深度学习对抗样本的防御方法综述

张嘉楠[1]，赵镇东[1]，宣晶[2,3]，常晓林[1]

（1.北京交通大学智能交通数据安全与隐私保护技术北京市重点实验室，北京 100044；2.北京京投卓越科技发展有限公司，北京 100101；3.北京京投信安科技发展有限公司，北京 100101）

摘要：深度学习技术的出现给许多领域带来了突破，被广泛地应用于多个实际场景中。在解决许多复杂问题方面，深度学习的表现已经超过了人类水平。但研究表明，深度学习模型容易受到对抗样本的攻击而产生不正确的输出，进而被攻击者加以利用，这影响到实际应用系统的可靠性和安全性。面对对抗样本的不同攻击方法，本文从模型和数据两个方面对防御方法进行了分类，总结了不同分类下防御方法的研究思路和研究进展，并给出了下一步对抗深度学习的发展方向。

关键词：深度学习；对抗样本；防御技术

中图法分类号：TP309.2 **文献标识码：**A

Survey of Defense of deep learning against adversarial examples

Zhang Jianan[1], Zhao Zhendong[1], Xuan Jing[2,3], Chang Xiaolin[1]

(1.Beijing Key Laboratory of Security and Privacy in Intelligent Transportation, Beijing Jiaotong University, Beijing 100044; 2.Beijing Jingtou Zhuoyue Technology Development Co., Ltd, Beijing 100101; 3.Beijing Jingtou Xin'an Technology Development Co., Ltd, Beijing 100101)

Abstract: The emergence of deep learning technology has brought breakthroughs in many fields, and it is widely used in multiple real-world scenarios. In terms of solving various complex problems, deep learning has outperformed humans. However, studies have shown that the deep learning model is vulnerable to attacks from adversarial examples and produces incorrect output, which is then exploited by the attacker to affect the reliability and security of the actual application system. In the face of different attack methods of adversarial examples, this paper classifies the defense methods from two aspects of model and data, summarizes the research ideas and research progress of defense methods under different classifications, and gives the development direction of the next step of adversarial deep learning.

Keywords: machine learning; adversarial examples; defense technology

删除的内容: many

删除的内容: be

删除的内容: ed

删除的内容: by the

删除的内容: makes

1 引言

近年来，深度学习理论技术不断成熟，在人工智能、大数据分析以及安全检测等方面都取得了很好的应用成果，它突破性地被应用在现实生活的很多领域中，在促进社会进步层面起到了关键作用。然而在带来便利的同时，深度学习本身也存在着一定的安全问题。如敌手的攻击和数据隐私的问题，这引起了安全领域的极大关注。

对抗样本指的是攻击者在数据集原始输入样本通过添加人类无法察觉的细微扰动来形成新的输入样本，导致模型以高置信度给出一个错误的输出，以欺骗机器学习模型。2013 年，Szegedy 等人[1]首先通过添加轻微扰动来干扰输入样本，使基于深度神经网络（Deep neural network, DNN）的图片识别系统输出攻击者想要的任意错误结果，此过程称为对抗攻击（Adversarial attack）。研究人员表明，现代深度神经网络模型极易受到人类视觉系统几乎无法察觉的微小扰动的对抗攻击。这种攻击可以造成神经网络分类器对原始图像进行错误预测。如果将其应用于现实世界，如恶意代码检测、无人驾驶系统、生物医学领域，将会带来极为严重的后果。

在之前的研究里，我们基于敌手知识和对抗特异性对对抗样本的攻击方法进行了细致的分类，本文根据前一部分研究提出的攻击方法，从模型和数据两个方面对防御方法分类，分析总结了分类方法

下各种防御技术的研究进展，并给出了对未来防御技术研究方向的展望。

本文组织结构如下：第 2 节简要介绍了深度学习的相关知识并回顾了对抗样本的攻击方法分类。第 3 节对现有的防御方法进行分类并介绍了防御方法的基本原理和研究现状。第 4 节基于所调查的文献提供了对未来对抗样本防御方法研究方向的展望。

2 背景知识

2.1 深度学习技术

深度学习是机器学习衍生出的一个研究方向，更偏近机器学习最初的目标——人工智能。它是一种基于对数据进行表征学习的方法，可以学习样本数据的内在规律及表示层次，通过建立模拟人脑进行学习分析的神经网络来模仿人脑机制处理数据。其解决问题的过程分为训练和预测两个阶段。

深度神经网络（DNN）是典型的深度学习模型，其他深度学习模型在其基础上进行扩展。DNN 本质是一个函数链，是由多个神经网络层累加起来的结构，神经网络层由一个个人工神经元构成，每个神经元都是一个感知器，可以将一组输入映射到具有激活功能的输出值上。DNN 每个函数是由每一层上的神经元组成，其目标是使训练的模型与真实的数据生成过程相匹配。函数表达如下所示：

$$f(x)=f^{(k)}\left(\cdots f^{(i)}\left(\cdots f^{(2)}\left(f^{(1)}\right)\cdots\right)\cdots\right)$$

其中$f^{(i)}$是第i层的函数，$i=1,2\ldots k$。

深度学习根据在训练过程中是否给定标签可以分为监督学习和无监督学习，其中监督学习算法有卷积神经网络（CNN）、循环神经网络（RNN）；无监督学习算法有深度置信网络（DBN）。

深度神经网络的强表达能力使其在许多领域取得了巨大的成功。CNN 被广泛应用于计算机视觉领域，RNN 在处理具有可变长度的顺序输入数据上，具有很好的处理效果。深度学习解决某些复杂问题的能力已经超出了人类水平，但研究表明，深度学习技术也面临多种安全性威胁。

2.2 对抗攻击方法总结

对抗样本最早由 Szegedy 等人[1]提出，通过在数据集添加轻微扰动干扰原始样本，导致模型以高置信度给出错误输出。在许多情况下，人类不会察觉原始样本和对抗样本之间的差异，但是神经网络会做出很大差异性的错误预测。

很多研究者对对抗样本的生成与攻击方法展开了研究，我们之前的研究根据敌手知识和对抗特异性，对攻击方法进行了细致的分类。根据敌手知识可分为白盒攻击和黑盒攻击，白盒攻击指攻击者完全了解神经网络模型和参数；黑盒攻击指攻击者无法获取模型全部信息，只能通过对模型的使用来观察输入输出并展开攻击。根据对抗特异性可以分为针对目标攻击和非针对目标攻击，针对目标攻击中对抗样本的分类结果会错分到指定分类；而非针对目标攻击的对抗类输出是任意的。表 1 给出对典型的对抗攻击方法的分类。

表 1 典型对抗攻击方法的分类

Table1 Classification of typical attack methods

	白盒攻击	黑盒攻击
非针对目标攻击	FGSM R+FGSM BIM&ILCM PGD CPPN EA Fool DeepFool UAP L-BFGS	FGSM U-MI-FGSM UAP One Pixel ZOO 结合 GAN 的对抗攻击
针对目标攻击	FGSM JSMA C&W Hot/Cold ATNs	JSMA T-MI-FGSM ATNs One Pixel ZOO UPSET ANGRI Houdini EAD 基于模型的集成攻击

3 对抗深度学习的防御方法

针对深度学习中的对抗样本攻击，本文对对抗样本的典型防御方法从模型和数据两个方向展开研究。模型层面的防御策略可分类为修改网络和使用附加网络，通过在训练阶段修改原始 DNN 模型的结构，或者不改变原始模型，用外部模型作为附加网络，使得防御后的 DNN 分类器能够检测出对抗样本或将其识别为正确标签；数据层面的防御策略主要通过在训练阶段将对抗样本注入训练数据集后重新训练模型，或预测阶段对样本进行修改，进行重建并将转换后的对抗样本输入到原模型来进行预测。

3.1 模型层面防御方法

从模型入手的防御方法主要分为以下两种：

（1）修改网络：仅修改原始模型的结构。

（2）使用附加网络：在保持原始模型所有信息的情况下，用外部模型作为附加网络。

3.1.1 修改网络

（1）防御蒸馏

Distillation（蒸馏）最早由 Hinton[2]提出，是指将复杂网络的知识迁移到简单网络上。该知识以训

练数据的类概率向量形式提取，并反馈给原始模型。Papernot[3]提出了防御蒸馏，是蒸馏算法的扩展，如图 1 所示，利用蒸馏算法为原始模型训练一个蒸馏模型。训练蒸馏模型时，输入是训练原始模型所需的样本集合。作者在 MNIST 和 CIFAR-10 两个数据集上测试应用防御蒸馏技术前后对抗样本的欺骗率得到：针对 MNIST 数据集，对抗样本的成功欺骗率从 95.86%降到 0.45%，CIFAR-10 数据集从 87.89%降到 5.11%。

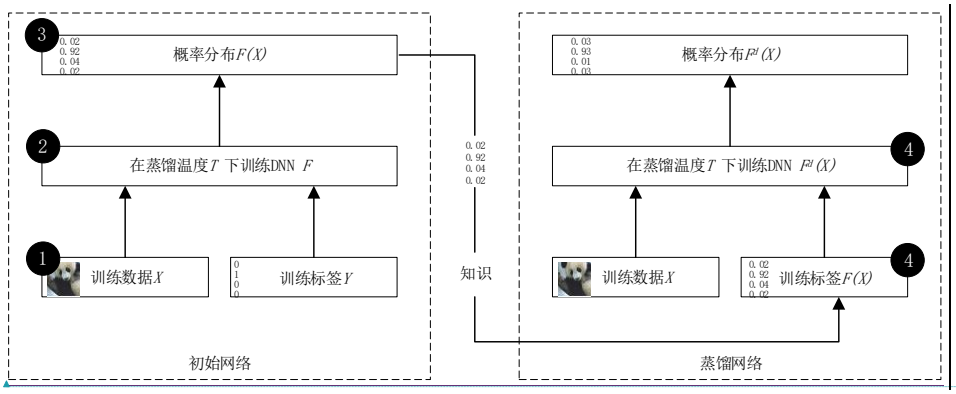

图 1 防御蒸馏原理图

Figure 1 Process of defensive distillation

Anil 等人[4]将蒸馏技术与分布式随机梯度下降（Stochastic gradient descent, SGD）相结合，分布式环境中的每个节点之间都可以互为教师模型和学生模型，并且互相提取内在知识，用以提升其他节点的模型性能。使用在线蒸馏方法降低了分布式 SGD 的通信开销，有效提高了模型预测的准确性。可以证明防御蒸馏技术可以显著降低损失函数梯度值，抵抗小幅度扰动的对抗攻击，但在黑盒攻击和未知模型函数的情况下，特征值的改变不能有效抵抗对抗攻击[5]。该方法的局限性在于只能对抗有限的对抗样本，研究者还需进一步研究更有效的安全防御算法。

（2）正则化

正则化方法是指在训练过程中在目标函数上惩罚输出对于输入的变化程度，可以在一定程度上使小的对抗扰动不会对输出有显著影响。Lyu 等人[6]使用一组联合的正则化方法对模型进行训练，以对抗基于 L-BFGS 和 FGSM 的攻击。Ross 等人[7]使用输入梯度正则化以提高对抗攻击的鲁棒性，在训练的目标函数上惩罚输出相对于输入的变化程度，产生的小的对抗性扰动不会对模型的预测结果造成显著影响。由于对抗训练需要增加训练数据量，Miyato 等人[8]提出虚拟对抗训练（Virtual adversarial training, VAT）方法。VAT 方法是一种新颖的半监督学习正则化方法，该方法在对抗训练的基础上，实现局部分布平滑。Moosavi-Dezfooli[10]指出对抗训练的主要作用之一是使损失函数的曲率和分类器的决策边界显著减小，基于此，他们提出了一种新的正则化策略，即曲率正则化，可以直接最小化损失面的曲率。这种正则化方法被证明可以显著提高神经网络的鲁棒性，甚至达到与对抗训练相当的性能，但可能在一定程度上会使模型的效果（如准确度）变差。此外，正则化方法与对抗训练结合会有很好的效果，但计算复杂度太高。

（3）深度压缩网络

Gu[11]引入了深度压缩网络方法（Deep contractive network, DCN），在训练过程中采用正则化方法使用压缩自编码器的平滑惩罚项，使得模型的输出更加平滑。Osadchy 等人[12]应用一组滤波器来消除对抗噪声，例如中值滤波器、均值滤波器、高斯低通滤波器。Liao 等人[13]使用高阶表征引导去噪器（High-level representation guided denoiser, HGD），训练一个基于神经网络的去噪器来消除对抗扰动。该方法使用 U-Net 作为去噪网络，相较于编码器和解码器结构，U-Net 在同分辨率的编码层和解码层之间直接相连，网络只需学习如何去除噪音，而无需重建整张图。

3.1.2 使用附加网络

（1）防御通用扰动

Akhtar 等人[14]提出了一种防御框架，该框架将额外的神经网络层附加到目标网络中，并重新训练网络来完成对对抗样本的校正，使目标网络对图像

域代码已更改

删除的内容: [8,9]

的原始版本与相应对抗样本的预测相同。通过这种方式不需要调整系数，而且能有效防御对抗样本。

（2）基于 GAN 的防御

Lee 等人[15]利用生成对抗网络（GAN）来训练一个可以抵抗 FGSM 攻击的网络。作者直接在生成网络上训练，在训练过程中，生成网络不断尝试对原始和对抗图像进行正确分类。Shen 等人[16]使用网络的生成器部分来修正一个受干扰的图像。

Samangouei 等人[17]提出一种新的防御策略 Defense-GAN，利用 GAN 来增强分类模型对白盒和黑盒对抗攻击的鲁棒性。实验表明，Defense-GAN 可以有效抵抗对抗攻击，但如果 GAN 没有得到适当的训练和调整，Defense-GAN 会受到原始输入样本和对抗样本的影响。

（3）对抗样本检测

上述使用附加网络的防御方法可以使得防御后的 DNN 分类器能够将对抗样本识别为正确的标签，而对抗样本检测只需判断输入样本是否为对抗样本，而无须将对抗样本识别为正确标签。

Feature Squeezing 方法[18]通过对输入样本压缩简化来检测输入样本是否为对抗样本，该方法在 DNN 分类器中添加了两个外部模型，分别用来减少每个像素的颜色位深度和进行像素值的空间平滑。将原始输入图片和用两种 Squeezing 方法压缩后的图片经过分类器预测后的两个结果进行比较，如果距离很大，则输入样本会被认为是对抗样本。文献[18]指出 Feature Squeezing 方法与对抗训练结合，分类结果会有更高的准确性。

Meng 等人[19]提出了一个框架使用一个或多个外部探测器将输入图像分类为对抗图像或原始图像。在训练期间，该框架的目的是学习各种无扰动的原始图像。MagNet 首先使用 Detector 将扰动量大的对抗样本直接丢弃；然后针对扰动量小的对抗样本，使用 Reformer 努力将其转化成正常样本，最后再交由模型识别。

SafetyNet[20]由分类神经网络和检测器组成，如果检测器检测出其输入样本为对抗样本，则该样本被检测器拒绝，不再被分类器分类。

3.2 数据层面防御方法

（1）训练阶段修改模型参数（对抗训练）

自从发现深度神经网络的对抗样本以来，相关文献中普遍认为，防御对抗样本的神经网络的鲁棒性会随着对抗训练而提高。对抗训练方法从训练数据集入手，在每个训练步骤中产生对抗样本，并将它们注入训练集，构建鲁棒性更好的模型。Goodfellow 等人[21]和 Huang 等人[22]使用对抗训练防御方法 MNIST 数据集上进行评估，实验表明，这种混合了合法样本和对抗样本训练出的模型有更强的鲁棒性。

Kuraki 等人[23]在 ImageNet 数据集上对对抗训练防御方法进行综合分析，即一般对抗训练。结果显示，对抗训练增加了神经网络对于单步攻击（如 FGSM）的鲁棒性，但对迭代攻击无效。Dong 等人[24]在对抗训练期间最小化了交叉熵损失和内部表示距离，即 PGD 对抗训练。

一般而言，对抗训练防御方法是在训练阶段添加由其自身模型所产生的对抗样本。而集成对抗训练[3]方法训练模型使用由其他模型生成的对抗样本，增加对抗样本的多样性，从而提高模型的鲁棒性。

Kannan 等人[25]介绍了一种基于 Logit pairing 的方法，Logit pairing 是基于对抗训练防御方法的扩展，在对抗训练的基础上，加入一个正则项。作者用 Logit pairing 进行了三组实验，分别对应于 MNIST，SVHN 和 ImageNet 数据集。对比实验是 PGD 对抗训练，利用 PGD 对抗样本建立训练集，进行对抗训练。实验分别测量了分类器对原始样本、白盒与黑盒场景下的对抗样本的识别准确率。结果表明经过 Logit pairing 方法的分类器具有更高的准确性。

对抗训练在训练过程中只能加入由已知攻击产生的特定类别的对抗样本，因此对抗训练防御通常不具备对其他攻击产生对抗样本的泛化能力。此外，对抗训练防御方法在训练阶段需要大量的正常样本和对抗样本，训练的成本较高，使得该方法很难在大规模数据集上使用，这是对抗训练防御方法亟待解决的难题。

（2）测试阶段修改输入样本

1）输入转换

输入转换方法不需要改变训练数据集和模型结构，而是对预测样本进行各种转换方法来减少可能存在的扰动，之后将转换后的样本输入到原模型中预测，使对抗样本重新被正确分类。Guo 等人[26]提出集成输入转换方法，对输入样本同时进行 5 种最常用的图像转换方法，降低对抗样本对模型的欺骗率。PixelDefense 防御[27]利用 PixelCNN 生成模型改变每个通道的所有像素，将对抗样本转换到正常样本，再放入原模型预测。VectorDefense[28]在分类之前将位图输入图像转换为矢量图像空间并返回，以避免被对抗结构所欺骗。

输入转换防御方法需要对预测样本进行转换处理，实验表明，目前这种方法在对抗样本预测上的误报率和漏报率较大。

2）数据压缩

Dziugaite 等人[29]发现在图像领域应用中最广泛的图像压缩技术是 JPG 图像压缩技术。受此启发，他们研究了 JPG 压缩技术对由于 FGSM 攻击扰动带来的网络模型识别率的影响。Das 等人[30]使用 JPEG 压缩方法，提出一种针对 FGSM 和 DeepFool 攻击方法的集成防御手段，但这种图像压缩技术无法面对更加强力的攻击，如 C&W 等。由于图像的局部结构中相邻像素之间具有很强的相似性和相关性，因此图像压缩可以在保留显著信息的同时减少图像的冗余信息。在此基础上，Jia 等人[31]设计了一种端到端的图像压缩模型 ComDefend 来抵御对抗样本。该模型由压缩卷积神经网络（ComCNN）和重建卷积神经网络（ResCNN）组成。该方法极大地提高了模型对各种攻击方法的抵抗力，有效地保护分类器免受对抗攻击。

Li 和 Wang[32]提出了一种新的深度去噪神经网络，用于消除对抗样本上的噪声。Liu 等人[33]运用数据压缩技术来防御对抗图像的攻击，实验结果表明所提出的防御策略只有在添加的扰动较小时才有一定的效果。

表 2 对抗样本防御方法

Table2 Defense of adversarial examples

	防御策略	防御方法	发生阶段	方法原理
针对模型	修改网络	防御蒸馏	训练阶段	用蒸馏算法为原始模型训练一个蒸馏模型
		正则化		隐藏模型梯度信息
		深度压缩网络		使用压缩自编码器的平滑惩罚项
	使用附加网络	防御通用扰动		添加外部网络重新训练，完成对抗样本校正
		基于 GAN 防御		在生成网络上训练
		对抗样本检测		判断样本是否为对抗样本
针对数据	训练阶段添加训练集	对抗训练		将生成的对抗样本注入训练集
	测试阶段修改样本	输入转换	测试阶段	使用转换方法减少测试样本中可能存在的扰动量
		数据压缩		压缩以消除对抗样本的噪声

3.3 其他对抗深度学习防御方法

3.1 和 3.2 两节讨论了从模型和数据两个角度入手的一些经典对抗样本防御方法。此外，还有一些其它应对对抗攻击的防御手段。例如，Ma 等人[34]利用对抗样本的局部本征维数（Local intrinsic dimensionality, LID）值大于正常样本的特性来识别对抗样本和正常样本，提出基于 LID 的检测方法。Buckman 等人[35]提出使用温度计编码（Thermometer encoding）将连续的输入样本进行离散化。论文[36]中提出了利用线性数据变换来抵御对抗样本攻击，即运用主成分分析作为防御机制，将高维数据投影到低维空间。Norton 等人[37]建立了一个基于网站的可视化工具，称为对抗平台，为对抗样本的生成与防御提供经验。Prakash 等人[39]将像素偏转和小波去噪技术结合提出了新的集成防御方法，利用小波域中的自适应软阈值使模型的输出平滑，该防御方法可以有效抵御最新的对抗攻击。Hosseini[40]模仿人类推理的训练方法，使分类器平滑地输出低置信度的原始标签，将对抗样本分类成空标签来拒绝对抗样本，保护黑盒系统免遭对抗样本迁移性的干扰。

4 结束语

对抗样本攻击与防御共同发展、相辅相成。本文根据对抗攻击的分类，对深度学习中的对抗样本防御方法展开了深入调查，针对模型和数据层面的几种防御策略进行了分类和介绍。随着深度学习在图像处理、自然语言处理、语音识别、医疗诊断等多个领域的深入应用，深度学习模型面临的安全威胁也日趋严重，现有的基于分类的机器学习模型极易受到对抗攻击。但另一方面，对抗样本的存在也可以激发更多关于如何防御对抗攻击的研究，从而获得具有更好鲁棒性的神经网络。尽管当前对抗样本的防御方法研究取得了一定的效果，但是也存在很多的挑战，目前面临的主要问题有四个方面：

（1）对抗攻击的防御存在对目标模型参数的依赖问题，模型使用的白盒防御策略为改变目标模型梯度传递过程，而黑盒攻击使用替代模型构造对抗样本，其本身的可迁移性属性使其在黑盒攻击中具有很好的泛化性，使模型使用的白盒防御策略失效。

（2）几乎所有的防御方法只能对有限的对抗攻击有效，不能够解决来自未知攻击带来的风险，并且很容易被不断演化的对抗样本绕过。

（3）大多数防御都是针对计算机视觉任务中的对抗样本，随着其他领域对抗样本的发展，迫切需要研究这些领域存在的问题。例如在网络空间安全领域，一些深度学习的网络空间安全应用存在的最大

删除的内容: [37,38]

问题是健壮性差，容易受到对抗攻击。

（4）正如本文介绍，对抗攻击在物理世界也十分有效，所以研究其在物理世界的防御方法也非常有必要。

由此可知，如果提高模型自身的健壮性，可以很有效地防御对抗攻击。本文对相关工作进行调研和分析，认为未来针对对抗深度学习可以从以下几个角度来进行研究：

（1）由于防御方法的局限性，针对不同类型的攻击、模型和应用，建立跨领域深度学习系统的安全性评估和防御的理论体系，结合各种对抗样本产生方法，以此来检验目标系统面对不同攻击时的健壮性，从总体上评估深度学习系统的鲁棒性，以制定更好的防御策略来应对对抗攻击。

（2）随着深度学习在各个领域的发展，将深度学习应用于网络空间安全和物理世界等安全性要求较高的应用场景时，应全面检测模型对于不同对抗攻击和隐私窃取攻击的抵抗能力，在提高模型的健壮性的同时保护其机密性。

（3）引入安全验证方法以保证对安全性有较高要求的深度学习系统的可靠性。对神经网络进行安全验证是未来可关注的一个方向。

深度学习是当前科研领域中的一大研究热点，但是其中的安全问题仍然给它的应用带来了很多隐患。如何兼顾深度学习模型的效率及其安全性是一个值得探索和研究的方向。

参考文献

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

[2] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[J]. arXiv preprint arXiv:1503.02531, 2015.

[3] Papernot N, McDaniel P, Wu X, Jha S, Swami A. Distillation as a defense to adversarial perturbations against deep neural networks[C]. 2016 IEEE Symposium on Security and Privacy (SP). IEEE, 2016: 582-597.

[4] Anil R, Pereyra G, Passos A, Ormandi R, Dahl G E, Hinton G E. Large scale distributed neural network training through online distillation[J]. arXiv preprint arXiv:1804.03235, 2018.

[5] 张思思, 左信, 刘建伟. 深度学习中的对抗样本问题[J]. 计算机学报, 2019 (8): 15. .

[6] Lyu C, Huang K, Liang H N. A unified gradient regularization family for adversarial examples[C]. 2015 IEEE International Conference on Data Mining. IEEE, 2015: 301-309.

[7] Ross A S, Doshi-Velez F. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients[C]. Thirty-second AAAI conference on artificial intelligence. 2018.

[8] Miyato T, Dai A M, Goodfellow I. Adversarial training methods for semi-supervised text classification[C]. ICLR (Poster). 2017.

[9] Miyato T, Maeda S, Koyama M, Ishii S. Virtual adversarial training: a regularization method for supervised and semi-supervised learning[J]. IEEE transactions on pattern analysis and machine intelligence, 2018, 41(8): 1979-1993.

[10] Moosavi-Dezfooli S M, Fawzi A, Uesato J, Frossard P. Robustness via curvature regularization, and vice versa[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 9078-9086.

[11] Gu S, Rigazio L. Towards deep neural network architectures robust to adversarial examples[C]. ICLR (Workshop). 2015.

[12] Osadchy M, Hernandez-Castro J, Gibson S J, Dunkelman O, Pérez-Cabo D. No Bot Expects the DeepCAPTCHA! Introducing Immutable Adversarial Examples with Applications to CAPTCHA[J]. IACR Cryptology ePrint Archive, 2016, 2016: 336.

[13] Liao F, Liang M, Dong Y, Pang T, Hu X, Zhu J. Defense against adversarial attacks using high-level representation guided denoiser[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 1778-1787.

[14] Akhtar N, Liu J, Mian A. Defense against universal adversarial perturbations[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 3389-3398.

[15] Lee H, Han S, Lee J. Generative adversarial trainer: Defense to adversarial perturbations with gan[J]. arXiv preprint arXiv:1705.03387, 2017.

[16] Shen S, Jin G, Gao K, Zhang Y. Ape-gan: Adversarial perturbation elimination with gan[J]. arXiv preprint arXiv:1707.05474,

2017.
[17] Samangouei P, Kabkab M, Chellappa R. Defense-gan: Protecting classifiers against adversarial attacks using generative models[J]. arXiv preprint arXiv:1805.06605, 2018.
[18] Xu W, Evans D, Qi Y. Feature squeezing: Detecting adversarial examples in deep neural networks[C]. Proceedings of. The Network and Distributed System Security Symposium. (NDSS), 2018.
[19] 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. ACM, 2017: 135-147.
[20] Lu J, Issaranon T, Forsyth D. Safetynet: Detecting and rejecting adversarial examples robustly[C]. Proceedings of the IEEE International Conference on Computer Vision. 2017: 446-454.
[21] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[C]. ICLR (Poster). 2015.
[22] Huang R, Xu B, Schuurmans D, et al. Learning with a strong adversary. 2015[J]. arXiv preprint arXiv:1511.03034.
[23] Kurakin A, Goodfellow I, Bengio S. Adversarial machine learning at scale[J]. arXiv preprint arXiv:1611.01236, 2016.
[24] Dong Y, Su H, Zhu J, Bao F. Towards interpretable deep neural networks by leveraging adversarial examples[J]. arXiv preprint arXiv:1708.05493, 2017.
[25] Kannan H, Kurakin A, Goodfellow I. Adversarial logit pairing[J]. arXiv preprint arXiv:1803.06373, 2018.
[26] Guo C, Rana M, Cisse M, Van Der Maaten, L. Countering adversarial images using input transformations[C]. ICLR (Poster). 2018
[27] Song Y, Kim T, Nowozin S, Ermon S, Kushman N. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples[C]. ICLR (Poster). 2018
[28] Kabilan V M, Morris B, Nguyen A. VectorDefense: Vectorization as a Defense to Adversarial Examples[J]. arXiv preprint arXiv:1804.08529, 2018.
[29] Dziugaite G K, Ghahramani Z, Roy D M. A study of the effect of jpg compression on adversarial images[J]. arXiv preprint arXiv:1608.00853, 2016.
[30] Das N, Shanbhogue M, Chen S T, Hohman F, Chen L, Kounavis M E, Chau D H. Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression[J]. arXiv preprint arXiv:1705.02900, 2017.
[31] Jia X, Wei X, Cao X, Foroosh, H. ComDefend: An Efficient Image Compression Model to Defend Adversarial Examples[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 6084-6092.
[32] Li Y, Wang Y. Defense Against Adversarial Attacks in Deep Learning[J]. Applied Sciences, 2019, 9(1): 76.
[33] Liu T, Liu Z, Liu Q, Wen W. Enhancing the Robustness of Deep Neural Networks from" Smart" Compression[C]. 2018 IEEE Computer Society Annual Symposium on VLSI (ISVLSI). IEEE, 2018: 528-532.
[34] Ma X, Li B, Wang Y, Erfani S M, Wijewickrema S, et al. Characterizing adversarial subspaces using local intrinsic dimensionality[J]. arXiv preprint arXiv:1801.02613, 2018.
[35] Buckman J, Roy A, Raffel C, et al. Thermometer encoding: One hot way to resist adversarial examples[J]. 2018.
[36] 吴嫚, 刘笑嶂. 基于 PCA 的对抗样本攻击防御研究[J]. 海南大学学报:自然科学版, 2019(2):134-139.
[37] Norton A P, Qi Y. Adversarial-Playground: A visualization suite showing how adversarial examples fool deep learning[C]//2017 IEEE Symposium on Visualization for Cyber Security (VizSec). IEEE, 2017: 1-4.
[38] Norton A, Qi Y. Adversarial-Playground: A Visualization Suite for Adversarial Sample Generation[J]. arXiv preprint arXiv:1706.01763, 2017.
[39] Prakash A, Moran N, Garber S, DiLillo A, Storer J. Deflecting adversarial attacks with pixel deflection[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8571-8580.
[40] Hosseini H, Chen Y, Kannan S, Zhang B, Poovendran R. Blocking transferability of adversarial examples in black-box learning systems[J]. arXiv preprint arXiv: 1703.04318, 2017.

作者简介

张嘉楠，北京交通大学，硕士在读；主要研究方向和关注领域：网络空间安全。

赵镇东，北京交通大学，硕士在读；主要研究方向和关注领域：网络空间安全。

宣晶，北京京投卓越科技发展有限公司，总经理，北京京投信安科技发展有限公司，董事长；主要研究方向和关注领域：网络空间安全。

常晓林，香港科技大学，博士，北京交通大学，教授；主要研究方向和关注领域：可信智能软件，网络安全和云边计算。