计算机科学与探索
Journal of Frontiers of Computer Science and Technology
1673-9418/2023/17(12)-2827-13
doi: 10.3778/j.issn.1673-9418.2303080

图像对抗样本防御技术研究综述

刘瑞祺,李　虎,王东霞+,赵重阳,李博宇
军事科学院 系统工程研究院 信息系统安全技术重点实验室,北京 100101
+ 通信作者 E-mail: wdxpaper@126.com

摘　要:人工智能的快速发展和广泛应用带来了新的安全性问题,针对深度神经网络的对抗样本生成与防御是其中的热点之一。深度神经网络在图像领域应用最广也最容易被图像对抗样本欺骗,针对图像对抗样本的防御技术研究是提升人工智能应用安全的重要手段。图像对抗样本的存在原因尚无统一解释,但可从不同维度加以观察与理解,从而为提出针对性的防御技术方法提供启示。对当前主流的盲区假说、线性假说、决策边界假说、特征假说等对抗样本存在原因假说,以及各种假说与典型对抗样本生成方法之间的关联关系进行了梳理分析。以此为基础,从基于模型和基于数据两个维度对图像对抗样本防御技术进行了总结归纳,对比分析了不同技术方法的适应场景与优缺点。现有的图像对抗样本防御技术方法大多针对具体的对抗样本生成方法进行防御,尚无统一的防御理论与方法。现实应用中需综合考虑具体的应用场景、潜在的安全风险等,在现有的防御技术方法中进行优化组合配置。后续可从泛化防御理论、防御效果评价、体系化防护策略等方面深化技术研究。

关键词:对抗样本;人工智能安全;对抗防御
文献标志码:A　**中图分类号:**TP181

Survey of Image Adversarial Example Defense Techniques

LIU Ruiqi, LI Hu, WANG Dongxia+, ZHAO Chongyang, LI Boyu
National Key Laboratory of Science and Technology on Information System Security, Institute of System and Engineering, Academy of Military Sciences, Beijing 100101, China

Abstract: The rapid and extensive growth of artificial intelligence introduces new security challenges. The generation and defense of adversarial examples for deep neural networks is one of the hot spots. Deep neural networks are most widely used in the field of images and most easily cheated by image adversarial examples. The research on the defense techniques for image adversarial examples is an important tool to improve the security of AI applications. There is no standard explanation for the existence of image adversarial examples, but it can be observed and understood from different dimensions, which can provide insights for proposing targeted defense approaches. This paper sorts out and analyzes current mainstream hypotheses of the reason for the existence of adversarial examples, such as the blind spot hypothesis, linear hypothesis, decision boundary hypothesis, and feature hypothesis, and the correlations between various hypotheses and typical adversarial example generation methods. Based on this, this paper summarizes the image adversarial example defense techniques in two dimensions, model-based and data-based, and compares and analyzes the adaptation scenarios, advantages and disadvantages of different technical methods. Most of the existing image adversarial example defense techniques are aimed at defending against specific adversarial

基金项目:国家部委科技重点实验室基金。
This work was supported by the Foundation of Key Laboratory of Science and Technology of China Ministries and Commissions.
收稿日期:2023-03-24　**修回日期:**2023-08-21

example generation methods, and there is no universal defense theory and method yet. In the real application, it needs to consider the specific application scenarios, potential security risks and other factors, optimize and combine the configuration in the existing defense methods. Future researchers can deepen their technical research in terms of generalized defense theory, evaluation of defense effectiveness, and systematic protection strategies.

Key words: adversarial examples; artificial intelligence security; adversarial defense

深度神经网络(deep neural network,DNN)助推当前阶段人工智能(artificial intelligence,AI)飞速发展,但其安全性问题也逐渐凸显[1]。如Szegedy等人[2]发现,对测试正常图像添加一个人眼难以察觉的非随机扰动后形成的对抗样本(adversarial example,AE),可以使DNN模型以较高的置信度输出错误结果。对抗样本的存在增强了公众对DNN模型安全性的重视,也为AI的安全应用提出了新挑战。尤其在自动驾驶[3]、医学诊断[4]、安全监控[5]等敏感业务场景下,如何高效检测或消除对抗样本对DNN模型带来的安全威胁十分必要。本文聚焦DNN应用最广的图像领域,围绕图像对抗样本防御技术的发展脉络,分析图像对抗样本存在的各类假说及典型攻击方法,进而对图像对抗样本防御技术进行归纳总结,以期为本领域研究人员提供技术参考。

已有部分研究围绕对抗样本的生成与防御技术方法从不同的角度进行了梳理归纳[6-8],如Wiyatno等人[6]调查了早期的对抗样本攻防技术方法并分析了这些技术方法存在的问题,白祉旭等人[7]从白盒攻击和黑盒攻击角度梳理分析了对抗样本生成技术,张田等人[8]从特征学习、分布统计等方面归纳了对抗样本检测与防御技术。随着新的攻防技术方法的不断迭代进化,亟待对不同技术方法之间的关联性,尤其从对抗样本的不同存在假说角度进行关联分析,进而为对抗样本的防御提供新的技术思路。如快速梯度符号法(fast gradient sign method,FGSM)[9]提出后,通过引入迭代[10]、随机[11]、动量[12]、动态学习[13]等策略产生了系列对抗样本生成方法。同时,对抗样本防御技术也在不断发展,如Papernot等人提出了防御蒸馏方法[14],通过降低模型对于对抗梯度的敏感度实现对FGSM等基于梯度信息的对抗样本防御。此后Carlini等人又针对防御蒸馏方法提出了C&W攻击方法[15]。围绕对抗样本的攻防对抗技术研究呈现快速迭代演化发展的趋势。因此,本文从对抗样本的存在假说入手,分析假说与相关技术的内在联系,归纳总结假说对攻防研究的指导启示,系统梳理代表性对抗样本防御技术的最新迭代演进趋势。图1按时间顺序列举了部分图像对抗样本攻防研究领域的典型技术方法,对技术方法之间的相关关系做了部分标注。

本文组织逻辑架构如图2所示。首先梳理了各种主流的对抗样本存在原因假说及相应的对抗样本生成方法;然后围绕对抗样本的检测和消除对防御技术方法进行了归纳总结;最后总结了图像对抗样本防御技术当前的发展现状,对未来的可能的发展方向进行了展望。

1 对抗样本的存在原因假说与典型生成方法

深度神经网络结构复杂,内部参数空间大,其决策过程具有不可解释性,对其面临的安全威胁进行深入分析也更加困难。学术界普遍认为针对DNN的对抗样本必然存在,但对其存在机理或存在原因尚无统一认识。研究者提出了多种存在原因假设,分别从不同的角度探讨了对抗样本产生的可能原因,并据此提出了部分对抗样本生成和防御的技术方法与意见建议。当前的对抗样本防御方法大部分针对某一种或某一类攻击方法进行针对性防御,普遍泛化性较差。基于不同的对抗样本存在原因假说对当前的攻击与防御方法进行梳理,可提供新的观察视角。

1.1 对抗样本存在原因与假说

1.1.1 盲区假说

DNN的不可解释性导致其决策过程具有非直观特性,其内部结构与训练数据分布的关联并不直观,可能存在一定的盲区,在盲区内的样本会使得DNN出错。为寻找这些盲区,Szegedy等人提出了L-BFGS(limited memory Broden-Fletcher-Goldfarb-Shanno)攻击[2],通过盒约束(box constraint)优化在输入空间中寻找对抗样本,其优化函数为:

$$\min c\|\delta\|_2+\mathcal{L}(x+\delta,t)\ \text{ s.t. }\ x+\delta\in[0,1]^n \tag{1}$$

其中,$x+\delta\in[0,1]^n$以盒约束形式来约束扰动的大小,确保生成合理的图像;常数$c>0$用于调节扰动对优化结果的影响,通常通过线性搜索方式选取最优值,并迭代优化来寻找生成对抗样本的最小扰动。L-BFGS攻击的目标就是找到并利用这些“盲区”,通过

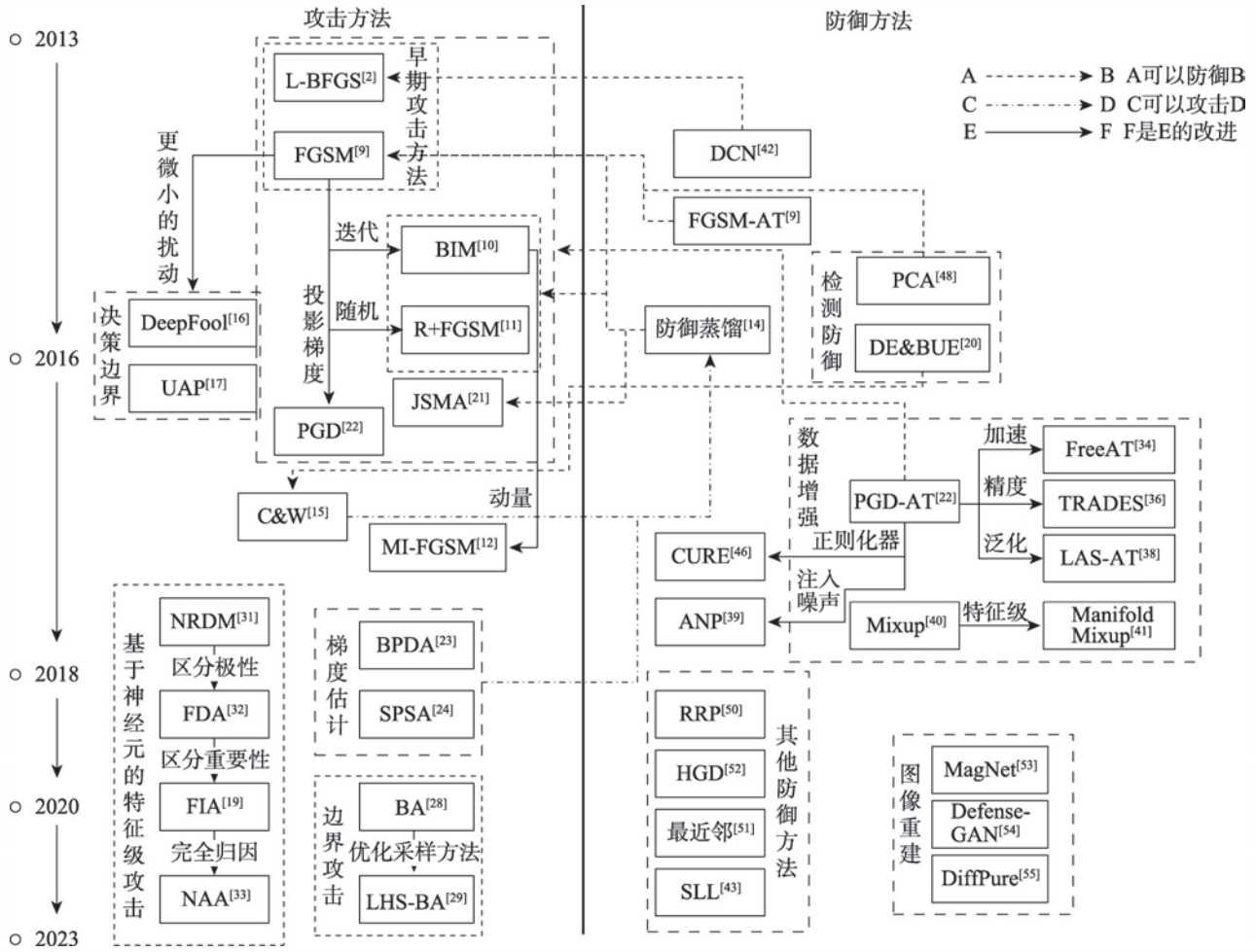

图1 对抗样本攻防技术迭代演进示意

Fig.1 Schematic of development and evolution of attacks and defenses of adversarial examples

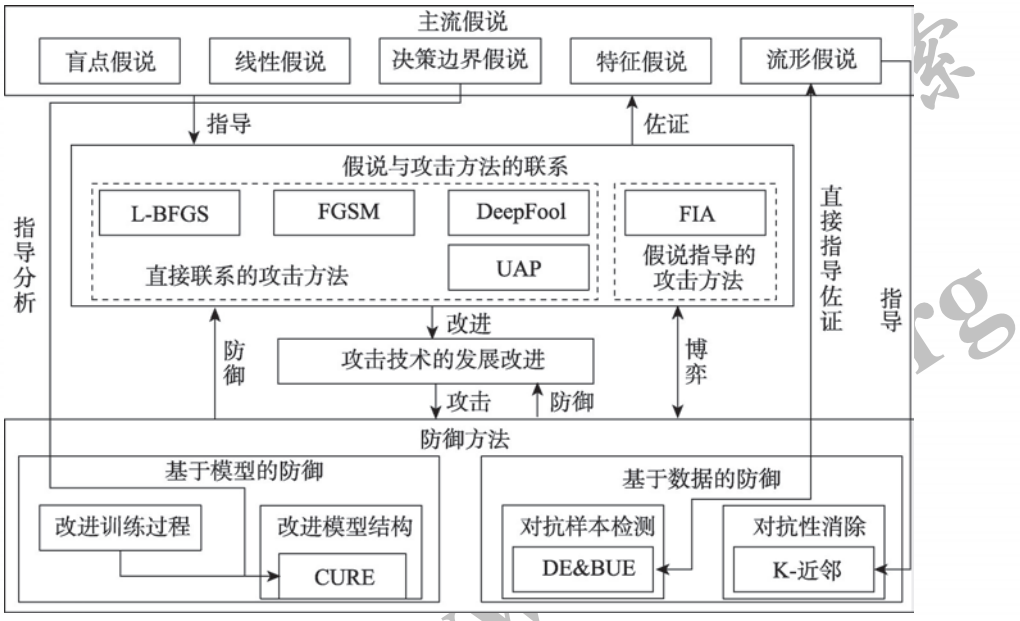

图2 本文组织逻辑架构

Fig.2 Logic structure of this paper

优化输入扰动，将输入样本扰动至“盲区”，改变模型预测结果，从而生成有效的对抗样本。

1.1.2 线性假说

DNN通常存在于高维空间，在整体上具有非线性特性，但针对具体的激活函数或局部区域仍然具有线性特性。当局部添加的扰动不断线性累积，则最终可能形成欺骗DNN的对抗样本。基于线性假说，Goodfellow等人提出了FGSM攻击，通过梯度下降来生成扰动：

$$x' = x + \varepsilon \cdot \text{sign}(\nabla_x \mathcal{L}(x,y)) \tag{2}$$

其中，x 为原始样本；x' 为添加扰动后的对抗样本；∇_x 是损失函数 $\mathcal{L}(x,y)$ 相对于输入 x 的梯度，即 $\mathcal{L}(x,y)$ 对 x 求偏导的结果；sign()是符号函数，将输入转换为 −1和+1两类符号输出；ε 为扰动的步长。FGSM攻击利用了模型的局部线性特性，沿输入空间的梯度方向进行扰动，简单高效，能够快速找到添加扰动的最优方向，但其有效性在一定程度上依赖于扰动步长的大小。

1.1.3 决策边界假说

DNN的分类决策边界通常是一个非线性曲线/面/体，通过在原始样本上添加扰动，其可能会跨过决策边界，进而使得DNN分类错误[16-17]。二分类问题的决策边界简化示意如图3，当在真实类别为 A 的样本上添加扰动后，其跨过DNN模型的决策边界，被错误识别为类别 B 。

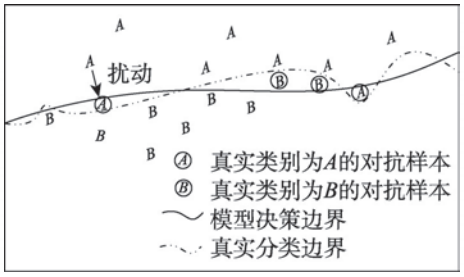

图3 决策边界示意图

Fig.3 Decision boundary schematic

根据决策边界假说，Moosavi-Dezfooli等人提出了针对单个决策边界的DeepFool攻击[16]，通过不断逼近模型的决策边界来寻找最优对抗扰动，可以更准确地量化模型的鲁棒性以及通用对抗扰动（universal adversarial perturbations，UAP）攻击[17]，通过捕捉多个决策边界不同区域之间的相关性，使得添加扰动后的对抗样本能够同时跨过多个决策边界。DeepFool攻击和UAP攻击的目标都是通过迭代计算寻找符合目标的最小扰动，DeepFool通过迭代地使输入越过最近的决策边界，可以生成比FGSM攻击更微小的扰动，UAP攻击对一组训练样本进行迭代，在每次迭代中根据当前模型的决策边界调整通用扰动，可以用于推动多个输入跨越其决策边界。

1.1.4 特征假说

DNN通过对图像特征的逐层提取来学习知识并做出决策，但不同的图像特征在决策过程中起到的作用具有差异。Ilyas等人将不易受对抗扰动影响的特征称为鲁棒性特征，将容易受到对抗扰动影响的特征称为非鲁棒特征[18]。在此基础上，Wang等人提出了特征重要性感知攻击方法（feature importance-aware attack，FIA）[19]，从分类模型的中间层提取特征图并计算评估特征的重要性，实现对重要特征的选择。

1.1.5 其他存在假说

对抗样本的存在原因目前尚无统一解释，除了前述各类假说，还有很多从其他角度提出的假设。如Feinman等人通过数据流形（data manifold）来解释对抗样本，认为训练数据存在于高维空间，但在模型中以低维流形的形式存在，添加扰动后的对抗样本不在低维数据流形上，因而会决策错误[20]。基于流形假设，Feinman等人设计了基于密度估计和贝叶斯不确定性估计的对抗性样本检测方法DE&BUE（density estimation & Bayesian uncertainty estimation）[20]。

深入理解对抗样本为何存在有助于理解各类对抗样本的生成方法之间的关联，也有助于设计更好的对抗样本防御措施。表1梳理总结了几种代表性的对抗样本存在原因假说及其对DNN攻防技术研究的启示。

需要说明的是，当前尚无关于对抗样本存在原因的统一解释，一种对抗样本攻击或防御方法的背后可能蕴含多个假说的思想，且不同的假说之间并非互斥，而是存在部分重叠，需要根据实际深入理解。

1.2 典型的对抗样本生成方法

本文的主旨是对图像对抗样本的防御技术进行梳理分析，但防御本身并非孤立存在，而是与攻击技术迭代演进。本节对典型的对抗样本生成技术方法进行对比分析，为后面的防御技术方法分析提供技术参考。

1.2.1 直接优化目标函数

早期的对抗样本生成方法通常默认DNN没有采取防御措施，但随着大家对DNN安全性的重视，提出了很多防御技术方法。如何生成能够规避已有防御

表1 对抗样本存在原因假说总结

Table 1 Summary of hypothesis of reason for existence of adversarial example

假说	观点	相关技术	对攻击技术的启示	对防御技术的启示
盲区假说	DNN存在固有盲区	L-BFGS[2]	找到神经网络的固有盲点	模糊或减少盲区
线性假说	DNN的局部线性特性容易被利用	FGSM[9]	计算梯度快速找到扰动方向	模糊或减少梯度信息输出
决策边界假说	对抗样本跨越了DNN的决策边界	DeepFool[16]、UAP[17]	优化扰动与决策边界的距离	平滑模型的决策边界
特征假说	非鲁棒特征上添加扰动更容易欺骗DNN	FIA[19]	区分非鲁棒特征以增强扰动添加的针对性	降低模型对局部特征的敏感性
流形假说	偏离低维数据流形的对抗样本容易误导DNN	DE&BUE[20]	优化对抗扰动与数据流形的距离	主动检测远离数据流形的输入样本

措施的对抗样本成为了新的研究方向，如Carlini等人[15]针对防御蒸馏防御技术[14]提出了基于优化的C&W攻击方法，其优化函数为：

$$\min \|\delta\|_p + c \cdot f(x+\delta) \quad \text{s.t.} \quad x+\delta \in [0,1]^n \tag{3}$$

其中，$\|\delta\|_p$可以是L_0、L_2或L_∞范数，用于度量扰动δ的大小；$f(x+\delta)$对添加扰动后的样本进行预测。与L-BFGS攻击[2]一样，C&W攻击[15]也是通过直接优化目标函数生成对抗样本，但其优化算法通常是梯度下降。且C&W攻击[15]的目标函数由两部分组成，一部分显式地约束扰动的大小，另一部分保证模型预测错误，因此生成的对抗样本更难以察觉。

1.2.2 计算导数寻找扰动

对抗样本的生成方法不断更新，针对FGSM攻击虽然计算速度快，但是生成的对抗样本质量较差、扰动大小难以控制等问题，Papernot等人提出了基于雅可比显著图的攻击方法（Jacobian saliency map attack，JSMA），通过计算模型输出对输入的偏导数构建雅可比矩阵，形成描述输出对输入的敏感度显著图，对模型输出影响最显著的部分添加针对性的扰动[21]。JSMA攻击在进行每次迭代时都需要重新计算显著图，因此需要较高的计算成本。

引入迭代、随机、动量、动态策略等可以在原有方法基础上进行改进。BIM（basic iterative method）攻击以较小的步长多次应用FGSM攻击，每步迭代后剪裁中间结果的像素值，使其始终处于原始图像的邻域范围内[10]。R+FGSM攻击（ramdom fast gradient sign method）在FGSM攻击的基础上引入随机起点，以避免陷入局部最大值[11]。PGD（projected gradient descent）攻击可以同时包含随机和迭代的思想，并通过投影控制扰动的大小[22]。MI-FGSM（momentum iterative fast gradient sign method）攻击则是在迭代中积累损失函数梯度方向的矢量以稳定更新方向[12]。DSNGD（dynamically sampled nonlocal gradient descent）攻击将梯度方向计算为优化历史中之前梯度的加权平均值，通过纳入非局部梯度信息，对噪声和非凸损失表面的全局下降方向给出更准确的估计[13]。

FGSM攻击及其系列改进的有效性建立在攻击者可以获取准确梯度信息的基础上，当攻击者无法获取准确的梯度信息时，此类攻击就会失效。Athalye等人进一步细分了防御蒸馏等通过混淆梯度来阻止基于梯度优化损失函数的防御方法，针对此类防御，提出了后向微分近似（backward pass differentiable approximation，BPDA）[23]，使用函数的可微分近似计算梯度信息。与之相似的，Uesato等人提出了同步扰动随机逼近（simultaneous perturbation stochastic approximtion，SPSA）[24]，用随机方向上的有限差分估计逼近梯度，同样可以绕过混淆梯度类的防御。

1.2.3 基于模型的决策边界

初始阶段对抗样本的研究主要集中在白盒攻击领域（攻击者能够获取模型的全部知识），但实际场景中，攻击者通常无法获取完整的模型信息，因此，在黑盒设置下生成对抗样本也是一个重要的研究方向[25-27]。仅依赖模型决策的攻击和迁移攻击是两类代表性的黑盒攻击方法。

仅能获取到模型输入和最终决策的设置最贴近实际应用场景，Brendel等人提出了仅依赖模型最终决策的边界攻击（boundary attack，BA）[28]，从一个扰动较大的对抗样本点开始，沿决策边界随机游走，逐步找到对抗区域内与良性样本距离最近的对抗样本。此类攻击最大的问题在于需要大量的查询来估计决策边界的梯度，其改进关键在于如何减少查询次数，如基于拉丁超立方体抽样的边界攻击（Latin hypercube sampling based boundary attack，LHS-BA），使用二进制搜索算法来搜索初始攻击位置，通过观

察网络决策结果来估计决策边界的梯度,最后将对抗性示例投射到边界,以促进下一个梯度估计[29]。

1.2.4 基于对抗样本的迁移性

迁移攻击的原理是利用对抗样本的迁移性,训练一个替代模型来模仿黑盒模型,使用白盒攻击方法在替代模型上生成对抗样本[30]。根据特征假说[18],对抗样本的迁移性来自于非鲁棒特征,因此基于特征实施迁移攻击的效果更好,其核心是如何破坏模型中间层的特征映射,最大化内部特征失真。Naseer等人提出了NRDM(neural representation distortion method)攻击[31],最大化攻击后所有神经元激活值的变化。FDA(feature disruptive attack)[32]通过平均激活值区分了神经元重要性的极性(即区分了中间层存在促进和抑制模型正确预测的积极和消极特征),FIA[19]对特征进行了更准确的评估,神经元归因攻击(neuron attribution-based attacks,NAA)进一步将输出完全归因于中间层的每个神经元[33]。

表2梳理总结了各类代表性攻击方法的特点。目前尚无普适的对抗样本防御技术方法,因而现有的防御技术方法主要针对表中攻击方法或相关衍生方法。

2 对抗样本防御技术

对抗样本防御技术可以从不同维度进行划分,通常以防御技术所针对的目标对象为依据,将其分为基于模型的防御和基于数据的防御两大类。基于模型的防御通过修改模型的架构或训练方法,通过提升模型本身的鲁棒性实现防御。基于数据的防御通过检测数据中的对抗样本或消除其对抗性扰动,在对抗样本进入到模型前实施防御。基于假说对防御的启示,分析代表性防御技术的防御机理,进行归纳总结。

2.1 基于模型的防御技术

对抗样本攻击的主要思想在于找到使模型的输出类别发生改变的微小扰动。因此,就模型层面而言,需要尽可能降低模型对此类微小扰动的敏感性。具体而言,可以从改进模型训练方法和改进模型结构两方面展开进行防御。

2.1.1 改进模型训练方法

(1)对抗训练。改进模型训练主要是对模型训练过程进行增强,主动生成对抗样本并将其加入到训练数据中进行对抗训练是一类典型方法。Goodfellow等人提出FGSM攻击时,也提出了使用FGSM攻击生成的对抗样本进行对抗训练(fast gradient sign method adversarial training,FGSM-AT)来进行防御[9]。Madry等人提出了使用PGD攻击方法生成对抗样本以进行对抗训练的PGD-AT(projected gradient descent adversarial training)防御方法[22],并提出了对抗训练的最小-最大(min-max)框架:

$$\min_{\theta}\rho(\theta),\rho(\theta)=\mathbb{E}_{(x,y)\sim\mathcal{D}}[\max_{\delta\in S}\mathcal{L}(\theta,x+\delta,y)] \tag{4}$$

其中,$\max_{\delta\in S}\mathcal{L}(\theta,x+\delta,y)$对应寻找损失最大的对抗样本,$S$用于约束扰动的最大范围。PGD-AT通过内部最大化问题优化参数θ,外部最小化模型训练的损失函数提升模型的鲁棒性。PGD-AT对FGSM攻击等

表2 典型对抗样本生成方法关联比较

Table 2 Connection and comparison of representative adversarial example generation methods

方法类别	攻击方法	提出动机	方法原理	主要特点
直接优化目标函数	L-BFGS[2]	仅通过简单随机采样难以找到“盲点”	将优化问题转化为更好解决的盒约束优化问题	迭代优化的计算开销较大
	C&W[15]	攻破防御蒸馏,重新审视防御方法	显式约束扰动的大小,通过logits输出保证模型预测错误	更好地控制扰动大小,可以绕过防御蒸馏
计算导数寻找扰动	FGSM[9]	佐证线性假说,快速找到扰动方向	沿损失函数的梯度方向施加扰动	速度快,容易防御
	JSMA[21]	有针对性地扰动对模型输出影响最显著的部分	基于雅可比矩阵构建输出对输入的敏感度的显著图	每次迭代都重新计算显著图,速度较慢
基于模型的决策边界	DeepFool[16]	更准确地计算攻击所需的扰动,量化模型鲁棒性	沿最短距离跨过分类超平面	扰动更不易察觉
	BA[28]	解决真实场景下如何进行攻击的问题	从一个大的扰动开始逐渐减小扰动,同时保持对抗性	适用于仅能获取模型最终决策的场景
基于对抗样本的迁移性	FIA[33]	对中间层特征进行区分,提升对抗样本的迁移性	构建加权特征图以选择性扭曲重要特征	针对特定特征进行攻击,具备更好的迁移性
	NAA[33]		将输出完全归因于中间层神经元,据此加权进行攻击	

代表性攻击方法均有较好的防御效果。

(2)对抗训练系列改进。对抗训练具有较好的防御效果,然而生成对抗样本并将其纳入模型的训练过程会明显增加计算成本。为此,Shafahi等人提出了Free-AT(free adversarial training)方法[34],通过循环利用梯度信息来降低计算开销。PGD-AT在每步迭代中都会计算参数θ的梯度和输入x的梯度,在处理外部最小化问题时只利用θ的梯度,在处理内部最大化问题时只利用x的梯度,梯度信息只会利用一次。而Free-AT通过循环利用更新模型参数时的梯度信息来减少计算梯度的次数。

min-max框架[22]明确了理想鲁棒分类器应该达到的目标,给出了模型鲁棒性的定量度量,但是提升模型对抗鲁棒性和标准训练的训练目标不同。从决策边界的角度看,对抗训练将能够改变模型预测结果的对抗扰动纳入到模型的学习过程中,修改了原本的决策边界。因此,需要在标准精度和鲁棒性之间做出权衡[35]。Zhang等人提出了TRADES(tradeoff-inspired adversarial defense via surrogate-loss minimization)[36],将鲁棒性误差分解为自然误差和边界误差之和,来描述分类问题的准确性和鲁棒性之间的权衡,并提出了新的损失。该损失由两项组成:经验风险最小化项鼓励算法最大限度地提高标准精度,正则化项鼓励算法将决策边界从数据中推开,以提高对抗鲁棒性。

对抗训练的模型还会表现出在训练数据上表现良好,但在未见过的对抗样本上表现更差,这种现象被称为鲁棒过拟合[37]。引入动态学习策略缓解这种现象,Jia等人提出了LAS-AT(adversarial training with learnable attack strategy)[38],LAS-AT由一个目标网络和一个策略网络组成,目标网络使用对抗样本进行对抗训练以提升模型鲁棒性,策略网络生成攻击策略以创建对抗样本来攻击目标网络。在这种博弈机制下,在训练的早期阶段,弱攻击可以成功地攻击目标网络。随着模型鲁棒性的提高,策略网络会学习生成策略以产生更强的攻击[38]。

(3)引入噪声。除了直接将对抗样本纳入训练过程,Liu等人提出了ANP(adversarial noise propagation)[39],在训练期间向神经网络的隐藏层注入对抗性噪声,起到类似对抗训练的效果,并且无需对抗训练的高计算成本。ANP还可以与其他对抗训练方法结合起来,进一步提高模型的鲁棒性。

(4)其他典型数据增强方法。对抗训练本质上是一种数据增强方法,通过增加训练数据的数量和多样性来提升模型的学习能力。结合特征向量的线性插值应该导致相关目标的线性插值的先验知识,Zhang等人提出了mixup[40],在训练样本之间进行线性插值来生成新的训练样本。Manifold Mixup[41]将mixup扩展到特征级,在隐藏表示上进行插值。通过在隐藏层应用mixup,鼓励模型在整个模型的决策边界上都保持平滑,而不仅仅是在输入空间[41]。

2.1.2 改进模型结构

(1)添加正则化层。改进模型结构主要是对模型本身的结构进行增强以增强模型抵御对抗样本的能力。如Gu等人对L-BFGS攻击生成的对抗样本进行了研究,将对抗样本防御问题等价为提高对每个样本最小对抗性噪声的发现能力,提出了深度收缩网络(deep contractive networks,DCN)[42]。DCN在神经网络的每一层都添加了惩罚层,最小化输出相对于输入扰动的方差,使模型在训练数据点周围达到“平坦”。DCN针对早期的对抗样本提出,对更强大的攻击方法防御效果不足,会产生较高的计算成本,且涉及到收缩率等复杂的超参数调整。

(2)可认证鲁棒性。通过量化模型鲁棒性,可以更好地理解和评估模型对于对抗样本攻击的敏感性,进而找到防御此类攻击的策略。Muthukumar等人提出了稀疏局部Lipschitzness(sparse local Lipschitzness,SLL)指标[43],用于测量局部敏感性,对于深度神经网络来说,Lipschitz常数较小时模型对微小扰动的敏感性也较小,引入正则化项鼓励模型拥有较小的Lipschitz常数可以提高模型的鲁棒性。SLL可以在稀疏程度和局部敏感性之间进行调节权衡,提供了一个可认证的鲁棒半径[43]。

(3)约束损失景观/决策边界。也可以从模型的损失景观(loss landscape)或决策边界入手。损失景观是模型损失函数的几何表示,向损失函数中添加正则化项,可以使得模型的损失景观变得更加平滑,进而降低模型对输入变化的敏感度。无论神经网络的权重是如何训练的,其输入损失景观的曲率程度与内在鲁棒性高度相关[44]。AdvRush[45]定义了一个神经架构搜索(neural architecture search,NAS)空间,然后引入一个倾向于选择具有更平滑损失景观的候选架构的正则化器,成功发现了具有高内在鲁棒性的神经网络架构。Moosavi-Dezfooli等人[46]研究了对抗训练对损失景观和决策边界几何结构的影响,对抗训练使损失景观的曲率明显降低,使模型表现出更线性的行为,这种线性是增强鲁棒性的来

源。Moosavi-Dezfooli等人还提出了一个正则化器CURE(curvature regularization)[46],直接将损失景观的曲率最小化,模仿对抗训练的效果,可以视为一种对抗训练的替代方案。

(4)混淆梯度。攻击者在优化攻击策略时,往往需要计算损失函数的梯度,如果梯度信息被混淆,就很难生成有效的对抗样本。防御蒸馏[14]也属于混淆梯度的防御[23],其原理是在高温下训练教师模型,使模型的输出更为“软”(即类别之间的概率差异更小),再使用这些“软标签”来训练学生模型,在实际使用中,将模型的温度设为低温,使得模型的输出更接近硬标签。通过这种方式,模型在训练过程中考虑了更多的信息,但是类别间的概率差异变小会导致反向传播的梯度变平滑,也就是梯度信息被混淆。尽管混淆梯度可能提高模型在一些攻击测试中的表现,但只是虚假的防御,对FGSM攻击和JSMA攻击等需要获取准确梯度信息的攻击有效,并非真正提升了模型的鲁棒性[23]。

2.2 基于数据的防御技术

基于数据的防御技术主要关注如何处理输入数据,以减少对抗样本对模型的影响。此类防御与模型的训练过程和模型结构无关,可以在不修改现有模型的情况下部署。具体可以分为对抗样本检测和对抗性消除。

2.2.1 对抗样本检测

(1)特征检测。对抗样本检测的主要目标是在对抗样本进入到模型前识别出来,核心在于提取对抗样本与良性样本的特征差异。高维空间中包含复杂的冗余信息,直接分析图像特征差异十分困难,因此,早期的对抗样本检测方法通常通过降维手段对图像的特征差异进行分析。如Hendrycks等人[47]采用主成分分析(principal components analysis,PCA)[48]对FGSM攻击生成的对抗样本和良性样本的主成分进行分析,发现对抗样本更强调低阶主成分,具有比良性样本更大的系数方差。PCA能否有效检测出对抗样本取决于样本中主要成分的分布是否被扰动,对FGSM攻击这种直接在输入级别添加噪声的攻击,PCA有一定的检测效果,但对如DeepFool攻击或C&W攻击等在原始特征空间中寻找扰动的更复杂的攻击,PCA的效果较差。

(2)分布统计特征。根据流形假说,对抗样本处于真实的数据流形之外,Feinman等人设计了密度估计(density estimates,DE)和贝叶斯不确定性估计(Bayesian uncertainty estimates,BUE)两个新的特征来检测对抗性样本[20]。DE在最后一个隐藏层的特征空间中用训练集计算,目标是检测远离数据流形的点,BUE基于贝叶斯推理考虑模型对于其预测的不确定性,模型对对抗样本预测的不确定性会比较大,BUE可以在DE无法检测的情况下检测出对抗性样本。将DE和BUE结合起来可以获得更好的检测效果,对FGSM攻击、JSMA攻击和C&W攻击等代表性攻击方法均有效。

(3)辅助分类器及图像重建。也可以直接通过辅助分类器对对抗样本进行检测,Metzen等人[49]训练了一个二分类器,专门用于区分输入是否为对抗样本。Hendrycks等人[47]还提供一种图像重建方法,将辅助模型用于图像重建,根据重建图像与原始图像差异来区分对抗样本与良性样本,通常来说对抗样本与其重建图像之间的差异更大。

图4将代表性的对抗样本检测方法归纳为三种典型检测架构。对抗样本检测通常基于一些直观的观察和假设,如对抗样本异常的统计特性或者它们可能位于模型决策边界的附近,往往特定的攻击策略设计,对其他攻击策略可能效果不佳。此外,还可能会漏报对抗样本或将良性样本误报为对抗样本,并且一旦攻击者了解了检测器的检测策略,很容易设计出可以绕过检测的对抗样本。

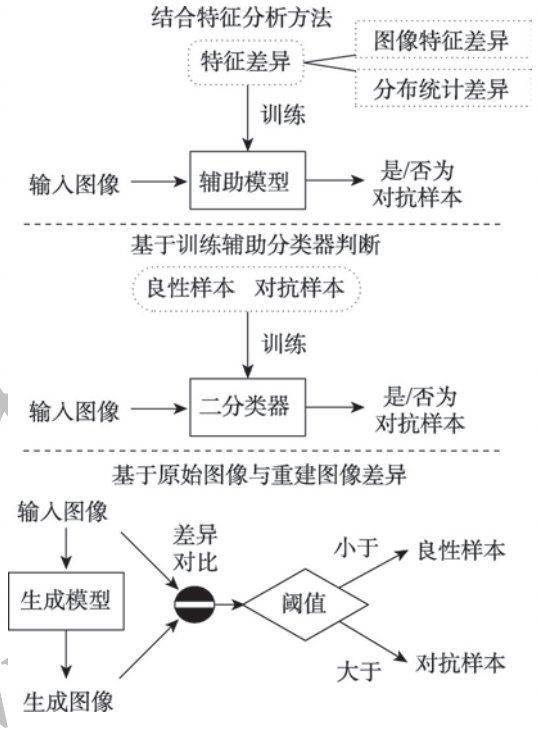

图4 典型检测架构

Fig.4 Typical detection architecture

2.2.2 对抗性消除

对抗性消除的目标是对输入数据进行处理，消除或减少对抗性扰动，使模型产生正确预测。和对抗样本检测类的方法相似，对抗性消除方法也是基于已有攻击方法生成的对抗样本进行研究。

（1）引入随机性。针对简单的攻击方法可以采用随机化的防御措施，如Xie等人[50]认为FGSM等单步攻击不够强大，无法欺骗网络，DeepFool等迭代攻击泛化能力较弱，调整大小、填充、压缩等低级别的图像变换就可能会破坏对抗性扰动的特定结构，提出了随机调整大小和随机填充（random resizing and padding，RRP），通过添加两个随机化层进行防御。这种方法对良性样本的精度几乎没有影响，并且几乎没有增加额外的计算成本，非常适合作为一个基本模块与其他防御方法相结合。

（2）流形约束。假设对抗样本位于正确标签的数据流形之外，那么针对对抗性样本的成功防御机制应旨在将图像投射回正确的数据流形上[51]。Dubey等人[51]提出在图像数据库中寻找最近的“邻居”来逼近对抗样本在图像流形上的投影，对投影进行分类。这种防御方法在完全白盒攻击场景中并不有效，但在更贴近现实的场景中，即使攻击者知道防御策略，也无法访问用于寻找邻居拟合流形的图像数据库，可以进行有效的防御。

（3）去噪。由于对抗样本是在原始图像上添加噪声构建的，去噪是消除图像对抗性噪声的一个直接思路，然而，去噪器可能无法消除所有的扰动，小的残余扰动在目标模型的顶层可能被放大到很大的幅度[52]。为了解决这个问题，Liao等人提出了HGD（high-level representation guided denoiser）[52]，将损失函数设计为良性样本和对抗样本在目标模型的顶层输出之间的差异。HGD的防御效果较好，与对抗训练等有效的防御方法相比，计算成本更低，并且适用于更广泛的数据集和攻击方法[52]。

（4）图像重建。随着生成式模型技术的发展，许多防御致力于借助生成式方法将输入样本重建为良性样本，其核心是通过训练生成模型来理解数据的潜在分布，将对抗样本映射到良性样本：Meng等人提出了MagNet[53]，用一对自编码器实施图像重构；Samangouei等人提出了Defense-GAN[54]，用生成对抗网络（generative adversarial networks，GAN）学习良性样本的分布；Nie等人受扩散模型反向生成过程与对抗样本去噪过程相似的启发，提出了DiffPure[55]，通过扩散模型进行图像重建。此类防御最大的问题是训练生成模型需要大量的数据和计算资源。

表3汇总梳理了当前代表性防御技术。分析各类防御技术方法的优缺点，根据实际需求设计更有针对性的防御策略，提升防御性能效果。

表3 代表性防御技术比较

Table 3 Comparison of representative defense technology

类型	目标	手段	原理	防御技术	实现手段	防御机理	优点	缺点
基于模型的防御	降低模型对对抗性扰动的敏感性	改进训练方法	增加训练数据的数量和多样性提升模型的鲁棒性	对抗训练[22,34,36-37]	将对抗样本纳入训练过程	减少模型盲区 平滑决策边界	效果较好	训练成本较高，标准精度下降等
				ANP[39]	注入噪声			
				mixup[40]	插值			
		改进模型结构	通过添加网络层等修改模型结构的方式降低模型对对抗性扰动的敏感性	DCN[42]	引入惩罚项	平滑决策边界	对于特定的模型结构、攻击类型，防御效果较好	模型变得更复杂，只针对特定的模型结构，无法防御未知攻击
				SLL[43]	可认证鲁棒半径			
				CURE[46]	正则化器			
				防御蒸馏[14]	二次训练（蒸馏）	减少梯度信息输出		
基于数据的防御	减少数据中扰动对模型输出的影响	对抗样本检测	在对抗样本进入到模型前识别出来	PCA检测[47]	系数方差	提取/放大特征差异	无需改变模型结构，易与其他方法结合使用	不易检测微小扰动
				DE[20]	流形空间位置			
				BUE[20]	分类不确定性			
		减少或消除对抗性	在对抗样本进入到模型前消除或减少对抗性扰动	RRP[50]	引入随机性	增加梯度计算的难度		
				最近邻[51]	K-近邻	投影回非对抗流形		
				HGD[52]	特征级去噪	移除噪声		
				图像重建[53-55]	重建良性样本	生成模型重建图像		

3 总结与展望

3.1 现状总结

针对图像对抗样本的攻防技术研究是当前人工智能安全领域的研究热点。从维护人工智能安全发展的角度而言,需要综合梳理分析现有的图像对抗样本防御技术方法,对比分析其特点,为后续研究提供技术参考。

现有的对抗样本攻击和防御方法通常针对特定的模型和数据集,防御方法通常针对已有的攻击方法进行设计,大致可分为基于模型的防御和基于数据的防御。

基于模型的防御施加于具体的模型之上,因此在特定范围的数据集上对指定范围的攻击方法防御效果通常更好,但在新的、未知的对抗样本面前,可能表现不佳。此类防御方法在模型的训练阶段将防御机制整合到模型中,在预测阶段无需增加额外的计算负担,但是可能在训练阶段产生高昂的计算成本。如对抗训练是当前较为有效的防御技术之一,然而其计算成本较高。很多改进工作都在对其进行改进,降低其计算成本,或权衡模型精度和鲁棒性,或提升模型泛化等。

相比之下,基于数据的防御不需要改变原有的模型结构或训练过程,更容易实施,并且更易于与其他防御策略进行结合,进一步提高模型的整体防御能力。在设计相关策略时,理想情况是覆盖所有可能的扰动,但对抗样本的潜在空间非常大,很难全面覆盖。此外,检测方法可能会将一些正常样本错误地判定为对抗样本,对于其中涉及到训练检测器或净化器的方法,训练这些检测或净化模块需要大量标记的对抗样本,成本也较高。

整体而言,对抗样本的防御是一个复杂且具有挑战性的问题。虽然已有许多防御策略,但是泛化性均不理想,目前尚没有一种策略可以完全解决这个问题。从现实应用的角度而言,需要综合考虑具体的应用场景、模型的潜在安全风险、输入数据的格式等因素来选择适合的一个或多个防御方法进行防护,在已有的防护策略中进行最优配置。

3.2 未来展望

随着各界对人工智能安全越来越重视,围绕图像对抗样本的防御技术研究正在快速发展。防御技术的研究既需要根据不断出现的新攻击方法进行针对性的单点防御,更需要不断探索更具泛化性的全面防御。结合现有技术发展历程及作者开展的研究实际,图像对抗样本防御技术的未来发展需要关注诸多问题:

(1)建立更具泛化性的防御理论

当前已有的防御方法大多根据具体的攻击方法特点进行针对性的防御,难以实现对不同类型攻击方法的泛化防御,容易被更新的攻击方法所突破。亟待结合图像对抗样本存在机理的研究进展建立泛化性更好的防御理论,以确保智能模型在面对各类对抗样本攻击时能安全运行。

(2)构建更加客观的防御效果评价体系

当前对于图像对抗样本防御技术的评价主要基于公开数据集对比模型性能的下降程度,难以反映各类复杂场景下的防御效果。亟待构建涵盖多种应用场景和不同类型图像数据以及跨图像模态数据,从多个维度对防御效果进行评价的基准测评体系。

(3)设计智能系统体系化安全防御策略

当前研究主要针对数字空间的智能模型本身进行对抗样本防御,而现实世界中的真实智能系统通常包括大量的外围非智能部件,智能系统面临的安全问题更加复杂。亟待统筹考虑智能系统中的智能因素和非智能因素,体系化设计安全防御策略,提升应对现实安全威胁的能力。

4 结束语

图像对抗样本防御技术研究是当前人工智能安全领域的研究热点,处在快速发展阶段。本文结合图像对抗样本防御技术最新研究成果,系统梳理分析了对抗样本的各类存在原因假说、攻击方法和防御方法,及其之间的关联关系与发展脉络,以期为领域相关人员提供参考。

参考文献:

[1] AMODEI D, OLAH C, STEINHARDT J, et al. Concrete problems in AI safety[J]. arXiv:1606.06565, 2016.

[2] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[J]. arXiv:1312.6199, 2013.

[3] ZHANG Q, HU S, SUN J, et al. On adversarial robustness of trajectory prediction for autonomous vehicles[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, Jun 18-24, 2022. Piscataway: IEEE, 2022: 15159-15168.

[4] ALI W, QURESHI E, FAROOQI O A, et al. Pneumonia detection in chest X-ray images: handling class imbalance[J]. arXiv:2301.08479, 2023.

[5] TIAN Y, NI Z, CHEN B, et al. Just noticeable difference modeling for face recognition system[J]. arXiv:2209.05856, 2022.

[6] WIYATNO R R, XU A, DIA O, et al. Adversarial examples in modern machine learning: a review[J]. arXiv:1911.05268, 2019.

[7] 白祉旭, 王衡军, 郭可翔. 基于深度神经网络的对抗样本技术综述[J]. 计算机工程与应用, 2021, 57(23): 61-70.
BAI Z X, WANG H J, GUO K X. Summary of adversarial examples techniques based on deep neural networks[J]. Computer Engineering and Applications, 2021, 57(23): 61-70.

[8] 张田, 杨奎武, 魏江宏, 等. 面向图像数据的对抗样本检测与防御技术综述[J]. 计算机研究与发展, 2022, 59(6):1315-1328.
ZHANG T, YANG G W, WEI J H, et al. Survey on detecting and defending adversarial examples for image data[J]. Journal of Computer Research and Development, 2022, 59 (6): 1315-1328.

[9] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv:1412.6572, 2014.

[10] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial examples in the physical world[C]//Proceedings of the 5th International Conference on Learning Representations, Toulon, Apr 24-26, 2017: 1-15.

[11] TRAMÈR F, KURAKIN A, PAPERNOT N, et al. Ensemble adversarial training: attacks and defenses[J]. arXiv:1705.07204, 2017.

[12] DONG Y, LIAO F, PANG T, et al. Boosting adversarial attacks with momentum[C]//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, Jun 18-23, 2018. Washington: IEEE Computer Society, 2018: 9185-9193.

[13] SCHWINN L, NGUYEN A, RAAB R, et al. Dynamically sampled nonlocal gradients for stronger adversarial attacks [C]//Proceedings of the 2021 International Joint Conference on Neural Networks, Shenzhen, Jul 18-22, 2021. Piscataway: IEEE, 2021: 1-8.

[14] PAPERNOT N, MCDANIEL P, WU X, et al. Distillation as a defense to adversarial perturbations against deep neural networks[C]//Proceedings of the 2016 IEEE Symposium on Security and Privacy, California, May 23-25, 2016. Piscataway: IEEE, 2016: 582-597.

[15] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//Proceedings of the 2017 IEEE Symposium on Security and Privacy, San Jose, May 22-26, 2017. Washington: IEEE Computer Society, 2017: 39-57.

[16] MOOSAVI-DEZFOOLI S M, FAWZI A, FROSSARD P. DeepFool: a simple and accurate method to fool deep neural networks[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, Jun 27-30, 2016. Washington: IEEE Computer Society, 2016: 2574-2582.

[17] MOOSAVI-DEZFOOLI S M, FAWZI A, FAWZI O, et al. Universal adversarial perturbations[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Jul 21-26, 2017. Washington: IEEE Computer Society, 2017: 1765-1773.

[18] ILYAS A, SANTURKAR S, TSIPRAS D, et al. Adversarial examples are not bugs, they are features[C]//Advances in Neural Information Processing Systems 32, Vancouver, Dec 8-14, 2019: 125-136.

[19] WANG Z, GUO H, ZHANG Z, et al. Feature importance-aware transferable adversarial attacks[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, Montreal, Oct 10-17, 2021. Piscataway: IEEE, 2021: 7639-7648.

[20] FEINMAN R, CURTIN R R, SHINTRE S, et al. Detecting adversarial samples from artifacts[J]. arXiv:1703.00410, 2017.

[21] PAPERNOT N, MCDANIEL P, JHA S, et al. The limitations of deep learning in adversarial settings[C]//Proceedings of the 2016 IEEE European Symposium on Security and Privacy, Saarbrücken, Mar 21-24, 2016. Piscataway: IEEE, 2016: 372-387.

[22] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[J]. arXiv: 1706.06083, 2017.

[23] ATHALYE A, CARLINI N, WAGNER D. Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples[C]//Proceedings of the 2018 International Conference on Machine Learning, Stockholm, Jul 10-15, 2018. New York: ACM, 2018: 274-283.

[24] UESATO J, O'DONOGHUE B, KOHLI P, et al. Adversarial risk and the dangers of evaluating against weak attacks [C]//Proceedings of the 2018 International Conference on Machine Learning, Stockholm, Jul 10-15, 2018. New York: ACM, 2018: 5025-5034.

[25] INKAWHICH N, LIANG K J, CARIN L, et al. Transferable perturbations of deep feature distributions[J]. arXiv: 2004.12519, 2020.

[26] SHI C, HOLTZ C, MISHNE G. Online adversarial purifica-

tion based on self-supervision[J]. arXiv:2101.09387, 2021.

[27] ANDRIUSHCHENKO M, CROCE F, FLAMMARION N, et al. Square attack: a query-efficient black-box adversarial attack via random search[C]//LNCS 12368: Proceedings of the 2020 European Conference on Computer Vision, Glasgow, Aug 23-28, 2020. Cham: Springer, 2020: 484-501.

[28] BRENDEL W, RAUBER J, BETHGE M. Decision-based adversarial attacks: reliable attacks against black-box machine learning models[J]. arXiv:1712.04248, 2017.

[29] WANG D, LIN J, WANG Y G. Query-efficient adversarial attack based on Latin hypercube sampling[C]//Proceedings of the 2022 IEEE International Conference on Image Processing, Bordeaux, Oct 16-19, 2022. Piscataway: IEEE, 2022: 546-550.

[30] PAPERNOT N, MCDANIEL P, GOODFELLOW I, et al. Practical black-box attacks against machine learning[C]// Proceedings of the 2017 ACM Asia Conference on Computer and Communications Security, Abu Dhabi, Apr 2-6, 2017. New York: ACM, 2017: 506-519.

[31] NASEER M, KHAN S H, RAHMAN S, et al. Task-generalizable adversarial attack based on perceptual metric[J]. arXiv:1811.09020, 2018.

[32] GANESHAN A, BS V, BABU R V. FDA: feature disruptive attack[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Oct 27-Nov 2, 2019. Piscataway: IEEE, 2019: 8069-8079.

[33] ZHANG J, WU W, HUANG J, et al. Improving adversarial transferability via neuron attribution-based attacks[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, Jun 18-24, 2022. Piscataway: IEEE, 2022: 14993-15002.

[34] SHAFAHI A, NAJIBI M, GHIASI A, et al. Adversarial training for free![C]//Advances in Neural Information Processing Systems 32, Vancouver, Dec 8-14, 2019: 3353-3364.

[35] TSIPRAS D, SANTURKAR S, ENGSTROM L, et al. Robustness may be at odds with accuracy[J]. arXiv:1805.12152, 2018.

[36] ZHANG H, YU Y, JIAO J, et al. Theoretically principled trade-off between robustness and accuracy[C]//Proceedings of the 36th International Conference on Machine Learning, Long Beach, Jun 9-15, 2019: 7472-7482.

[37] CHEN T, ZHANG Z, LIU S, et al. Robust overfitting may be mitigated by properly learned smoothening[C]//Proceedings of the 9th International Conference on Learning Representations, Austria, May 3-7, 2021: 1-19.

[38] JIA X, ZHANG Y, WU B, et al. LAS-AT: adversarial training with learnable attack strategy[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, Jun 18-24, 2022. Piscataway: IEEE, 2022: 13398-13408.

[39] LIU A, LIU X, YU H, et al. Training robust deep neural networks via adversarial noise propagation[J]. IEEE Transactions on Image Processing, 2021, 30: 5769-5781.

[40] ZHANG H, CISSE M, DAUPHIN Y N, et al. mixup: beyond empirical risk minimization[J]. arXiv:1710.09412, 2017.

[41] VERMA V, LAMB A, BECKHAM C, et al. Manifold mixup: better representations by interpolating hidden states[C]//Proceedings of the 2019 International Conference on Machine Learning, Long Beach, Jun 9-15, 2019. New York: ACM, 2019: 6438-6447.

[42] GU S, RIGAZIO L. Towards deep neural network architectures robust to adversarial examples[J]. arXiv:1412.5068, 2014.

[43] MUTHUKUMAR R, SULAM J. Adversarial robustness of sparse local lipschitz predictors[J]. arXiv:2202.13216, 2022.

[44] ZHAO P, CHEN P Y, DAS P, et al. Bridging mode connectivity in loss landscapes and adversarial robustness[J]. arXiv:2005.00060, 2020.

[45] MOK J, NA B, CHOE H, et al. AdvRush: searching for adversarially robust neural architectures[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision, Montreal, Oct 10-17, 2021. Piscataway: IEEE, 2021: 12302-12312.

[46] MOOSAVI-DEZFOOLI S M, FAWZI A, UESATO J, et al. Robustness via curvature regularization, and vice versa[C]// Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, Jun 15-20, 2019. Piscataway: IEEE, 2019: 9078-9086.

[47] HENDRYCKS D, GIMPEL K. Early methods for detecting adversarial images[J]. arXiv:1608.00530, 2016.

[48] LEVER J, KRZYWINSKI M, ALTMAN N. Points of significance: principal component analysis[J]. Nature Methods, 2017, 14(7): 641-643.

[49] METZEN J H, GENEWEIN T, FISCHER V, et al. On detecting adversarial perturbations[J]. arXiv:1702.04267, 2017.

[50] XIE C, WANG J, ZHANG Z, et al. Mitigating adversarial effects through randomization[J]. arXiv:1711.01991, 2017.

[51] DUBEY A, MAATEN L, YALNIZ Z, et al. Defense against adversarial images using web-scale nearest-neighbor search [C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, Jun 15-20, 2019. Piscataway: IEEE, 2019: 8767-8776.

[52] LIAO F, LIANG M, DONG Y, et al. Defense against adve-

rsarial attacks using high-level representation guided denoiser[C]//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, Jun 18-23, 2018. Washington: IEEE Computer Society, 2018: 1778-1787.

[53] MENG D, CHEN H. MagNet: a two-pronged defense against adversarial examples[C]//Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, New York, Oct 30-Nov 3, 2017. New York: ACM, 2017: 135-147.

[54] SAMANGOUEI P, KABKAB M, CHELLAPPA R. Defense-GAN: protecting classifiers against adversarial attacks using generative models[J]. arXiv:1805.06605, 2018.

[55] NIE W, GUO B, HUANG Y, et al. Diffusion models for adversarial purification[J]. arXiv:2205.07460, 2022.

刘瑞祺(1994—),女,辽宁沈阳人,硕士研究生,主要研究方向为网络空间安全、人工智能等。
LIU Ruiqi, born in 1994, M.S. candidate. Her research interests include cyberspace security, artificial intelligence, etc.

李虎(1987—),男,甘肃定西人,博士,工程师,主要研究方向为人工智能安全、机器学习等。
LI Hu, born in 1987, Ph.D., engineer. His research interests include artificial intelligence security, machine learning, etc.

王东霞(1974—),女,河南焦作人,博士,研究员,博士生导师,主要研究方向为网络空间安全。
WANG Dongxia, born in 1974, Ph.D., professor, Ph.D. supervisor. Her research interest is cyberspace security.

赵重阳(1998—),男,山东聊城人,硕士研究生,主要研究方向为人工智能安全。
ZHAO Chongyang, born in 1998, M.S. candidate. His research interest is artificial intelligence security.

李博宇(1993—),男,河北石家庄人,硕士研究生,主要研究方向为信息安全、攻击检测等。
LI Boyu, born in 1993, M.S. candidate. His research interests include information security, attack detection, etc.