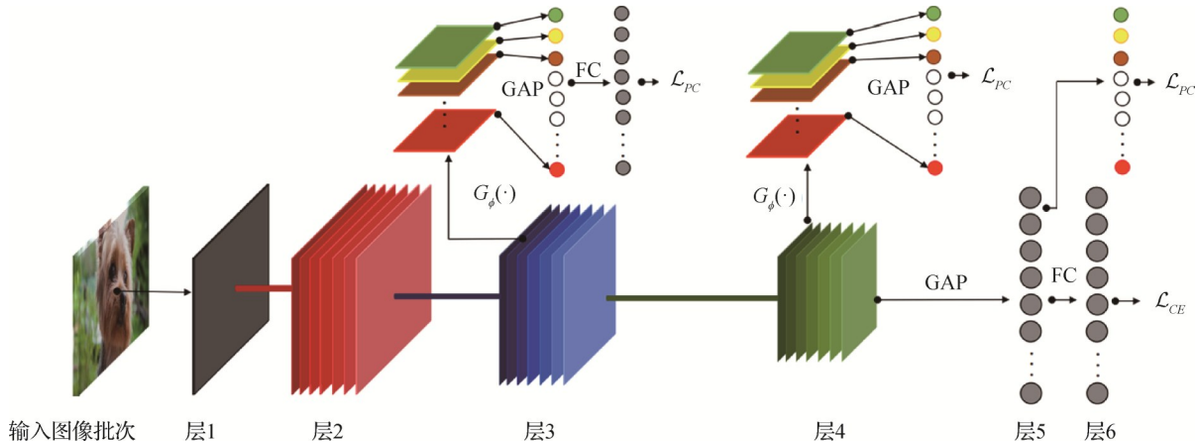


图 6　为模型引入辅助分类器(Mustafa 等,2019)

Fig. 6　Introduce an auxiliary classifier to the model(Mustafa et al. , 2019)

2)修改模型内部结构。Xie 等人(2020)改进了网络 BN 层,提出 Dual-BN,使原始样本与对抗样本分开利用不同的 BN,如图 7 所示。而 Singla 等人(2021)旨在研究不同的激活函数与模型鲁棒性的关系;此外,除了在输入上添加扰动,Wu 等人(2020a)还在模型权重上添加扰动。除了上述方法,Singh 等人(2019)提出鲁棒的子网络方法,即首先微调网络前几层生成的扰动,然后对网络剩余层数进行对抗训练。

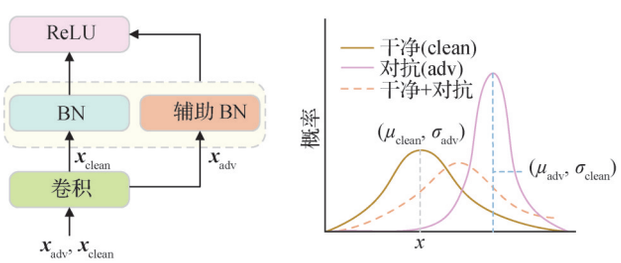


图 7　Dual-BN(Xie 等,2020)

Fig. 7　Dual-BN(Xie et al. , 2020)

3)添加预训练模型或使用双模型训练。研究发现,即使是仅使用干净样本进行预训练,也可以提高模型的鲁棒性,并且使用对抗预训练的模型效果更好(Hendrycks 等 , 2019; Xiao 和 Zheng, 2020; Chen 等,2020),如图 8 所示,Xiao 和 Zheng(2020)使用了轻量级网络预训练输入的干净样本,之后利用此网络生成对抗样本,最终输入到目标网络进行训练与测试。除此之外,Cui 等人(2021)提出双模型训练,同时训练自然模型与目标对抗模型。

3.3　参数自适应

在基于梯度生成对抗样本时,参数的设定对模型鲁棒性的提升有着重要影响,如使用 PGD 攻击

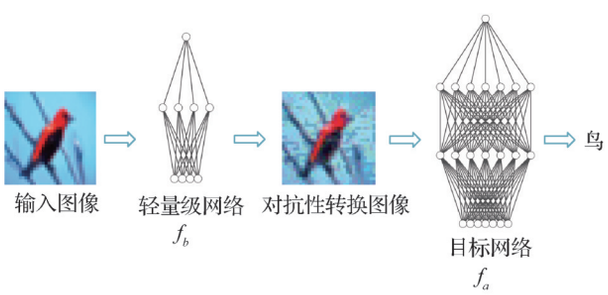


图 8　使用轻量级对抗预训练模型(Xiao 和 Zheng,2020)

Fig. 8　Using lightweight adversarial pretrained models (Xiao and Zheng, 2020)

时,其迭代次数 K、攻击步长 α 与扰动约束 ϵ 都需要人为设置。下列方法针对在设置此类参数时,采用一系列自适应方法,能够更好地提升模型鲁棒性。

Cheng 等人(2020)提出自适应扰动约束 ϵ,根据每个训练样本与决策边界的距离来分配相应的 ϵ,即自适应寻找能让模型错误分类的最小的 ϵ,具体来说,初始化扰动约束为 0,每次迭代后增加一个具体的常数,直到模型产生误判即停止。此外,还提出了自适应标签平滑方法,以反映每个样本上的不同扰动容限;Xiong 和 Hsieh(2020)采用 PGD 攻击方式生成对抗样本,对原始的固定攻击步长进行改进,采用了回溯线搜索(backtracking line search,BLS)方法自适应步长;Croce 和 Hein(2020a)同样提出自适应步长,具体来说,设置初始步长 $\eta^{(0)} = 2\epsilon$,设定检查点 $w_0 = 0,\ w_1, \cdots, w_n$,根据以下两个条件判断是否需要使当前步长减半,具体为

$$\sum_{i=w_j-1}^{w_j-1} 1_{f\left(\boldsymbol{x}^{(i+1)}\right)>f\left(\boldsymbol{x}^{(i)}\right)} < \rho\left(w_j - w_{j-1}\right) \tag{18}$$

$$\eta^{(w_{j-1})}=\eta^{(w_j)},\ f_{\max}^{(w_{j-1})}=f_{\max}^{(w_j)} \tag{19}$$

式中，$f_{\max}^{(k)}$是在前k次迭代中找到的最高目标值。如果其中一个条件为真，则迭代$k=w_j$的步长减半，并且对于每个$k=w_j,\cdots,w_{j+1}$，都有$\eta^{(k)}\eta^{(w_j)}/2$。

受学习率预热的启发，Liu等人(2020)提出了自适应扰动约束ϵ，同样采用预热的方法，具体来说，定义了一个余弦调度器$\epsilon_{\cos}$和一个线性调度器ϵ_{lin}，由$\epsilon_{\max}$和$\epsilon_{\min}$参数化，具体为

$$\begin{cases}\epsilon_{\cos}(d)=\dfrac{1}{2}\left(1-\cos\dfrac{d}{D}\pi\right)(\epsilon_{\max}-\epsilon_{\min})+\epsilon_{\max}\\ \epsilon_{\text{lin}}(d)=(\epsilon_{\max}-\epsilon_{\min})\dfrac{d}{D}+\epsilon_{\min}\end{cases} \tag{20}$$

把$\epsilon_{\cos}(d)$和$\epsilon_{\text{lin}}(d)$裁剪到0和ϵ_{target}(扰动约束的目标值)，如果$\epsilon_{\min}\leqslant 0$且$\epsilon_{\max}>\epsilon_{\text{target}}$，$\epsilon$的值从0逐渐增加到$\epsilon_{\text{target}}$，然后保持不变。

Ye等人(2021)为减少与对抗训练相关的开销，采用了退火机制，由于神经网络在训练初始阶段专注于学习特征，这可能不需要准确的对抗样本。因此，在训练开始时设置大步长α与迭代次数K进行内部最大化，然后逐渐增加K和减少α以提高内部最大化解的质量，第t次迭代产生的退火数量K_t和攻击步长α_t具体为

$$K_t=K_{\min}+(K_{\max}-K_{\min})\times\frac{t}{T} \tag{21}$$

$$\alpha_t=\frac{\tau}{K_t} \tag{22}$$

式中，$K_{\max}$和$K_{\min}$分别为对抗扰动的退火数量上限与下限，τ为某一常数。

此外，对于一些经典的对抗训练，如TRADES(Zhang等，2019b)、MART(misclassification aware adversarial training)(Wang等，2020)等，其目标函数中的参数λ组合了两部分准确度，当调节λ时，需要重新训练模型。为了避免这种繁重的过程，Wang等人(2020)提出了在推理阶段调节λ，使模型不需要因为参数的改变而反复训练。

3.4 早期停止

与标准训练不同，对抗训练会产生鲁棒过拟合现象，即过度适应对抗性强的训练会导致更差的测试集性能。Rice等人(2020)对对抗训练中的过拟合现象进行了全面的研究，发现鲁棒过拟合现象是普遍存在的。如图9所示，训练在初始阶段正常进行，但在学习率衰减之后，测试误差会短暂降低，随着训练进行，训练误差会持续降低，但是测试误差会增加，也就是说，训练结束后的误差不是最佳的，即训练结束后的模型鲁棒性不是最佳鲁棒性。

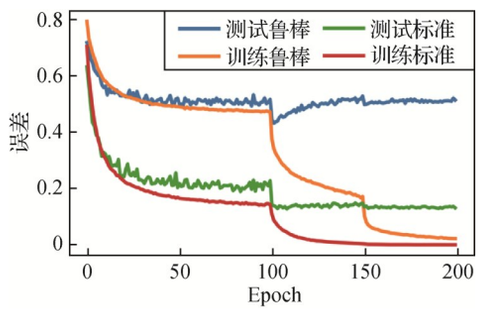


图9 鲁棒训练模型的学习曲线(Rice等，2020)

Fig. 9 Learning curves for robustly trained models (Rice et al., 2020)

针对鲁棒过拟合现象，采用早期停止策略是有必要的，而早期停止可以分为学习率早期停止和攻击强度早期停止(Pang等，2021)，如Zhang等人(2019b)在代码中将学习率设置为在75 epoch处衰减，并且训练在76 epoch处停止；Zhang等人(2020)设置了固定的攻击迭代次数$\tau(\tau<K)$，K为最大迭代次数，旨在模型迭代了τ次生成对抗样本时，就停止训练；Sitawarin等人(2021)定义了样本到决策边界的距离度量，一旦输入数据通过决策边界到达错误类别一侧，就停止生成对抗样本。如图10所示，灰色十字是最接近理想对抗样本点(蓝点)的对抗样本，但计算起来非常昂贵；红色十字表示标准对抗训练生成的对抗样本。对于早期停止方法，一旦越过决策边界(蓝色十字)，该过程就会停止，近似于灰色十字。但是标准对抗训练会继续更新，直到达到指定的最大步长(红色箭头和十字)。

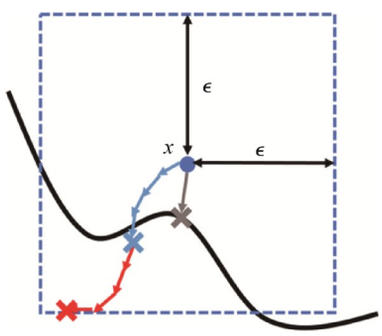


图10 早期停止方法(Sitawarin等，2021)

Fig. 10 Early-stop mothod (Sitawarin et al., 2021)

3.5 半监督或无监督扩展训练

理论证明，对抗训练需要使用比标准训练大得多的数据集，而这需要巨大的成本。因此，有研究希望仅通过增加无标注的数据来提升模型的对抗鲁棒性，即使用半监督或无监督扩展训练，如 Carmon 等人（2019）使用半监督自训练的方式提升模型鲁棒性。Zhai 等人（2019）同样采用大量无标注数据，提出 PASS 算法来提高模型的对抗鲁棒泛化性；上文中提到的 MART 等方法同样可以推广到半监督训练，而 Uesato 等人（2019）则使用无监督训练与监督训练加权结合的方法。

具体来说，通过设计不同的目标函数，作者提出 3 种策略：1）Online Targets 的无监督对抗训练（unsupervised adversarial training with online targets，UAT-OT）；2）无监督的固定目标对抗训练（unsupervised adversarial training with fixed targets，UAT-FT）；3）二者结合训练（UAT++）。Zhang 和 Wang（2019）使用了无监督训练方式，将注意力放在样本间结构上，采用最优传输（optimal transport，OT）距离衡量原始样本与干净样本的距离。

3.6 加速对抗训练

为了提升标准对抗训练的速度，Zhang 等人（2019a）提出了一种能够减少计算正反向传播次数的新方法；Shafahi 等人（2019）提出同步更新扰动和模型参数，连续 m 次在同一小批次（mini-batch）上训练，并在前一阶段的训练结果上继续训练；Zheng 等人（2020）通过逐 epochs 累积对抗扰动，以更少的迭代次数生成相似（甚至更强）的对抗样本。与此同时，有研究旨在利用单步对抗训练来加速，即对 FGSM 进行优化来生成对抗样本。但单步对抗训练会产生灾难性过拟合（catastroptic overfitting，CO）现象（Kim 等，2021），原因在于如果 FGSM 的攻击步长过大，模型会产生扭曲的决策边界（图 11），导致 CO 现象的产生，因此，Wong 等人（2020）提出了随机初始化的 FGSM 内部最大化攻击方式，在一定程度上缓解了 CO 现象；Andriushchenko 等人（2020）认为，当攻击步长较大时，随机初始化依旧对 CO 现象无效，在此基础上，他们提出了梯度对齐（GradAlign）方法，即在点 x 和围绕 x 的 l_∞-*ball* 内的随机扰动点 $x+\eta$ 处的梯度之间最大化梯度对齐。Vivek 和 Babu（2020）旨在通过在网络层增加 dropout 层来减轻 CO 现象，与传统的仅在全连接层与 ReLU（rectified linear unit）后增加 dropout 层（typical setting）不同，作者还在模型的每个非线性层之后引入 dropout 层（proposed setting）；Kim 等人（2021）认为，单步攻击的主要问题是内部最大化的线性近似的失败，如图 11 所示，因此，应重新考虑适当的攻击步长。

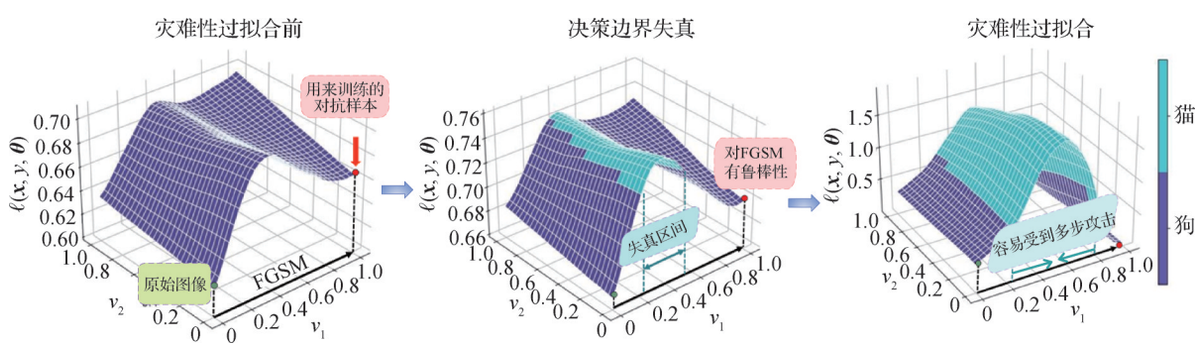


图 11　正常决策边界转变为扭曲决策边界的过程（Kim 等，2021）

Fig. 11　Process of normal decision boundary turns into distorted decision boundary（Kim et al.，2021）

表 4 为对抗训练加速方法的总结，可以看出，加速方法分为多步训练的加速以及利用单步训练加速。多步训练旨在达到同样的攻击强度时，使用更少的迭代次数；单步训练利用只需要迭代一次的优势加速对抗训练，但需要解决灾难性过拟合问题。

3.7 其他防御方法

将标准对抗训练与其他领域相结合，如 Cai 等人（2018）提出课程对抗训练，旨在将课程学习的思想融合进对抗训练中，使攻击强度逐步提升；同样，Sitawarin 等人（2021）提出了将 softmax 概率差距作为课程学习的难度度量，将概率差距定义为除正确类

表 4　加速对抗训练方法对比

Table 4　Comparison of accelerated adversarial training methods

加速方法	训练方式	算法思想	优缺点
YOPO(Zhang等,2019a)	多步训练	减少计算正反向传播速度	计算效率高,有效加速对抗训练
ATTA(Zheng 等,2020)	多步训练	重复使用来自前一epoch扰动,累积攻击强度	仅使用较少的攻击迭代就可以获得相同强度的对抗样本
Free(Shafahi 等,2019)	多步训练	生成扰动与更新模型参数同时进行	效率极大提升,但对FGSM单步攻击不鲁棒
Fast(Wong等,2020)	单步训练	在进行FGSM之前添加初始噪声	不需迭代多次生成扰动,但容易灾难性过拟合
GradAlign(Andriushchenko,2020)	单步训练	相邻两点最大化梯度对齐	加速的同时有效预防灾难性过拟合,但对大型数据集无效
Stable-single AT(Kim等,2021)	单步训练	使用较小扰动生成更强的对抗样本	加速的同时有效避免灾难性过拟合
SADS(Vivek 和 Babu,2020)	单步训练	在模型的每个非线性层之后引入dropout layer	有效避免灾难性过拟合

别之外的任何类别中的最大softmax概率与正确类别的softmax概率之差,旨在最小化其概率差距。

钱申诚等人(2019)希望使对抗样本远离错误类别,更接近真实类别,使用度量学习中的三元组损失(triple loss)函数,向模型添加了额外的约束。如图12所示,通过设置三元组损失中的至少一个元素为对抗样本,其他为原始干净图像,模型使用交叉熵损失与三元组损失交替训练。

图12　度量学习的三元组损失(Mao等,2019)

Fig. 12　Triple loss for metric learning(Mao et al., 2019)

Song等人(2019)考虑到原始样本与对抗样本的分布之间存在很大的领域差距,通过将原始样本与对抗样本分别划分为两个域,并将无监督和有监督的领域适应引入对抗训练中,以最小化原始样本和对抗样本分布之间的差距并增加它们之间的相似性。

Bashivan等人(2022)同样引入领域适应,在网络的逻辑输出层添加领域判别器,如图13所示,将训练数据区分为原始样本和对抗样本。Jia等人(2022)将强化学习思想引入对抗训练,构建了目标网络与策略网络结合的框架,如图14所示,前者使用对抗样本训练提高模型鲁棒性,后者通过学习自动产生依赖于样本的攻击策略,进而指导自动编码器生成器网络生成对抗样本。Dong等人(2020)试

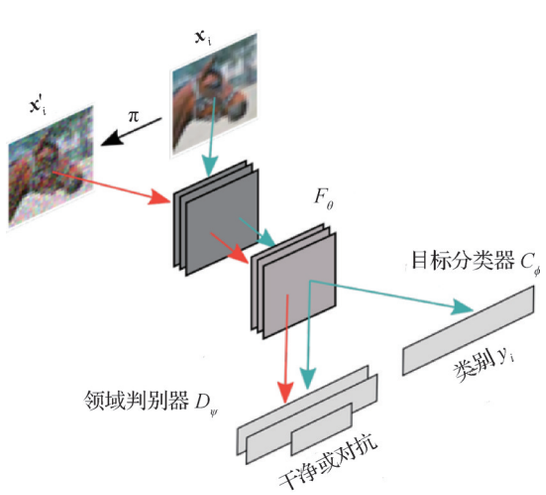


图13　为模型引入领域判别器(Bashivan等,2022)

Fig. 13　Introduce the domain discriminator to the model (Bashivan et al., 2022)

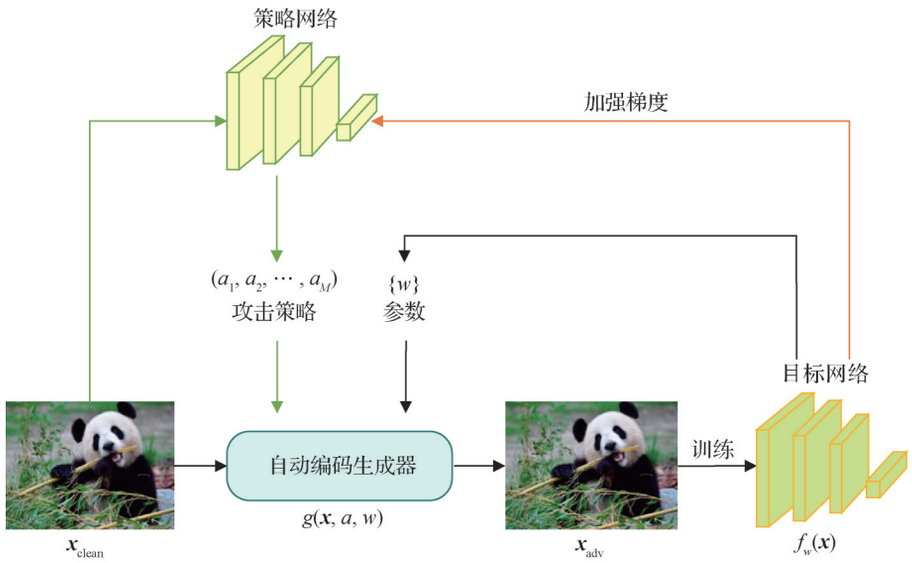


图 14　目标网络结合策略网络框架(Jia 等,2022)

Fig. 14　Target network combining policy network framework(Jia et al. , 2022)

图捕捉每个输入周围的对抗性扰动的分布,而不是像传统对抗训练一样寻找局部最具对抗性的点。

4　对抗训练评估

作为最有效的防御手段之一,对抗训练的评估必不可少,本节从两个方面介绍对抗训练的评估。

4.1　常用的数据集

对抗训练最常用的数据集是 CIFAR-10(Canadian Institute for Advanced Research)、CIFAR-100、MNIST(Modified National Institute of Standards and Technology)以及 SVHN(street view house numbers),由于对抗训练在大型数据集(如 ImageNet)上的效果不尽人意,有些研究使用了 Tiny-ImageNet 数据集来验证其对抗训练的效果。图 15、图 16 和表 5 总结了上述几种数据集的可视化与其基本信息。为了清楚起见,图 15 中所有数据集的图像设置成同一大小,图 16 中的 SVHN 数据集为原始图像大小。

4.1.1　CIFAR-10

CIFAR-10 是一个用于识别普适物体的小型彩色图像数据集,共 10 个类别,如飞机、鸟、汽车、猫等。每幅图像的尺寸为 32 × 32 像素,每个类别有

图 15　数据集总结

Fig. 15　Summary of the datasets ((a) CIFAR-10 dataset;(b) CIFAR-100 dataset;(c) MNIST dataset;(d) Tiny-ImageNet)

图16　SVHN数据集

Fig. 16　SVHN dataset

表5　常用数据集

Table 5　Commonly used datasets

常用数据集	性质	类别数	训练图像/幅	测试图像/幅	图像尺寸(预处理后)/像素
CIFAR-10	小型普适物体彩色图像数据集	10	50 000	10 000	32×32
CIFAR-100	小型普适物体彩色图像数据集	100	50 000	10 000	32×32
MNIST	手写字符灰度图像数据集	10	60 000	10 000	28×28
SVHN	彩色街道门牌号码数据集	10	73 257	26 032	32×32
Tiny-ImageNet	ImageNet数据集的子集	200	100 000	10 000	64×64

6 000幅图像,共有50 000幅训练图像与10 000幅测试图像,在训练时,该数据集将训练图像分为5个批次,每个批次有10 000幅图像,而测试图像的每个批次包含了每个类别的1 000幅随机选择的图像。

4.1.2　CIFAR-100

与CIFAR-10类似,CIFAR-100同样是小型彩色图像数据集,每幅图像尺寸为32×32像素。不同点在于CIFAR-100有20个超类,且被进一步区分为100个类别(例如,鱼为超类;水族馆的鱼、比目鱼、射线、鲨鱼、鳟鱼为类别),每个类别有600幅图像,且分为500幅训练图像与100幅测试图像。每幅图像含有两个标签,其一为100个类别所在的“精细标签”,其二为20个超类所在的“粗糙”标签。

4.1.3　MNIST

MNIST是传统的灰度手写字符数据集,由人口普查局的工作人员(SD-3)与大学生(SD-1)记录的共250种不同笔迹组成,两种笔迹同样包含30 000幅训练图像和5 000幅测试图像,该数据集共60 000幅训练图像和10 000幅测试图像。每幅图像的尺寸为28×28像素,且经过了归一化预处理。

4.1.4　SVHN

与MNIST类似,SVHN同样用于识别字符,但包含了更多数量级的标记数据。SVHN是由谷歌街景彩色图像中的门牌号码组成的数据集,共73 257幅训练图像,26 032幅测试图像以及531 131幅额外的、难度稍低的训练图像。经过预处理后的SVHN数据集图像为固定的32×32像素尺寸。

4.1.5　Tiny-ImageNet

ImageNet是目前世界上最大的图像识别数据集,可以用于目标分类、目标检测等各大领域。而Tiny-ImageNet是ImageNet的子集,共200个类别,每个类有500幅训练图像,50幅验证图像和50幅测试图像,每幅图像的尺寸为64×64像素。

4.2　常用攻击方法

对抗训练的评估方式可以分为白盒攻击与黑盒攻击,白盒攻击是指攻击者可以获得目标模型的一切信息,包括内部结构、训练参数和防御方法等;而黑盒攻击对模型一无所知,只能通过输入输出与模型进行交互。其中,白盒攻击又分为基于梯度的

FGSM与PGD-K攻击，K为攻击迭代次数。在PGD-K中，K通常设为20、50、100等；基于超平面分类的DeepFool（fool deep neural network）攻击（Moosavi-Dezfooli等，2016）；基于优化的C&W（Carlini and Wagner）攻击（Carlini和Wagner，2017）等。黑盒攻击可以分为基于近似梯度的攻击，如ZOO（zeroth order optimization）攻击（Chen等，2017）、使用同步扰动梯度近似的多元随机近似（multivariate stochastic approximation using a simultaneous perturbation gradient approximation，SPSA）攻击（Uesato等，2018）以及使用替代模型（Papernot等，2017）进行攻击评估。基于替代模型的黑盒攻击较为普遍，其主要思想是先训练一个与目标模型具有相似决策边界的替代模型，之后对替代模型进行白盒攻击得到对抗样本，再利用其迁移性实现对目标模型的攻击。此外，Dolatabadi等人（2020）将流模型引入对抗攻击，能够生成分布上与干净数据类似的对抗样本，因而具有较强的攻击性。表6总结了各种评估攻击方法以及对应的攻击强度。

表6　评估对抗训练的攻击方法比较

Table 6　Comparison of adversarial attack methods for evaluating adversarial training

对抗攻击方法	攻击方式	特点	攻击强度
FGSM（Goodfellow等，2015）	白盒攻击	基于梯度的攻击	**
PGD-K（Mardy等，2017）	白盒攻击	基于梯度的攻击	****
C&W（Carlini和Wagner，2017）	白盒攻击	基于优化的攻击	*****
DeepFool（Moosavi-Dezfooli等，2016）	白盒攻击	基于超平面分类的攻击	****
A-PGD（Croce和Hein等，2020b）	白盒攻击	基于梯度的攻击	****
ZOO（Chen等，2017）	黑盒攻击	基于近似梯度的攻击	-
SPSA（Uesato等，2018）	黑盒攻击	基于近似梯度的攻击	-
NATTACK（Li等，2019）	黑盒攻击	基于概率密度估计的攻击	-
Advflow（Dolatabadi等，2020）	黑盒攻击	基于标准化流的攻击	-
FAB（Croce和Hein，2020a）	黑盒攻击	基于快速边界的攻击	-
SquareAttack（Andriushchenko等，2020）	黑盒攻击	基于分数的攻击	-
AutoAttack（Croce和Hein，2020a）	组合攻击	基于两种模式下的A-PGD、FAB、Square Attack的组合攻击	******

注：攻击强度一列仅对白盒攻击（含）比较，“*”的个数代表攻击强度的大小。

另外，针对攻击是否包含特定目标，Croce和Hein（2020b）提出A-PGD攻击，将其两种模式（目标攻击与无目标攻击）与FAB（fast adaptive boundary）攻击（Croce和Hein，2020a）、SquareAttack（Andriushchenko等，2020）进行整合，得到相对复杂但攻击性强的AutoAttack攻击（Croce和Hein，2020a）。

5　结　语

由于深度神经网络脆弱、易受攻击的安全性问题，对抗训练这一典型的防御技术取得了重要进展，并在对抗样本生成、基于对抗样本的模型防御训练两大关键领域涌现出许多优秀的技术。本文不仅详细梳理了传统对抗训练框架下的典型对抗训练方法和关键技术，而且回顾了结合课程学习、度量学习等思想的新型对抗训练方法。

通过本文的梳理不难发现，由于对对抗攻击的本质成因仍不明确，现有的对抗训练技术仍面临缺乏强大有效的攻击方法来对对抗训练方法性能进行有效评估、难以应对多扰动综合的攻击及效率低等挑战，为此，对抗训练未来的发展趋势为：

1）探究数字空间及物理世界对抗样本的成因。针对数字空间与物理世界，探究导致模型输出错误的干扰类型、强度以及干扰方式等，为对抗样本生成提供可靠的理论指导；

2）设计更高效、强大的对抗攻击方法，一方面通过提升对抗样本的攻击能力，来增强模型的防御性

能,另一方面提升对抗训练的效率;

3)构建能自适应防御多攻击类型的通用对抗训练框架与方法,推动深度学习对抗性安全的发展。

参考文献(References)

Addepalli S, Jain S and Babu R V. 2022. Efficient and effective augmentation strategy for adversarial training [EB/OL]. [2022-8-27]. https://arxiv.org/pdf/2210.15318.pdf

Andriushchenko M, Croce F, Flammarion N and Hein M. 2020. Square attack: a query-efficient black-box adversarial attack via random search//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 484-501 [DOI: 10.1007/978-3-030-58592-1_29]

Andriushchenko M and Flammarion N. 2020. Understanding and improving fast adversarial training//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 16048-16059

Bai Y, Zeng Y Y, Jiang Y, Xia S T, Ma X J and Wang Y S. 2022. Improving adversarial robustness via channel-wise activation suppressing [EB/OL]. [2022-01-16]. https://arxiv.org/pdf/2103.08307.pdf

Bashivan P, Bayat R, Ibrahim A, Ahuja K, Faramarzi M, Laleh T, Richards B A and Rish I. 2022. Adversarial feature desensitization [EB/OL]. [2022-01-04]. https://arxiv.org/pdf/2006.04621.pdf

Brown T B, Mané D, Roy A, Abadi M and Gilmer J. 2018. Adversarial patch. [EB/OL]. [2022-10-27]. https://arxiv.org/pdf/1712.09665.pdf

Cai Q Z, Liu C and Song D. 2018. Curriculum adversarial training//Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden: AAAI Press: 3740-3747

Carlini N and Wagner D. 2017. Towards evaluating the robustness of neural networks//Proceedings of 2017 IEEE Symposium on Security and Privacy (SP). San Jose, USA: IEEE: 39-57 [DOI: 10.1109/SP.2017.49]

Carmon Y, Raghunathan A, Schmidt L, Liang P and Duchi J C. 2019. Unlabeled data improves adversarial robustness//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 11192-11203

Chan A, Tay Y, Ong Y S and Fu J. 2020. Jacobian adversarially regularized networks for robustness [EB/OL]. [2022-10-27]. https://arxiv.org/pdf/1912.10185.pdf

Chen P Y, Zhang H, Sharma Y, Yi J F and Hsieh C J. 2017. ZOO: zeroth order optimization based black-box attacks to deep neural networks without training substitute models//Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. Dallas, USA: ACM: 15-26 [DOI: 10.1145/3128572.3140448]

Chen T L, Liu S J, Chang S Y, Cheng Y, Amini L and Wang Z Y. 2020. Adversarial robustness: from self-supervised pre-training to fine-tuning//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 696-705 [DOI: 10.1109/CVPR42600.2020.00078]

Chen Z H, Jiang H M, Dai B and Zhao T. 2021. Learning to defense by learning to attack [EB/OL]. [2022-10-27]. https://arxiv.org/pdf/1811.01213.pdf

Cheng M H, Lei Q, Chen P Y, Dhillon I and Hsieh C J. 2020. CAT: customized adversarial training for improved robustness [EB/OL]. [2022-10-27]. https://arxiv.org/pdf/2002.06789.pdf

Croce F and Hein M. 2020a. Minimally distorted adversarial examples with a fast adaptive boundary attack//Proceedings of the 37th International Conference on Machine Learning. [s.l.]: JMLR.org: 2196-2205

Croce F and Hein M. 2020b. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks//Proceedings of the 37th International Conference on Machine Learning. [s.l.]: JMLR.org: 2206-2216

Cui J Q, Liu S, Wang L W and Jia J Y. 2021. Learnable boundary guided adversarial training//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 15701-15710 [DOI: 10.1109/ICCV48922.2021.01543]

Ding G W, Sharma Y, Lui K Y C and Huang R T. 2020. Mma training: direct input space margin maximization through adversarial training [EB/OL]. [2022-10-27]. https://arxiv.org/pdf/1812.02637.pdf

Dolatabadi H M, Erfani S and Leckie C. 2020. AdvFlow: inconspicuous black-box adversarial attacks using normalizing flows//Proceedings of the 34th International Conference on Advances in Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 15871-15884

Dong Y P, Deng Z J, Pang T Y, Zhu J and Su H. 2020. Adversarial distributional training for robust deep learning//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 8270-8283

Dong Y P, Liao F Z, Pang T Y, Su H, Zhu J, Hu X L and Li J G. 2018. Boosting adversarial attacks with momentum//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 9185-9193 [DOI: 10.1109/CVPR.2018.00957]

Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y. 2014. Generative adversarial networks [EB/OL]. [2022-10-27]. https://arxiv.org/pdf/1406.2661.pdf

Goodfellow I J, Shlens J and Szegedy C. 2015. Explaining and harnessing adversarial examples [EB/OL]. [2022-10-27]. https://arxiv.org/pdf/1412.6572.pdf

He Z, Rakin A S and Fan D L. 2019. Parametric noise injection: trainable randomness to improve deep neural network robustness against

adversarial attack//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 588-597 [DOI: 10.1109/CVPR.2019.00068]

Hendrycks D, Lee K and Mazeika M. 2019. Using pre-training can improve model robustness and uncertainty//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR: 2712-2721

Jang Y, Zhao T C, Hong S and Lee H. 2019. Adversarial defense via learning to generate diverse attacks//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 2740-2749 [DOI: 10.1109/ICCV.2019.00283]

Jeddi A, Shafiee M J, Karg M, Scharfenberger C and Wong A. 2020. Learn2Perturb: an end-to-end feature perturbation learning to improve adversarial robustness//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1238-1247 [DOI: 10.1109/CVPR42600.2020.00132]

Jia X J, Zhang Y, Wu B Y, Ma K, Wang J and Cao X C. 2022. LAS-AT: adversarial training with learnable attack strategy//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 13388-13398 [DOI: 10.1109/CVPR52688.2022.01304]

Kannan H, Kurakin A and Goodfellow I. 2018. Adversarial logit pairing [EB/OL]. [2022-10-27]. https://arxiv.org/pdf/1803.06373.pdf

Kariyappa S and Qureshi M K. 2019. Improving adversarial robustness of ensembles with diversity training [EB/OL]. [2022-10-27]. https://arxiv.org/pdf/1901.09981.pdf

Kim H, Lee W and Lee J. 2021. Understanding catastrophic overfitting in single-step adversarial training. Proceedings of the AAAI Conference on Artificial Intelligence, 35(9): 8119-8127 [DOI: 10.1609/aaai.v35i9.16989]

Kong R, Cai J C and Huang G. 2022. Defense to adversarial attack with generative adversarial network. Acta Automatica Sinica: 1-21 (孔锐, 蔡佳纯, 黄钢. 2022. 基于生成对抗网络的对抗攻击防御模型. 自动化学报: 1-21) [DOI: 10.16383/j.aas.2020.c200033]

Kurakin A, Goodfellow I and Bengio S. 2017. Adversarial examples in the physical world [EB/OL]. [2022-10-27]. https://arxiv.org/pdf/1607.02533.pdf

Lee S, Lee H and Yoon S. 2020. Adversarial vertex mixup: toward better adversarially robust generalization//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 269-278 [DOI: 10.1109/CVPR42600.2020.00035]

Li Q, Lin C H, Yang Y L, Shen C and Fang L M. 2022. Adversarial attacks and defenses against deep learning under the cloud-edge-terminal scenes. Journal of Computer Research and Development, 59(10): 2109-2129 (李前, 蔺琛皓, 杨雨龙, 沈超, 方黎明. 2022. 云边端全场景下深度学习模型对抗攻击和防御. 计算机研究与发展, 59(10): 2109-2129) [DOI: 10.7544/issn1000-1239.20220665]

Li Y D, Li L J, Wang L Q, Zhang T and Gong B Q. 2019. NATTACK: learning the distributions of adversarial examples for an improved black-box attack on deep neural networks [EB/OL]. [2022-10-27]. https://arxiv.org/pdf/1905.00441.pdf

Lin J D, Song C B, He K, Wang L W and Hopcroft J E. 2020. Nesterov accelerated gradient and scale invariance for adversarial attacks [EB/OL]. [2022-10-27]. https://arxiv.org/pdf/1908.06281.pdf

Liu C, Salzmann M, Lin T, Tomioka R and Süsstrunk S. 2020. On the loss landscape of adversarial training: identifying challenges and how to overcome them//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 21476-21487

Liu X M, Xie L H, Wang Y P and Li X R. 2020. Adversarial attacks and defenses in deep learning. Chinese Journal of Network and Information Security, 6(5): 36-53 (刘西蒙, 谢乐辉, 王耀鹏, 李旭如. 2020. 深度学习中的对抗攻击与防御. 网络与信息安全学报, 6(5): 36-53) [DOI: 10.11959/j.issn.2096-109x.2020071]

Liu Y P, Chen X Y, Liu C and Song D. 2017. Delving into transferable adversarial examples and black-box attacks [EB/OL]. [2022-10-27]. https://arxiv.org/pdf/1611.02770.pdf

Madry A, Makelov A, Schmidt L, Tsipras D and Vladu A. 2017. Towards deep learning models resistant to adversarial attacks [EB/OL]. [2022-10-27]. https://arxiv.org/pdf/1706.06083.pdf

Mao C Z, Zhong Z Y, Yang J F, Vondrick C and Ray B. 2019. Metric learning for adversarial robustness//Proceedings of the 33rd Conference on Neural Information Processing Systems. Virtual: NIPS: #32.

Moosavi-Dezfooli S M, Fawzi A and Frossard P. 2016. DeepFool: a simple and accurate method to fool deep neural networks//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 2574-2582 [DOI: 10.1109/CVPR.2016.282]

Mustafa A, Khan S, Hayat M, Goecke R, Shen J B and Shao L. 2019. Adversarial defense by restricting the hidden space of deep neural networks//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 3384-3393 [DOI: 10.1109/ICCV.2019.00348]

Najafi A, Maeda S, Koyama M and Miyato T. 2019. Robustness to adversarial perturbations in learning from incomplete data//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 5541-5551

Pang T Y, Xu K, Du C, Chen N and Zhu J. 2019. Improving adversarial robustness via promoting ensemble diversity//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR, 2019: 4970-4979

Pang T Y, Xu K and Zhu J. 2020a. Mixup inference: better exploiting mixup to defend adversarial attacks [EB/OL]. [2022-10-27].

https://arxiv.org/pdf/1909.11515.pdf

Pang T Y, Yang X, Dong Y P, Su H and Zhu J. 2021. Bag of tricks for adversarial training [EB/OL]. [2022-10-27]. https://arxiv.org/pdf/2010.00467.pdf

Pang T Y, Yang X, Dong Y P, Xu K, Zhu J and Su H. 2020b. Boosting adversarial training with hypersphere embedding//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 7779-7792

Papernot N, McDaniel P, Goodfellow I, Jha S, Celik Z B and Swami A. 2017. Practical black-box attacks against machine learning//Proceedings of 2017 ACM on Asia Conference on Computer and Communications Security. Abu Dhabi, United Arab Emirates: ACM: 506-519 [DOI: 10.1145/3052973.3053009]

Qian S C, Wen Y H, Ma Y F and Mao X W. 2022. Adversial sample attack and defense methods based on deep neural networks. Cyberspace Security, 13(5): 77-86 (钱申诚, 文宇恒, 马耀飞, 毛鑫唯. 2022. 基于深度神经网络的对抗样本攻击与防御方法研究. 网络空间安全, 13(5): 77-86)

Qin C L, Martens J, Gowal S, Krishnan D, Dvijotham K, Fawzi A, De S, Stanforth R and Kohli P. 2019. Adversarial robustness through local linearization//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 13842-13853

Rice L, Wong E and Kolter J Z. 2020. Overfitting in adversarially robust deep learning//Proceedings of the 37th International Conference on Machine Learning. [s.l.]: JMLR.org: 8093-8104

Salman H, Ilyas A, Engstrom L, Vemprala S, Madry A and Kapoor A. 2021. Unadversarial examples: designing objects for robust vision//Proceedings of the 35th International Conference on Neural Information Processing Systems. [s. l.]: Curran Associates, Inc.: 15270-15284

Shafahi A, Najibi M, Ghiasi A, Xu Z, Dickerson J, Studer C, Davis L S, Taylor G and Goldstein T. 2019. Adversarial training for free!//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 3358-3369

Sharif M, Bhagavatula S, Bauer L and Reiter M K. 2016. Accessorize to a crime: real and stealthy attacks on state-of-the-art face recognition//Proceedings of 2016 ACM SIGSAC Conference on Computer and Communications Security. Vienna, Austria: ACM: 1528-1540 [DOI: 10.1145/2976749.2978392]

Singh M, Sinha A, Kumari N, Machiraju H, Krishnamurthy B and Balasubramanian V N. 2019. Harnessing the vulnerability of latent layers in adversarially trained models [EB/OL]. [2022-10-27]. https://arxiv.org/pdf/1905.05186.pdf

Singla V, Singla S, Feizi S and Jacobs D. 2021. Low curvature activations reduce overfitting in adversarial training//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 16403-16413 [DOI: 10.1109/ICCV48922.2021.01611]

Sitawarin C, Chakraborty S and Wagner D. 2021. Improving adversarial robustness through progressive hardening [EB/OL]. [2022-10-27]. https://arxiv.org/pdf/2003.09347.pdf

Song C B, He K, Lin J D, Wang L W and Hopcroft J E. 2020. Robust local features for improving the generalization of adversarial training [EB/OL]. [2022-10-27]. https://arxiv.org/pdf/1909.10147.pdf

Song C B, He K, Wang L W and Hopcroft J E. 2019. Improving the generalization of adversarial training with domain adaptation [EB/OL]. [2022-10-27]. https://arxiv.org/pdf/1810.00740.pdf

Sriramanan G, Addepalli S, Baburaj A and Babu R V. 2020. Guided adversarial attack for evaluating and enhancing adversarial defenses//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 20297-20308

Tramèr F and Boneh D. 2019. Adversarial training and robustness for multiple perturbations//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 5866-5876

Tramèr F, Kurakin A, Papernot N, Goodfellow I, Boneh D and McDaniel P. 2020. Ensemble adversarial training: attacks and defenses [2022-10-27]. https://arxiv.org/pdf/1705.07204.pdf

Tsipras D, Santurkar S, Engstrom L, Turner A and Madry A. 2019. Robustness may be at odds with accuracy [EB/OL]. [2022-10-27]. https://arxiv.org/pdf/1805.12152.pdf

Uesato J, Alayrac J B, Huang P S, Stanforth R, Fawzi A and Kohli P. 2019. Are labels required for improving adversarial robustness?//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 12214-12223

Uesato J, O'Donoghue B, Kohli P and Oord A. 2018. Adversarial risk and the dangers of evaluating against weak attacks//Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden: PMLR: 5025-5034

Vivek B S and Babu R V. 2020. Single-step adversarial training with dropout scheduling//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 947-956 [DOI: 10.1109/CVPR42600.2020.00103]

Wan W T, Chen J S and Yang M H. 2020. Adversarial training with bi-directional likelihood regularization for visual classification//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 785-800 [DOI: 10.1007/978-3-030-58586-0_46]

Wang G Q, Wei X X and Yan H Q. 2022a. Improving adversarial transferability with spatial momentum [EB/OL]. [2022-10-27]. https://arxiv.org/pdf/2203.13479.pdf

Wang H T, Chen T L, Gui S P, Hu T K, Liu J and Wang Z Y. 2020. Once-for-all adversarial training: in-situ tradeoff between robustness and accuracy for free//Proceedings of the 34th International

Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 7449-7461

Wang J K, Yin Z X, Hu P F, Liu A S, Tao R S, Qin H T, Liu X L and Tao D C. 2022b. Defensive patches for robust recognition in the physical world//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 2446-2455 [DOI: 10.1109/CVPR52688.2022.00249]

Wang J Y and Zhang H C. 2019. Bilateral adversarial training: towards fast training of more robust models against adversarial attacks//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 6628-6637 [DOI: 10.1109/ICCV.2019.00673]

Wang X S and He K. 2021. Enhancing the transferability of adversarial attacks through variance tuning//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 1924-1933 [DOI: 10.1109/CVPR46437.2021.00196]

Wang Y S, Zou D F, Yi J F, Bailey J, Ma X J and Gu Q Q. 2019. Improving adversarial robustness requires revisiting misclassified examples//Proceedings of 2019 International Conference on Learning Representations. openreview

Wong E, Rice L and Kolter J Z. 2020. Fast is better than free: revisiting adversarial training [EB/OL]. [2020-01-12]. https://arxiv.org/pdf/2001.03994.pdf

Wu D X, Xia S T and Wang Y S. 2020a. Adversarial weight perturbation helps robust generalization//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 2958-2969

Wu W B, Su Y X, Lyu M R and King I. 2021. Improving the transferability of adversarial samples with adversarial transformations//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 9020-9029 [DOI: 10.1109/CVPR46437.2021.00891]

Wu Y T, Liu W, Yu H T and Cao X C. 2022. Adversarial attacks on graph neural network based on local influence analysis model. Journal of Electronics and Information Technology, 44(7): 2576-2583 (吴翼腾, 刘伟, 于洪涛, 操晓春. 2022. 基于局部影响分析模型的图神经网络对抗攻击. 电子与信息学报, 44(7): 2576-2583) [DOI: 10.11999/JEIT210448]

Wu Z X, Lim S N, Davis L S and Goldstein T. 2020b. Making an invisibility cloak: real world adversarial attacks on object detectors//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 1-17 [DOI: 10.1007/978-3-030-58548-8_1]

Xiao C and Zheng C X. 2020. One man's trash is another man's treasure: resisting adversarial examples by adversarial examples//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 408-418 [DOI: 10.1109/CVPR42600.2020.00049]

Xie C H, Tan M X, Gong B Q, Wang J, Yuille A L and Le Q V. 2020. Adversarial examples improve image recognition//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 816-825 [DOI: 10.1109/CVPR42600.2020.00090]

Xie C H, Wang J Y, Zhang Z S, Ren Z and Yuille A. 2018. Mitigating adversarial effects through randomization [EB/OL]. [2022-10-27]. https://arxiv.org/pdf/1711.01991.pdf

Xie C H, Wu Y X, van der Maaten L, Yuille A L and He K M. 2019. Feature denoising for improving adversarial robustness//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 501-509 [DOI: 10.1109/CVPR.2019.00059]

Xiong Y H and Hsieh C J. 2020. Improved adversarial training via learned optimizer//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 85-100 [DOI: 10.1007/978-3-030-58598-3_6]

Xu H, Liu X R, Li Y X, Jain A and Tang J L. 2021. To be robust or to be fair: towards fairness in adversarial training//Proceedings of the 38th International Conference on Machine Learning. [s. l.]: PMLR: 11492-11501

Xu Z, Shafahi A and Goldstein T. 2020. Exploring model robustness with adaptive networks and improved adversarial training [EB/OL]. [2022-10-27]. https://arxiv.org/pdf/2006.00387.pdf

Ye N Y, Li Q X, Zhou X Y and Zhu Z X. 2021. Amata: an annealing mechanism for adversarial training acceleration. Proceedings of the AAAI Conference on Artificial Intelligence, 35 (12): 10691-10699 [DOI: 10.1609/aaai.v35i12.17278]

Yu Y R, Gao X T and Xu C Z. 2021. LAFEAT: piercing through adversarial defenses with latent features//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 5731-5741 [DOI: 10.1109/CVPR46437.2021.00568]

Yu Z F, Yan Q and Zhou Y. 2022. A survey on adversarial machine learning for cyberspace defense. Acta Automatica Sinica, 48(7): 1625-1649 (余正飞, 闫巧, 周鋆. 2022. 面向网络空间防御的对抗机器学习研究综述. 自动化学报, 48(7): 1625-1649) [DOI: 10.16383/j.aas.c210089]

Yuan L, Li X M, Pan Z X, Sun J M and Xiao L. 2022. Review of adversarial examples for object detection. Journal of Image and Graphics, 27(10): 2873-2896 (袁珑, 李秀梅, 潘振雄, 孙军梅, 肖蕾. 2022. 面向目标检测的对抗样本综述. 中国图象图形学报, 27(10): 2873-2896) [DOI: 10.11834/jig.210209]

Zhai R T, Cai T L, He D, Dan C, He K, Hopcroft J and Wang L W. 2019. Adversarially robust generalization just requires more unlabeled data [EB/OL]. [2019-09-26]. https://arxiv.org/pdf/1906.00555.pdf

Zhang D H, Zhang T Y, Lu Y P, Zhu Z X and Dong B. 2019a. You only propagate once: accelerating adversarial training via maximal prin-

ciple//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 227-238

Zhang H C and Wang J Y. 2019. Defense against adversarial attacks using feature scattering-based adversarial training//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 1831-1841

Zhang H C and Xu W. 2020. Adversarial interpolation training: a simple approach for improving model robustness.//Proceedings of 2020 International Conference on Learning Representations. Virtual: ICLR

Zhang H Y, Yu Y D, Jiao J T, Xing E, El Ghaoui L and Jordan M. 2019b. Theoretically principled trade-off between robustness and accuracy//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR: 7472-7482

Zhang J F, Xu X L, Han B, Niu G, Cui L Z, Sugiyama M and Kankanhalli M. 2020. Attacks which do not kill training make adversarial learning stronger//Proceedings of the 37th International Conference on Machine Learning. [s.l.]: JMLR.org, 2020: 11258-11287

Zhang J P, Wu W B, Huang J T, Huang Y Z, Wang W X, Su Y X and Lyu M R. 2022. Improving adversarial transferability via neuron attribution-based attacks//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 14973-14982 [DOI: 10.1109/CVPR52688.2022.01457]

Zhao H, Chang Y K and Wang W J. 2022. Survey of adversarial attacks and defense methods for deep neural networks. Computer Science, 49(S2): #210900163 (赵宏, 常有康, 王伟杰. 2022. 深度神经网络的对抗攻击及防御方法综述. 计算机科学, 49(S2): #210900163) [DOI: 10.11896/jsjkx.210900163]

Zheng H Z, Zhang Z Q, Gu J C, Lee H and Prakash A. 2020. Efficient adversarial training with transferable adversarial examples//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1178-1187 [DOI: 10.1109/CVPR42600.2020.00126]

作者简介

隋晨红,女,副教授,主要研究方向为多模态数据融合、对抗攻击与防御和智能遥感图像处理。
E-mail: sui6662015@ytu.edu.cn
王海鹏,通信作者,男,教授,主要研究方向为多源数据融合、对抗攻击与防御。E-mail: whp5691@163.com
王奥,女,硕士研究生,主要研究方向为计算机视觉、对抗攻击与防御。E-mail: wangao1999@s.ytu.edu.cn
周圣文,男,硕士研究生,主要研究方向为计算机视觉、对抗攻击与防御。E-mail: zhoushengwen@s.ytu.edu.cn
臧安康,男,本科生,主要研究方向为计算机视觉、对抗攻击与防御。E-mail: 202057506310@s.ytu.edu.cn
潘云豪,男,硕士研究生,主要研究方向为计算机视觉、对抗攻击与防御。E-mail: panyunhao@s.ytu.edu.cn
刘颢,男,研究员,主要研究方向为信息融合和偏微分方程。
E-mail: liuhao2020@sjtu.edu.cn