深度学习的对抗攻击与鲁棒性测评

(申请清华大学工学博士学位论文)

培 养 单 位： 计算机科学与技术系

学　　科： 计算机科学与技术

研 究 生： 董 胤 蓬

指 导 教 师： 朱　军 教 授

二〇二一年十二月

Adversarial Attacks and Robustness Evaluation in Deep Learning

Dissertation Submitted to

Tsinghua University

in partial fulfillment of the requirement

for the degree of

Doctor of Philosophy

in

Computer Science and Technology

by

Dong Yinpeng

Dissertation Supervisor: Professor Zhu Jun

December, 2021

学位论文公开评阅人和答辩委员会名单

公开评阅人名单

张长水	教授	清华大学
王立威	教授	北京大学

答辩委员会名单

主席	刘成林	研究员	中国科学院自动化研究所
委员	张钹	教授	清华大学
	张长水	教授	清华大学
	朱军	教授	清华大学
	陶建华	研究员	中国科学院自动化研究所
	胡晓林	副教授	清华大学
秘书	李建民	副研究员	清华大学

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：

清华大学拥有在著作权法规定范围内学位论文的使用权，其中包括：（1）已获学位的研究生必须按学校规定提交学位论文，学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文；（2）为教学和科研目的，学校可以将公开的学位论文作为资料在图书馆、资料室等场所供校内师生阅读，或在校园网上供校内师生浏览部分内容；（3）根据《中华人民共和国学位条例暂行实施办法》及上级教育主管部门具体要求，向国家图书馆报送相应的学位论文。

本人保证遵守上述规定。

作者签名：_____　　导师签名：_____

日　　期：_____　　日　　期：_____

摘 要

以深度学习为代表的人工智能技术在计算机视觉、语音识别等众多领域均取得了显著进展，规模化应用已现曙光。但现有深度学习模型存在鲁棒性不足的问题，很容易被攻击者恶意构造的对抗样本欺骗，产生错误的预测结果。深度学习鲁棒性的不足已被证实会对一些与安全密切相关的领域带来威胁。同时，这一问题也阻碍了深度学习的进一步发展。对抗攻击与鲁棒性测评作为深度学习鲁棒性研究中的重要方向，旨在面向不同场景高效地生成对抗样本并针对深度学习模型的鲁棒性进行全面的测评。此方面研究有助于发现深度学习模型的脆弱性，比较不同模型的鲁棒性，以及发展更加鲁棒的深度学习模型。

对抗攻击与鲁棒性测评方面的研究仍然存在一些亟待解决的问题。第一，现有对抗攻击方法在无法获取模型结构和参数信息的黑盒场景下攻击成功率与效率较低，阻碍了模型脆弱性机理的分析。第二，现有对抗攻击方法生成对抗样本的多样性不足，限制了基于这些对抗样本训练所得模型的鲁棒性。第三，目前对抗鲁棒性测评的研究工作较为欠缺，研究者难以有效评估不同深度学习模型的鲁棒性以及对抗攻防算法的有效性。为解决上述关键问题，本文构建对抗攻防测评基准与平台，并面向不同场景研发高效对抗攻击算法。主要创新点概括如下：

1. 针对黑盒迁移攻击成功率较低的问题，提出动量迭代法与平移不变对抗攻击方法，分别通过引入动量项以及对一组经过平移变换的图片生成对抗样本，大幅提高黑盒迁移攻击成功率，为理解深度学习模型的脆弱性机理及发现模型的安全漏洞奠定了理论和方法基础。
2. 针对黑盒决策攻击效率较低的问题，面向人脸识别场景，提出进化攻击方法，通过建模搜索方向的局部几何结构和降低搜索空间的维度有效提升黑盒决策攻击的效率，为挖掘人脸识别模型的安全漏洞奠定了理论和方法基础。
3. 针对对抗训练模型鲁棒性不足的问题，提出对抗分布训练，利用对抗分布刻画原始样本周围多样化的对抗样本，并通过三种对抗攻击方式参数化建模对抗分布，为构建更加鲁棒的深度学习模型奠定了理论和方法基础。
4. 针对对抗鲁棒性测评较为欠缺的问题，面向图像分类任务构建对抗鲁棒性测评基准，采用鲁棒性曲线针对多个典型的对抗攻防算法进行公平、全面的鲁棒性测评，为今后对抗攻防模型及算法的开发奠定了测评基础。

关键词：深度学习；对抗样本；对抗攻击；对抗防御；鲁棒性测评

Abstract

Artificial intelligence technologies, especially deep learning, have made significant progress in numerous fields such as computer vision and speech recognition, while large-scale applications are now dawning. However, the existing deep learning models have the problem of insufficient robustness that they can be easily deceived by the adversarial examples maliciously generated by adversaries to produce wrong predictions. The lack of robustness of deep learning has been proven to pose threats to some areas closely related to security. Meanwhile, this problem hinders the further development of deep learning. Adversarial attacks and robustness evaluation are important directions of the research on deep learning robustness, aiming to efficiently generate adversarial examples under different scenarios and conduct comprehensive robustness evaluation of deep learning models. The research in this area helps to identify the vulnerabilities of deep learning models, compare the robustness of different models, and develop more robust deep learning models.

The research on adversarial attacks and robustness evaluation still has some problems that need to be solved urgently. First, the existing adversarial attack methods exhibit low attack success rate and inefficiency under the black-box scenarios where model structure and parameters cannot be obtained, which hinders the analysis of model's vulnerability mechanism. Second, the diversity of adversarial examples generated by the existing adversarial attack methods is insufficient, which limits the robustness of the models trained on these adversarial examples. Third, the current research on adversarial robustness evaluation is relatively lacking, such that it is difficult for researchers to effectively evaluate the robustness of different deep learning models and the effectiveness of adversarial attack and defense algorithms. To solve the above key problems, this dissertation builds a benchmark and a platform for evaluating adversarial attacks and defenses, and develops efficient adversarial attack algorithms under different scenarios. The main contributions are summarized as follows:

1. For the problem of low success rate of black-box transfer-based attacks, a momentum iterative method and a translation-invariant adversarial attack method are proposed. They introduce a momentum term and adopt a set of translated images, respectively, for generating adversarial examples, which greatly improve the suc-

cess rate of black-box transfer-based attacks. They lay the theoretical and methodological foundation for understanding the vulnerability mechanism of deep learning models and discovering model's security holes.

2. For the problem of inefficiency of black-box decision-based attacks, an evolutionary attack method is proposed for face recognition. It models the local geometry of the search direction and reduces the dimension of the search space in black-box decision-based attacks to effectively improve the efficiency of black-box decision-based attacks, which lays the theoretical and methodological foundation for digging the security holes of face recognition models.
3. For the problem of insufficient robustness of adversarial training models, an adversarial distributional training is proposed to use adversarial distributions to characterize the diverse adversarial examples around the original one, which parameterizes the adversarial distributions through three adversarial attacks, laying the theoretical and methodological foundation for building more robust deep learning models.
4. For the problem of the lack of adversarial robustness evaluations, an adversarial robustness benchmark is constructed for image classification, which uses robustness curves to conduct fair and comprehensive robustness evaluation of many typical adversarial attack and defense algorithms. It lays the evaluation foundation for further development of adversarial attack and defense algorithms.

Keywords: deep learning; adversarial example; adversarial attack; adversarial defense; robustness evaluation

目 录

插图和附表清单

符号和缩略语说明

AT	对抗训练（Adversarial Training）
FGSM	快速梯度符号法（Fast Gradient Sign Method）
BPDA	后向传递可微近似（Backward Pass Differentiable Approximation）
EoT	期望变换（Expectation over Transformation）
NES	自然进化策略（Natural Evolution Strategy）
BIM	基础迭代法（Basic Iterative Method）
MIM	动量迭代法（Momentum Iterative Method）
PGD	投影梯度下降法（Projected Gradient Descent）
TI-FGSM	平移不变快速梯度符号法（Translation-Invariant Fast Gradient Sign Method）
TI-MIM	平移不变动量迭代法（Translation-Invariant Momentum Iterative Method）
DIM	多样输入法（Diverse Inputs Method）
TI-DIM	平移不变多样输入法（Translation-Invariant Diverse Inputs Method）
CMA-ES	协方差矩阵自适应进化策略（Covariance Matrix Adaptation Evolution Strategy）
ADT	对抗分布训练（Adversarial Distributional Training）
EXP	显式对抗分布（EXPlicit adversarial distribution）
EXP-AM	均摊显式对抗分布（AMortized EXPlicit adversarial distribution）
IMP-AM	均摊隐式对抗分布（AMortized IMPlicit adversarial distribution）
$\mathbb{R}$	实数集
$\mathcal{X}$	输入数据空间
$\mathcal{Y}$	输出类别空间
x	原始样本
y	真实类别
x^*	对抗样本
y^*	攻击目标类别
C	分类器
$\|\cdot\|_p$	ℓ_p 范数
ϵ	最大扰动规模

$J(\cdot)$	损失函数
$C(\cdot)$	对抗判别准则
$\mathbb{I}(\cdot)$	指示函数
$\mathcal{N}$	高斯分布
$\mathbb{E}(\cdot)$	期望

第 1 章　绪论

深度学习的快速发展推动了人工智能在计算机视觉、语音识别、自然语言处理等多个领域的广泛应用。随着人工智能应用的范围与规模不断扩大，深度学习模型的鲁棒性与安全性问题成为了人们关注的焦点。在真实数据上表现良好的深度学习模型很容易被攻击者恶意构造的对抗样本欺骗，这会对安全性要求较高的深度学习应用带来严重的威胁。本文将针对深度学习的对抗攻击与鲁棒性测评开展研究，旨在提供高效的对抗攻击算法，并构建公平、全面的对抗鲁棒性测评基准。本文的研究内容在实际中可用于及时发现深度学习模型的脆弱性，比较不同模型的鲁棒性，并从原理上加深对深度学习机理的理解，以及发展鲁棒的深度学习理论与方法。

本章将介绍本文的研究背景与意义，详述对抗样本的定义与分类，总结国内外研究现状与关键研究问题，概括本文的研究内容与主要贡献，并给出本文的组织结构。

1.1　研究背景与意义

近年来，人工智能的飞速发展正在深刻改变着人类的生产和生活方式。相比于传统的机器学习方法，深度学习[1-2]借助层次化的深度神经网络提取对数据内容更加准确有效的特征表示，成为人工智能领域的代表方法。随着算法的演进、数据规模的扩大以及计算能力的增长，深度学习在图像分类[3-6]、目标检测[7-9]、人脸识别[10-12]、机器翻译[13-14]等实际应用中取得了显著进展，其性能甚至超越了人类。例如，基于深度神经网络的人工智能围棋系统 AlphaGo 战胜了人类顶尖选手[15]，这场人机大战引发全球广泛关注。最新研究将深度学习应用于蛋白质结构预测[16-17]，取得了出色的结果，体现出深度学习对于科学发展的推动作用。

尽管深度学习在实际应用中具有优越的性能，但是其在对抗场景下的鲁棒性（robustness）①存在严重不足。研究发现，在真实数据上表现良好的深度学习模型很容易被攻击者恶意构造的对抗样本（adversarial example）[18-20]欺骗，导致模型预测失准。如图1.1所示，针对图像分类任务，攻击者可以向原始图片中添加微小的扰动生成对抗样本，使人眼难以发现对抗样本与原始样本的区别，但是深度学习模型会以很高的概率将对抗样本错分为其他类别。产生此现象的原因是深度学习模型通过复杂的网络结构建模数据特征间的统计规律，而并未真正获取数据本质

① 也有文献译为稳健性、健壮性等。

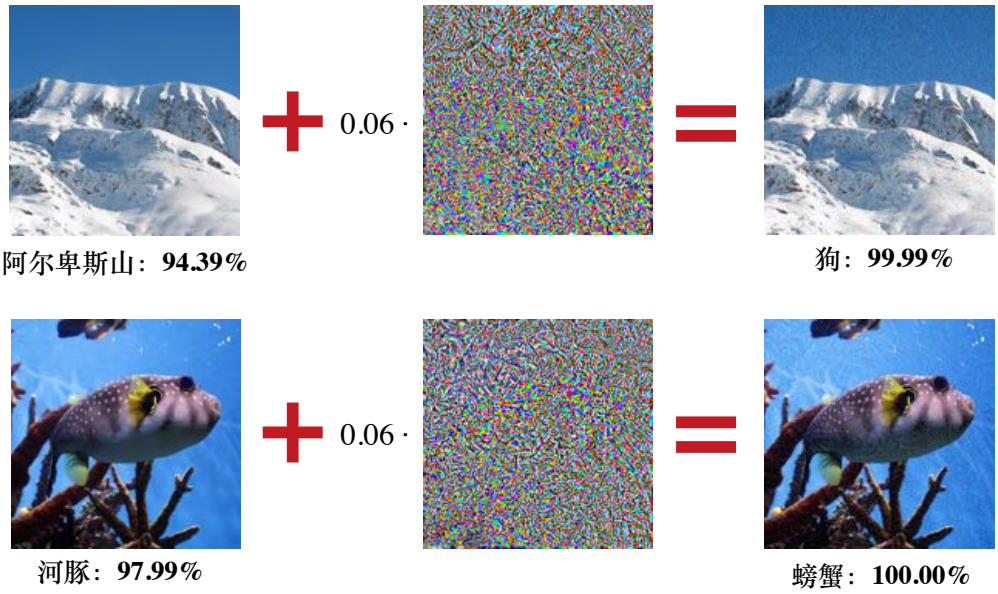
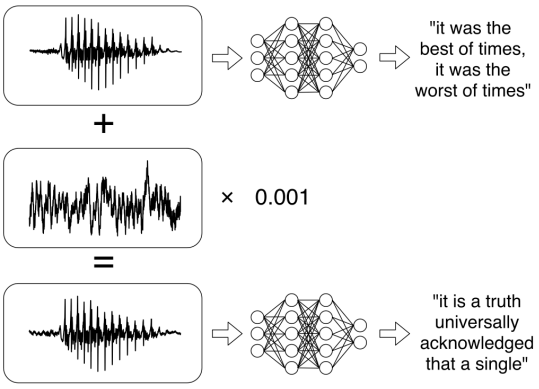


图 1.1　图像分类任务中对抗样本示例

的特征及因果关系，这就导致了深度学习模型很容易被人眼不可见的对抗扰动欺骗[21]。随着人工智能技术的创新，其与社会发展也进一步融合，在军事、工业、金融、医疗、人脸识别、自动驾驶等不同领域中人工智能技术均被广泛应用。这些领域也与社会安全密切相关，这就使得对抗样本带来的威胁愈发突出，在现实生活中可能会造成财产损失甚至威胁人身安全。同时，由于鲁棒性的不足，用户难以信赖人工智能系统做出的决策，这也阻碍了人工智能的进一步发展。

深度学习的鲁棒性与安全性问题层出不穷，针对图像数据生成的对抗样本只是其中的一种典型场景。除图像数据外，攻击者还可以针对视频[22]、语音[23-24]、文本[25-26]等各种类型的数据生成对抗样本，并相应地欺骗不同领域的深度学习模型，如图1.2所示。此外，深度学习模型在数据投毒攻击[27-29]、后门攻击[30-31]、数据自然变换[32-33]等其他场景下也存在鲁棒性问题。随着研究愈发深入，深度学习的缺陷不断暴露，其鲁棒性与安全性问题也得以全面分析。尽管如此，本文聚焦深度学习在处理图像数据时面临的对抗鲁棒性问题，即对于对抗样本的鲁棒性。此方向的研究对于探索深度学习的鲁棒性具有一定代表性。在不引起歧义的情况下，本文中的鲁棒性即代表对抗鲁棒性。

鉴于对抗样本会对深度学习模型及应用带来潜在的安全威胁，更多研究者开始关注深度学习的鲁棒性，其也成为人工智能领域的一大热点研究方向。该方向随着攻击与防御的博弈而不断发展[34-35]。一方面，对抗攻击（adversarial attack）研究在不同场景下生成对抗样本的高效算法，旨在发现深度学习模型中存在的缺陷[19-20,36]。另一方面，对抗防御（adversarial defense）研究增强模型鲁棒性的方式，以减轻对抗攻击所带来的危害，发展安全可靠的人工智能[20,37-38]。这种良性



(a) 语音数据的对抗样本[23]

Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

(b) 文本数据的对抗样本[25]

图 1.2　语音和文本数据的对抗样本示例

博弈促使越来越多的对抗攻击与防御算法被提出，而如何公平合理地测评深度学习模型的鲁棒性以及对抗攻防算法的有效性成为一个新的研究问题[39]，对反映此领域的进展具有重要作用。

综上所述，深度学习鲁棒性的研究对构建安全可靠的人工智能系统具有重要的理论与现实意义。

1.1.1　研究价值

本文将针对深度学习的对抗攻击与鲁棒性测评开展研究，这是深度学习鲁棒性研究中的重要方向，具有以下研究价值。

首先，对抗攻击可用于发现深度学习模型在不同场景下的脆弱性。随着科技的不断发展，人工智能逐步渗透到出行、医疗、金融等与人们日常生活密切相关的领域中。以深度学习模型为代表的人工智能技术在真实世界中需要处理各种复杂信息，如果其在部署前未经过充分的安全性测试，可能会对实际应用带来严重威胁。由于深度学习模型结构复杂且处理的数据维度较高（如图片），从理论上分析其脆弱性十分困难，而对抗攻击提供了一种便捷的方式。例如，2014 年 Szegedy 等人[19]首次通过对抗攻击发现深度学习模型中存在对抗样本，证实了深度学习的脆弱性。Jing 等人[40]通过对抗攻击发现商用自动驾驶系统中的车道检测模型十分脆弱，很容易被真实世界中的扰动欺骗。因此，对抗攻击对发现深度学习模型在不同场景下的脆弱性具有重要意义。

其次，对抗攻击与鲁棒性测评可用于比较不同深度学习模型的鲁棒性。随着深度学习的鲁棒性问题受到广泛关注，如何公平合理地测评不同模型的鲁棒性成为重要的研究问题[39,41]。相比于直接计算模型在真实测试数据上的准确率，模型的鲁棒性提供了另一个评估维度，有助于更加全面地比较不同模型的性能。同理，深度学习模型的结构复杂，鲁棒性的精确计算较为困难，因此需要采用对抗攻击

生成的对抗样本测试模型的鲁棒性。值得强调的是，由于大量对抗防御算法被相继提出，针对具有防御机制的深度学习模型进行公平、全面的鲁棒性测评具有重要意义[39,42]，可用于比较不同防御算法的有效性。综上所述，构建强大的对抗攻击算法与全面的鲁棒性测评基准能够更加精准地比较深度学习模型与算法的鲁棒性，具有重大研究价值。

再次，通过对深度学习鲁棒性的分析与测评可以加深对深度学习机理的理解。深度学习模型结构复杂，可解释性存在不足，因此往往被当作一个黑箱模型使用。在很大程度上，对深度学习机理的理解不足是导致模型在对抗攻击下鲁棒性欠缺的重要原因。通过对抗攻击发现模型的脆弱性，可以促进对深度学习机理的研究，有助于分析深度学习的缺陷。例如，已有研究发现深度学习模型的准确率与鲁棒性之间存在相互制约的关系[43]，说明其训练方式存在不足。深度学习的鲁棒性与可解释性之间也存在一定关联[44-45]，通过构建鲁棒的深度学习模型可以提升其可解释性。这些基于深度学习鲁棒性的分析与测评的研究加深了对深度学习机理的理解，为构建安全可靠的人工智能提供理论和方法支持。

最后，对抗攻击对于增强深度学习模型的鲁棒性也具有重要作用。对抗攻击可以发现深度学习模型的脆弱性，因此启发了不同类型的对抗防御，旨在消除对抗样本所带来的威胁。对抗攻击生成的对抗样本还可以作为训练数据直接增强模型的鲁棒性。采用对抗样本进行训练的方法被称为对抗训练（Adversarial Training，AT）[20,37]，是目前最有效的防御方式之一。相关研究通过设计合理的对抗攻击方法有效提升对抗训练模型在不同任务上的鲁棒性，保障人工智能模型和系统在一些高风险任务（如军事、金融、自动驾驶等）上的应用。因此，对抗攻击是保证人工智能系统安全可靠的重要方式。

1.2　对抗样本的定义与分类

本节将以图像分类任务为例介绍对抗样本的形式化定义与攻击者所处威胁模型（threat model）的分类。

1.2.1　对抗样本的定义

对于图像分类任务，攻击者可以向原始样本 x 中添加微小的扰动生成对抗样本 x^*，使得深度学习模型将对抗样本 x^* 分类错误，而人眼无法发现对抗样本与原始样本的区别[19-20]。给定基于深度学习的分类器模型 $C(x)$，假设原始样本 x 的真实类别为 y，对抗样本可以形式化表示为：

$$C(x^*) \neq y, \qquad \text{s.t. } D(x^*, x) \leqslant \epsilon, \tag{1.1}$$

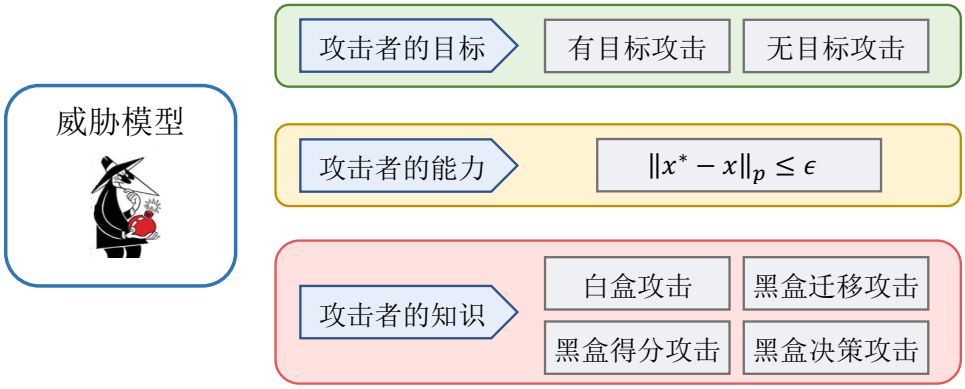


图 1.3　威胁模型分类示意图

其中 $\mathcal{D}(\cdot,\cdot)$ 代表距离度量函数，ϵ 代表在距离度量 $\mathcal{D}$ 下的扰动范围。公式（1.1）中的第一项表示分类器将对抗样本 x^* 分类错误，第二项表示对抗样本 x^* 与原始样本 x 在度量 $\mathcal{D}$ 下的距离不超过 ϵ，使人眼无法察觉对抗样本中添加的扰动。

对抗攻击通过定义攻击目标函数并采用不同优化方法对其进行求解，以生成满足公式（1.1）中两个条件的对抗样本。

1.2.2　威胁模型的分类

威胁模型描述攻击者针对深度学习模型生成对抗样本时所处的场景。根据威胁模型的不同设置，攻击者需要采用相应的对抗攻击方法生成对抗样本。清晰准确地定义威胁模型是进行鲁棒性测评的前提[39]。在一般情况下，威胁模型包含攻击者的目标、攻击者的能力和攻击者的知识三个方面，如图1.3所示。

1.2.2.1　攻击者的目标

攻击者在生成对抗样本时可能存在不同的攻击目标，以欺骗深度学习模型达到特定的目的。在图像分类任务中，可以根据攻击者的不同目标将对抗攻击分为有目标攻击（targeted attack）和无目标攻击（untargeted attack）。其中，有目标攻击是使得模型将对抗样本分类为指定的目标类别 y^*，即 $C(x^*) = y^*$；而无目标攻击是使得模型将对抗样本分类错误，即 $C(x^*) \neq y$。可以看到，如果有目标攻击成功，无目标攻击也会成功，因此有目标攻击相比于无目标攻击更具挑战。

在其他任务中攻击者也存在不同的目标。例如，在人脸识别任务中，对抗攻击可以被分为躲避攻击（dodging attack）和伪装攻击（impersonation attack），第4.3.2节会做进一步介绍。

1.2.2.2　攻击者的能力

为了使对抗样本中添加的扰动难以察觉，攻击者所具备的能力需要被合理地约束。如果不限制其能力，攻击者可以向原始样本中添加非常大的扰动甚至改变图片的语义，这就超出了深度学习鲁棒性的研究范畴。如公式（1.1）所示，攻击者在生成对抗样本时做出改变的大小通常会被距离度量 $\mathcal{D}$ 约束，使得生成的对抗样本满足 $\mathcal{D}(x^*, x) \leqslant \epsilon$。在一般情况下，距离度量 $\mathcal{D}$ 下的微小变化不会改变数据的真实类别，即对抗样本的类别与原始样本一致。

针对图像数据，距离度量 $\mathcal{D}$ 通常选取为 ℓ_p 范数，即对抗样本需要满足 $\|x^* - x\|_p \leqslant \epsilon$。$\ell_p$ 范数也具备很好的数学性质与意义。例如，采用 ℓ_∞ 范数意味着攻击者对图片中每个像素的修改不能超过 ϵ。虽然真实世界中的很多威胁并不在 ℓ_p 范数的限制范围内[32,46-47]，但是 ℓ_p 范数的定义较为直观且被大多数研究采用，因此本文针对 ℓ_p（包括 ℓ_∞ 和 ℓ_2）范数限制下的对抗样本进行研究。

1.2.2.3　攻击者的知识

威胁模型的最后一个方面描述了攻击者对深度学习模型所能获取的知识。根据获取知识的不同，对抗攻击可以被分为白盒攻击（white-box attack）、黑盒迁移攻击（black-box transfer-based attack）、黑盒得分攻击（black-box score-based attack）和黑盒决策攻击（black-box decision-based attack），其中后三种攻击类型均属于黑盒攻击（black-box attack）。

在能够获取模型所有信息（包括结构、参数、梯度等）的情况下，攻击者可以利用白盒攻击生成对抗样本。白盒攻击通常采用基于梯度的方法[20,48-49]优化攻击目标函数。在模型具有防御机制的情况下，攻击者可以针对特定的防御设计适应性攻击（adaptive attack）[36,42]，以攻破防御模型。尽管白盒攻击效果显著，但是攻击者往往无法获取实际应用中的深度学习模型，这导致白盒攻击的使用范围受到限制。

为了欺骗实际应用中的黑盒模型，攻击者可以采用三种不同的黑盒攻击方式生成对抗样本。黑盒迁移攻击针对替代模型生成对抗样本，并利用对抗样本的迁移能力（transferability）[50-51]欺骗黑盒模型，其中迁移能力是指对一个模型生成的对抗样本也有一定概率欺骗未知的黑盒模型。黑盒得分攻击通过查询（query）的方式获取黑盒模型对输入数据的预测概率分布，并利用该信息生成对抗样本。黑盒决策攻击同样可以查询黑盒模型，但只能获取模型对输入数据的预测类别，因此黑盒决策攻击相比于黑盒得分攻击更具挑战。

1.3　国内外研究现状

基于上述对抗样本的定义与分类，本节将综述深度学习鲁棒性的国内外研究现状，具体内容包括对抗攻击、对抗防御和鲁棒性测评三个方面。

1.3.1　对抗攻击

自 2014 年 Szegedy 等人[19]首次发现深度学习模型会被对抗样本欺骗后，针对深度学习模型的对抗攻击逐渐成为一大研究热点。本小节将分别介绍白盒攻击与黑盒攻击的典型方法，并阐述深度学习模型在真实世界中可能面临的安全风险。

1.3.1.1　白盒攻击

由于白盒攻击假设深度学习模型的信息完全公开，攻击者可以计算攻击目标函数对于输入数据的梯度，进而采用基于梯度的方法生成对抗样本。白盒攻击的代表方法是 Goodfellow 等人[20]提出的*快速梯度符号法*（Fast Gradient Sign Method，FGSM）。该方法通过单步梯度更新生成对抗样本，具有简单高效的特点。后续工作[37,48-49,52]采用基于多步梯度迭代的方法生成对抗样本，可以提升白盒攻击效果，在使用较小扰动的情况下仍可以达到接近 100% 的攻击成功率。

尽管基于梯度的白盒攻击方法可以有效攻破正常训练的深度学习模型，但很容易被防御。例如，对输入图片进行 JPEG 压缩的防御方式[53-54]会导致基于梯度的攻击方法无法生成对抗样本，这是因为 JPEG 压缩的过程无法求导。这种现象被称为*混淆梯度*（obfuscated gradients）[36]。为了解决此问题，Athalye 等人[36,55]提出了*后向传递可微近似*（Backward Pass Differentiable Approximation，BPDA）与*期望变换*（Expectation over Transformation，EoT）两种适应性攻击方法，分别用于模型的梯度无法直接计算或存在随机性的情况。然而，适应性攻击需要针对不同的模型手动设计，其灵活性较差。

白盒攻击的主要问题是在实际应用中模型的梯度信息难以获取。例如，针对手机的人脸解锁系统，攻击者无法得到其内部使用的人脸识别模型，因而不能计算模型的梯度。在实际应用中，攻击者只能采用黑盒攻击方式生成对抗样本，说明黑盒攻击的适用范围更广。

1.3.1.2　黑盒攻击

黑盒攻击包括黑盒迁移攻击、黑盒得分攻击和黑盒决策攻击三种类型，本小节将详细介绍每一类黑盒攻击的典型方法与挑战。

黑盒迁移攻击利用对抗样本的迁移能力欺骗黑盒模型。在此场景下，攻击者

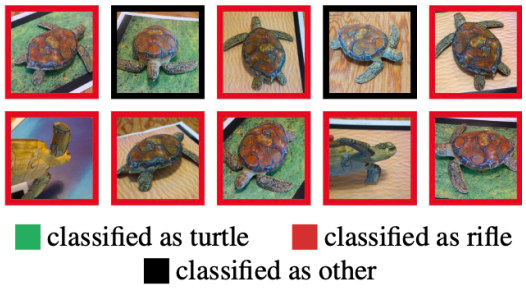


图 1.4　真实世界中的 3D 对抗样本[55]

图 1.5　真实世界中的对抗眼镜[65]

首先收集或训练本地的替代模型，然后采用上述白盒攻击方法针对替代模型生成对抗样本，最后直接使用生成的对抗样本攻击黑盒模型[50-51]。这个过程看似简单，但在实际应用中黑盒迁移攻击面临成功率较低的问题。一些研究[51,56-58]通过设计更有效的对抗样本生成方法与替代模型选取策略提高黑盒迁移攻击成功率。

然而，在模型较为复杂或输入维度较高时，攻击者难以找到合适的替代模型。在实际应用中，尽管攻击者无法获取被攻击模型的信息，但往往可以对黑盒模型进行查询得到模型对输入数据的预测结果，进而利用模型的预测结果生成对抗样本。黑盒得分攻击可以获取黑盒模型对输入数据的预测概率分布，然后通过黑盒优化（black-box optimization）估计模型的梯度以生成对抗样本[59-61]。Chen 等人[59]通过有限差分（finite difference）方法估计模型对于输入数据中每一个维度的导数，并使用随机坐标下降生成对抗样本。其他一些方法[60-61]基于自然进化策略（Natural Evolution Strategy，NES）[62]在高斯分布下优化对抗样本。

与黑盒得分攻击相比，黑盒决策攻击只能获取黑盒模型对输入数据的预测类别。Ilyas 等人[60]通过模型的预测类别估计其预测概率分布，将黑盒决策攻击转换为黑盒得分攻击，进而采用黑盒得分攻击方法生成对抗样本。Brendel 等人[63]提出基于启发式搜索策略的边界攻击（Boundary attack）方法，在搜索过程中不断减小对抗样本与原始样本之间的距离。Cheng 等人[64]将黑盒决策攻击建模为一个连续优化问题，并利用梯度估计方法进行优化。黑盒得分攻击与黑盒决策攻击面临的主要问题是其攻击效率较低，通常需要对黑盒模型进行大量查询才能生成扰动较小对抗样本，在实际中代价较高。

1.3.1.3　真实世界中的对抗攻击

目前大部分对抗攻击的研究主要集中于数字世界中的对抗样本，近期的研究表明对抗样本同样可以存在于真实世界中[48]。相比于数字世界中的对抗攻击，真实世界中的对抗攻击需要解决对环境变化、空间相对位置变化、对抗扰动的物理可实现性等挑战。Athalye 等人[55]提出期望变换（EoT）方法，将真实世界中可能存在的变换加入到攻击目标函数中，并优化攻击目标函数生成对抗样本。如图1.4所

示，该方法利用 3D 打印生成真实世界中的对抗样本，使得摄像机采集的图片可以欺骗深度学习模型。Sharif 等人[65]针对人脸识别任务进行研究，在对抗样本生成的过程中加入物理变换及打印机的色彩误差。如图1.5所示，该方法将生成的对抗样本打印到眼镜上，使得佩戴对抗眼镜的攻击者可以欺骗真实世界中的人脸识别系统。此外，针对路牌识别[66]、行人检测[67]、自动驾驶[40,68]等系统，攻击者也可以在真实世界中生成对抗样本，欺骗不同应用中的深度学习模型。真实世界中的对抗攻击也会对深度学习模型的实际应用带来更加现实的安全威胁。

1.3.2　对抗防御

由于对抗样本会对深度学习模型带来潜在的安全威胁，研究者围绕对抗防御开展了大量研究工作，多种类型的对抗防御方法被相继提出。本小节主要介绍鲁棒训练（robust training）、图像变换（image transformation）、随机化（randomization）、模型集成（model ensemble）和可证实防御（certified defense）五种防御类型。值得注意的是，这五种防御类型并不是相互排斥的，即某些防御模型可能同时包含多种类型的防御技术。

鲁棒训练是指通过改进深度学习模型的训练方式增强其鲁棒性的防御。对抗训练[20,37,69]作为鲁棒训练的代表方法，采用对抗攻击生成的对抗样本训练深度学习模型。鲁棒训练也可通过设计训练损失函数[70]或正则化[71-73]的方式实现，代表性方法包括控制模型的 Lipschitz 系数[71]、增大最小扰动的范数[73]等。

图像变换是指对模型的输入数据进行预处理的防御。图像变换旨在消除对抗扰动对模型的干扰，包括 JPEG 压缩[53-54]、图像去噪[38]等方法。此外，一些研究[74-75]认为对抗样本与真实样本的分布不同，提出利用深度生成模型将对抗样本映射到真实数据分布中进行防御。图像变换防御会引起混淆梯度[36]，很容易被适应性攻击攻破。

随机化是指向输入数据[76-77]或深度学习模型[78]中加入随机性的防御。此类防御使得攻击目标函数对于输入数据的梯度具有随机性，防止基于梯度的对抗攻击生成对抗样本。随机化防御会被基于期望变换（EoT）的适应性攻击攻破[36]。

模型集成是指对多个深度学习模型进行集成的防御，以构建更加有效的防御模型。Liu 等人[78]提出随机自集成（random self-ensemble）方式，向模型中加入随机性并将不同随机模型的预测结果进行平均。Pang 等人[79]提出促进模型多样性的正则化方法并将具有多样性的模型进行集成。

可证实防御是指通过理论分析保证模型在某个扰动范围内不会存在对抗样本的防御[80-84]。目前，可证实防御[85-86]可用于大规模图像数据集 ImageNet[87]上，展示出此类防御方式的可扩展性。

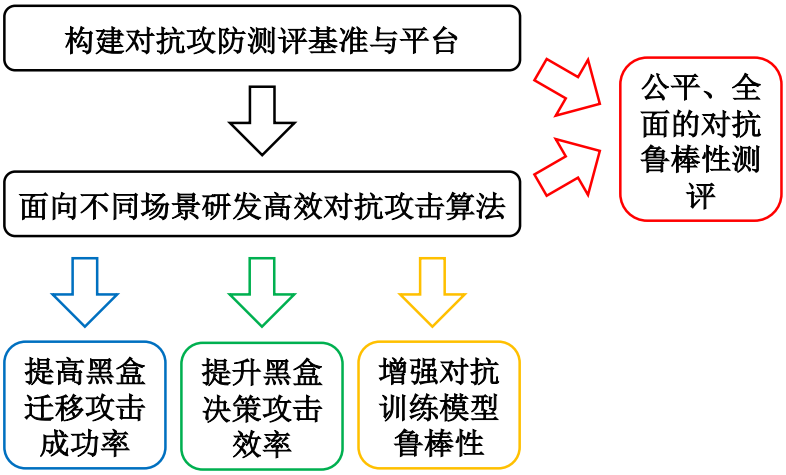


图 1.6　对抗攻击与鲁棒性测评中有待解决的关键研究问题

在多种对抗防御类型中，对抗训练是目前增强深度学习模型鲁棒性最有效的方式[36,88]。然而，对抗训练通常采用单一的攻击方法生成对抗样本，导致模型对未知攻击无法有效防御。对抗训练面临着在不同攻击下泛化能力以及在测试数据上鲁棒性不足的问题。

1.3.3　鲁棒性测评

目前关于深度学习鲁棒性测评的研究较为欠缺，仅有一些开源平台可支持不同模型的鲁棒性测评，包括 CleverHans[89]、Foolbox[90]、ART[91]、Advbox[92]等。这些平台大多只实现了典型的对抗攻防算法，而没有对已有攻防算法进行鲁棒性测评。Carlini 等人[39]提出对防御模型进行鲁棒性测评的若干要求，以提高测评结果的可靠性，但也没有提供具体的测评结果。目前大多数对抗攻防算法在测评时采用的评估指标过于简单，难以反映攻防算法在不同场景下的性能。因此，如何公平、全面地测评深度学习模型的鲁棒性以及攻防算法的有效性仍然是亟待解决的重要研究问题。

1.4　关键研究问题

综上所述，已有研究针对深度学习的鲁棒性问题提出了多种对抗攻击与防御算法，旨在全面分析当前深度学习模型存在的问题并进一步构建安全可靠的深度学习模型。本文围绕基于黑盒对抗攻击的深度学习脆弱性机理分析、基于高效对抗攻击的深度学习鲁棒性增强方法、以及深度学习模型的对抗鲁棒性测评与基准这三方面科学问题进行研究，具体解决现有工作存在的黑盒攻击成功率与效率较低，对抗训练模型鲁棒性不足，以及对抗鲁棒性测评欠缺的问题。本文通过构建对抗攻防测评基准与平台，并面向不同场景研发高效对抗攻击算法解决上述问题。

图1.6总结了本文所解决的关键研究问题。

在黑盒迁移攻击成功率方面，已有攻击方法生成对抗样本的白盒攻击效果与其迁移能力相互制约，造成黑盒迁移攻击成功率较低的问题。一方面，快速梯度符号法[20]只进行一步梯度更新，因此其白盒攻击效果较差；另一方面，基于多步梯度迭代的方法[37,48]生成的对抗样本迁移能力较差。当前攻击方法的黑盒迁移攻击成功率较低，可能导致对深度学习模型的鲁棒性与安全性评估出现偏差。因此，提高黑盒迁移攻击成功率对发现深度学习模型在此场景下的脆弱性及比较不同模型的鲁棒性具有重要研究意义。

在黑盒决策攻击效率方面，已有攻击方法需要对黑盒模型进行大量查询才能生成扰动较小的对抗样本，攻击效率较低。在实际应用中，攻击者对黑盒模型查询次数过多需要付出较高代价（如付费查询），也有可能导致该应用检测到攻击者的恶意行为。因此，降低黑盒决策攻击的查询次数以提升攻击效率是需要解决的问题，此方面的研究同样可用于更好地发现深度学习模型的脆弱性以及比较不同模型的鲁棒性。

在对抗训练模型鲁棒性方面，一些模型在不同攻击下泛化能力较差，同时在测试数据上的鲁棒性存在不足，使得当前对抗训练的效果不及预期。对抗训练需要使用对抗攻击生成的对抗样本，因此攻击方法的不足是导致对抗训练存在问题的主要原因。针对对抗训练设计更加有效的对抗攻击方法用于增强模型的鲁棒性成为一个关键研究问题。相关研究也会进一步加深对深度学习机理的理解。

在对抗鲁棒性测评方面，相关工作有所欠缺，研究者难以对深度学习模型的鲁棒性及对抗攻防算法的有效性进行准确评估。一方面，已有对抗攻防平台只实现了典型算法，而没有对这些算法进行全面的鲁棒性测评。另一方面，常用的鲁棒性评估指标过于简单，难以全面反映模型的性能。因此，构建公平、全面的对抗鲁棒性测评基准可用于更好地比较深度学习模型的鲁棒性，有助于加深对深度学习机理的理解。

1.5　研究内容与主要贡献

基于上述问题，本文针对深度学习的对抗攻击与鲁棒性测评开展系统性研究，包括构建对抗攻防测评基准与平台以及面向不同场景研发高效对抗攻击算法。根据研究问题的不同，本文的研究内容与主要贡献可以总结为以下四个部分。表1.1概括了各章节研究内容的研究问题、解决思路和应用任务。

第一部分研究内容包括第2章和第3章，旨在提高黑盒迁移攻击成功率。受传统优化领域中动量法[93]的启发，第2章提出动量迭代法，该方法在对抗样本生成

表 1.1　各章节研究内容在研究问题、解决思路和应用任务三个方面的总结

章节	研究问题	解决思路	应用任务
第2章	提高黑盒迁移攻击成功率	动量迭代法	图像分类
第3章	提高黑盒迁移攻击成功率	平移不变攻击方法	图像分类
第4章	提升黑盒决策攻击效率	进化攻击方法	人脸识别
第5章	增强对抗训练模型鲁棒性	对抗分布训练	图像分类
第6章	公平、全面的对抗鲁棒性测评	对抗鲁棒性测评基准	图像分类

的迭代优化过程中引入动量项，用于记录梯度更新的历史方向信息。动量迭代法使对抗攻击的优化过程中梯度更新方向更加平稳，同时避免对抗样本落入较差的局部极值点，有效缓解对抗样本的白盒攻击效果与迁移能力的相互制约。针对具有防御机制的深度学习模型，第3章提出*平移不变攻击方法*。该方法构建平移不变攻击目标函数，同时对一组经过平移变换的图片生成对抗样本。平移不变攻击方法可以在不增加计算复杂度的情况下与所有基于梯度的攻击方法相结合，大幅提高针对防御模型的黑盒迁移攻击成功率。该部分的研究为理解深度学习模型的脆弱性机理及发现模型的安全漏洞奠定了理论和方法基础，研究成果发表在人工智能领域顶级国际会议 CVPR 2018[94] 和 CVPR 2019[95] 上。

第二部分研究内容为第4章，旨在提升黑盒决策攻击效率。人脸识别作为人工智能领域应用最广的任务之一，其鲁棒性与安全性的不足可能会对真实世界中的人脸识别相关应用带来严重的安全威胁。在真实世界中，对人脸识别系统进行黑盒决策攻击更为现实。针对这一任务，第4章提出*进化攻击方法*，在黑盒决策攻击的过程中建模搜索方向的局部几何结构，并降低搜索空间的维度。与已有方法相比，进化攻击方法可以通过更少的模型查询次数生成扰动更小的对抗样本，展示出更高的攻击效率。该方法成功攻破了商用人脸识别系统，说明当前使用广泛的人脸识别模型存在安全隐患。该部分的研究为挖掘人脸识别模型和系统的安全漏洞奠定了理论和方法基础，研究成果发表在人工智能领域顶级国际会议 CVPR 2019[96] 上。此方面的研究同时启发了在黑盒得分攻击场景下的高效对抗攻击算法，提出迁移先验引导的梯度估计方法，发表在顶级国际会议 NeurIPS 2019[100] 和顶级国际期刊 TPAMI 2021[101] 上。

第三部分研究内容为第5章，旨在增强对抗训练模型鲁棒性。由于大多数对抗训练方法采用特定的攻击生成对抗样本，其训练得到的模型无法有效防御未知攻击。同时，单一的攻击方式对输入空间中不同扰动的探索不够充分，导致模型在测试数据上鲁棒性不足。为了提升对抗攻击生成对抗样本的多样性，第5章提出对

抗分布训练，其利用对抗分布刻画每一个原始样本邻域内的对抗样本。通过三种对抗攻击方式参数化建模对抗分布，对抗分布训练有效增强了模型的鲁棒性，取得了比一般对抗训练方法更加优异的性能。第5章进一步说明对抗训练可以提升深度学习模型的可解释性。该部分的研究为构建更加鲁棒的深度学习模型奠定了理论和方法基础，研究成果发表在人工智能领域顶级国际会议 NeurIPS 2020[97]和中文核心期刊自动化学报[98]上。

第四部分研究内容为第6章，旨在提供公平、全面的对抗鲁棒性测评。针对图像分类任务，第6章构建对抗鲁棒性测评基准，其中包含 15 种对抗攻击方法与 16 个对抗防御模型。对抗攻击覆盖白盒攻击、黑盒迁移攻击、黑盒得分攻击和黑盒决策攻击中的典型方法。对抗防御包含鲁棒训练、图像变换、随机化、模型集成和可证实防御中的代表模型。鲁棒性基准采用两条鲁棒性曲线作为公正的评估指标，并在多种威胁模型下针对攻防算法进行大规模实验。通过分析实验结果得出一些重要结论：1）防御模型在同样攻击的不同参数下鲁棒性的相对好坏存在差异；2）对抗训练是增强模型鲁棒性最有效的方式，其鲁棒性可以泛化到其他威胁模型下；3）随机化防御在黑盒查询攻击下防御能力较好。该部分的研究为今后对抗攻防模型及算法的开发奠定了测评基础，研究成果发表在人工智能领域顶级国际会议 CVPR 2020[99]上。

除了上述研究内容，本人针对鲁棒高效的深度学习还进行过其他方面研究并取得了一些成果，主要包含：1）利用对抗攻击的思想，提出在黑盒场景下检测深度学习模型是否存在后门的方法，发表在顶级国际会议 ICCV 2021[102]上；2）针对深度学习模型所需存储空间较大，推理速度较慢的问题，提出随机量化方法训练低比特神经网络，发表在顶级国际期刊 IJCV 2019[103]上。由于内容与篇幅限制，本文对这两方面的研究内容不再详细介绍。

1.6　本文组织结构

本文的章节组织结构如图1.7所示，总共包含 7 个章节。

第1章为绪论，介绍本文的研究背景与意义，综述国内外研究现状，总结关键研究问题，并概括本文的研究内容与主要贡献。

第2章提出动量迭代法，在对抗攻击的迭代优化过程中引入动量项，有效缓解对抗样本的白盒攻击效果与迁移能力的相互制约，大幅提高黑盒迁移攻击成功率，发现了现有深度学习模型鲁棒性不足的问题。

第3章提出平移不变对抗攻击方法，对一组经过平移变换的图片生成对抗样本。该方法可以在不增加计算复杂度的情况下与所有基于梯度的攻击方法相结合，有

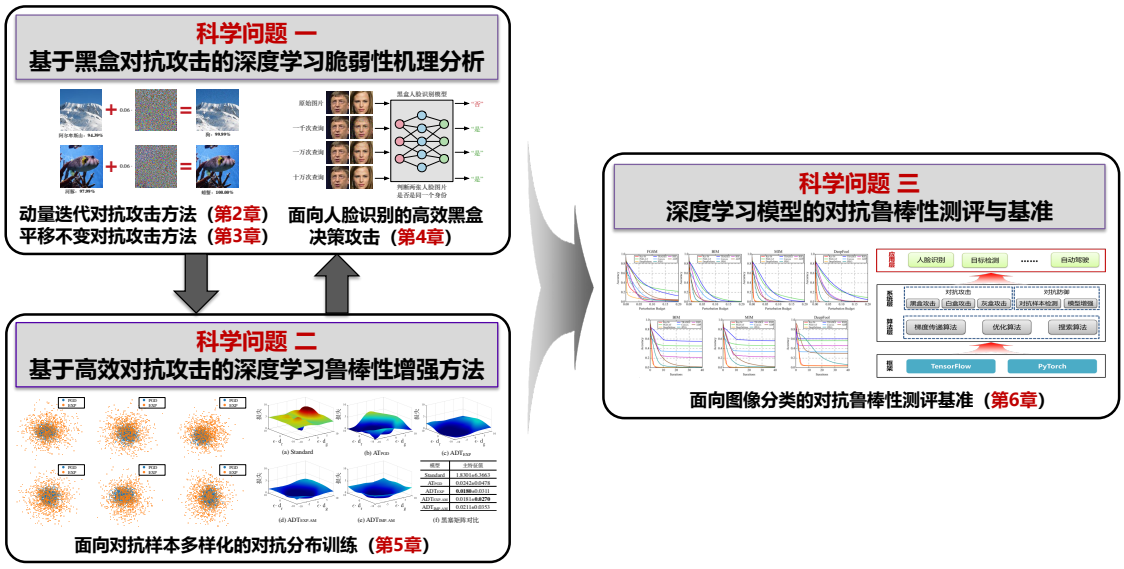


图 1.7　本文的章节组织结构

效提高针对防御模型的黑盒迁移攻击成功率，说明了典型对抗防御模型鲁棒性不足的问题。

第4章针对人脸识别任务提出进化攻击方法，可以在黑盒决策攻击场景下更加高效地生成对抗样本。通过建模搜索方向的局部几何结构并降低搜索空间的维度，该方法需要更少的模型查询次数即可生成对抗样本，并攻破了商用人脸识别系统，验证了人脸识别模型的脆弱性。

第5章提出对抗分布训练框架，通过对抗分布建模原始样本周围多样的对抗样本，有效增强模型在不同攻击下的泛化能力以及在测试数据集上的鲁棒性。该章也进一步分析对抗训练对深度学习模型可解释性的作用。

第6章针对图像分类任务构建公平、全面的对抗鲁棒性测评基准，针对多种典型的对抗攻防算法采用鲁棒性曲线进行测评，对比这些算法在不同威胁模型下的性能，并得出一些重要结论。

第7章回顾并总结本文的研究内容与主要贡献，并对未来研究方向进行展望。

第 2 章　动量迭代对抗攻击方法

对抗样本具有一定的迁移能力，使得对白盒模型生成的对抗样本也可以欺骗未知的黑盒模型。然而，已有对抗攻击方法存在黑盒迁移攻击成功率较低的问题，这会导致对模型鲁棒性的评估不准确。受传统优化领域中动量法的启发，本章提出基于动量的迭代式对抗攻击方法，被称为动量迭代法。该方法在对抗样本生成的迭代优化过程中引入动量项，用于记录对抗样本更新方向的历史信息。动量迭代法可以使对抗攻击的优化过程更加平稳，同时避免对抗样本落入优化问题的局部极值点，有效提高黑盒迁移攻击成功率。本章进一步提出针对多个模型同时进行对抗攻击的策略，增强对黑盒模型的迁移攻击效果。实验结果表明本章所提出的方法可以大幅提高黑盒迁移攻击成功率，并说明现有深度神经网络在此场景下鲁棒性不足的问题。

2.1　本章引言

深度学习模型很容易被攻击者恶意构造的对抗样本欺骗[18-20]。很多对抗攻击方法可以在获取模型结构和参数的基础上生成对抗样本，这些方法大致可被分为三类：基于单步梯度更新的方法[20]、基于多步梯度迭代的方法[37,48]和基于优化的方法[19,49]。具体而言，基于单步梯度更新的方法通过计算一次梯度生成对抗样本，其典型例子为快速梯度符号法（Fast Gradient Sign Method，FGSM）[20]；基于多步梯度迭代的方法通过计算多次梯度迭代地生成对抗样本，其典型例子为基础迭代法（Basic Iterative Method，BIM）[48]；基于优化的方法则直接求解无约束优化问题生成对抗样本，其典型例子为 Carlini 和 Wagner 提出的方法[49]。这些对抗攻击方法在生成对抗样本的过程中均需计算攻击目标函数对于输入数据的梯度，因此被称为白盒攻击。

对抗样本还具有一种被称为*迁移能力*（transferability）的重要特征[50-51]，即对一个模型生成的对抗样本也可以成功欺骗其他模型。这样，攻击者就可以在不获取黑盒模型结构和参数信息的情况下，针对本地构建的替代模型生成对抗样本以欺骗黑盒模型，使黑盒攻击得以实现，如图2.1所示。在实际应用中攻击者很难获取模型的信息，因此黑盒攻击相比于白盒攻击更加现实，也会对深度学习的实际应用带来更加严重的安全威胁[50-51]。对抗样本具有迁移能力的原因是不同深度学习模型由于结构和训练数据的相似性，在训练过程中会学习到相似的决策边界，这使得对其中某一个模型生成的对抗样本也能够成功欺骗其余模型。

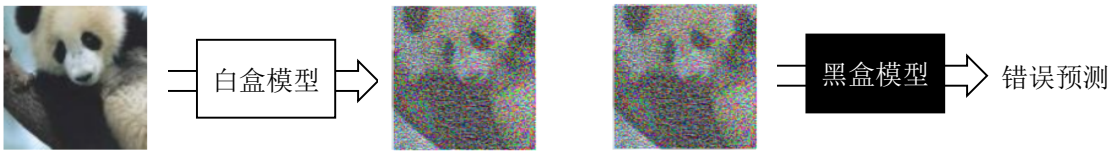


图 2.1　黑盒迁移攻击示意图

尽管如此，已有对抗攻击方法的黑盒迁移攻击成功率较低，尤其是在黑盒模型具有一定防御机制的情况下，不能有效攻破黑盒模型[69,104]。为了成功攻破黑盒模型，对抗样本需要同时满足两个条件：其一是对抗样本可以成功欺骗所采用的白盒模型；其二是对抗样本基于迁移能力可以成功攻破黑盒模型。虽然基于单步梯度更新的方法生成对抗样本的迁移能力很强，但是其对白盒模型的攻击成功率较低。而已有基于多步梯度迭代的方法和基于优化的方法生成对抗样本的白盒攻击效果很好，但是其迁移能力较差。综上所述，已有攻击方法生成对抗样本的白盒攻击效果与其迁移能力相互制约，从而导致这些方法存在黑盒迁移攻击成功率较低的问题。

为解决上述问题，提高对抗样本的黑盒迁移攻击成功率，本章提出基于动量的迭代式对抗攻击方法，被称为动量迭代法（Momentum Iterative Method，MIM）。受传统优化领域中动量法[93]的启发，动量迭代法在对抗样本生成的迭代优化过程中引入动量项，用于记录梯度更新的历史信息。该方法采用历史梯度方向生成对抗样本，可以使梯度方向更加平稳，同时避免对抗样本落入较差的局部极值点。动量迭代法有效缓解了白盒攻击效果与迁移能力的相互制约，其生成的对抗样本在白盒攻击和黑盒迁移攻击场景下均表现出更好的攻击效果。为了进一步提升对抗样本的迁移能力，攻击者通常可以针对多个白盒模型同时进行攻击，这是因为同时欺骗多个白盒模型的对抗样本更有可能欺骗黑盒模型[51]。本章研究几种在对抗攻击中对多个模型进行融合的策略，有效提高黑盒迁移攻击成功率。

本章在广泛使用的 ImageNet[87]数据集上针对典型的深度神经网络测试攻击效果，第2.4节中的实验结果表明动量迭代法可以大幅提高黑盒迁移攻击成功率，其相比于已有方法提升一倍左右。同时，本章也验证了集成对抗训练（ensemble adversarial training）[104]这类防御方式在黑盒迁移攻击下的脆弱性。实验结果说明现有深度神经网络鲁棒性的不足，对于构建更加鲁棒的深度学习模型带来了新的挑战。

2.2　背景知识

本节将首先介绍对抗攻击相关的符号表示与问题定义，然后介绍已有的典型攻击方法。

2.2.1　符号表示与问题定义

本章使用 x 代表输入数据，在图像分类任务中 x 为一张图片。输入数据 x 对应的真实类别用 $y \in \{1, 2, \cdots, L\}$ 表示，其中 L 为总共的类别数量。给定一个基于深度神经网络的分类器模型 $f(x)$，假设其输出为所有 L 个类别上的概率分布，即 $f(x) \in [0, 1]^L$，并且满足 $\sum_{i=1}^{L} f(x)_i = 1$，其中 $f(x)_i$ 为分类器对第 i 个类别的预测概率。通常情况下，分类器的最后一层通过 softmax 函数将网络的 logits 层进行归一化处理，即 $f(x)_i := \frac{e^{l(x)_i}}{\sum_{j=1}^{L} e^{l(x)_j}}$，其中 $l(x)$ 代表分类器的 logits 层输出。分类器的最终分类结果为预测概率最大的类别，可以表示为 $C_f(x) = \arg\max_{i \in \{1,2,\cdots,L\}} f(x)_i$。

对抗攻击的目标是在原始样本 x 的邻域内寻找可以被分类器 f 分类错误的对抗样本 x^*。为了叙述简洁，本章主要针对无目标攻击进行介绍，所述方法可以被简单地扩展到有目标攻击场景。对于无目标攻击，对抗样本可以通过求解约束优化问题（constrained optimization problem）生成，如下所示：

$$\arg\max_{x^*} J(f(x^*), y), \qquad \text{s.t. } \|x^* - x\|_p \leqslant \epsilon, \tag{2.1}$$

其中 $J(\cdot, \cdot)$ 代表模型的损失函数，同时也是攻击目标函数，通常选取为交叉熵损失（cross-entropy loss），其具体计算公式为：

$$J(f(x), y) := -\log f(x)_y. \tag{2.2}$$

已有攻击方法可以针对约束优化问题（2.1）进行近似求解，下面将介绍一些典型的方法。

2.2.2　已有对抗攻击方法

白盒对抗攻击主要包括基于单步梯度更新的方法[20]、基于多步梯度迭代的方法[37,48]和基于优化的方法[19,49]。下面将介绍这三类对抗攻击的典型方法。

快速梯度符号法（FGSM）[20]：作为基于单步梯度更新的典型方法，快速梯度符号法通过计算损失函数 J 对于输入数据 x 的梯度生成对抗样本。在 ℓ_∞ 范数的限制下，该方法生成对抗样本的过程可以被描述为：

$$x^* = x + \epsilon \cdot \text{sign}(\nabla_x J(f(x), y)), \tag{2.3}$$

其中 $\nabla_x J(f(x), y)$ 代表损失函数对于输入数据的梯度。快速梯度符号法通过符号函

数 sign(·) 对梯度进行归一化，使得添加的扰动满足 ℓ_∞ 范数的限制。在损失关于输入是线性函数的假设下，快速梯度符号法是求解优化问题（2.1）的最优方式。在 ℓ_2 范数的限制下，快速梯度符号法可以被扩展为*快速梯度法*（Fast Gradient Method，FGM），其具体过程可以被描述为：

$$x^* = x + \epsilon \cdot \frac{\nabla_x \boldsymbol{J}(f(x), y)}{\|\nabla_x \boldsymbol{J}(f(x), y)\|_2}. \tag{2.4}$$

在实际中，深度神经网络是高度非线性的，因此快速梯度符号法依赖的线性假设难以成立，这就导致该方法生成的对抗样本对白盒模型的攻击成功率较低。

基础迭代法（BIM）[48]：基础迭代法属于基于多步梯度迭代的方法，利用损失函数的梯度多次更新对抗样本，达到更强的攻击效果。在 ℓ_∞ 范数的限制下，基础迭代法生成对抗样本的过程可以被描述为：

$$x_0^* = x, \quad x_{t+1}^* = x_t^* + \alpha \cdot \text{sign}(\nabla_x \boldsymbol{J}(f(x_t^*), y)), \tag{2.5}$$

其中 x_t^* 代表第 t 轮迭代后生成的对抗样本，α 代表迭代过程中的步长。假设总共迭代 $\boldsymbol{T}$ 轮，为了使对抗样本满足 ℓ_∞ 范数的限制，通常可以设置步长为 $\alpha = \frac{\epsilon}{T}$ 或者在每一轮迭代后通过裁剪（clip）函数将对抗样本映射到限制的范围内。可以看到，当 $\boldsymbol{T} = 1$ 时，基础迭代法退化为快速梯度符号法。基础迭代法在 ℓ_2 范数的限制下生成对抗样本的过程类似，不再赘述。此外，*投影梯度下降法*（Projected Gradient Descent，PGD）[37] 也是基于多步梯度迭代的典型方法，并且和基础迭代法十分相似。这两个方法唯一的区别是投影梯度下降法将对抗样本进行随机初始化，即 $x_0^* = x + n$，其中 n 代表在 ℓ_∞ 扰动范围内均匀采样的随机噪声。以基础迭代法和投影梯度下降法为代表的基于多步梯度迭代的攻击方法在每一轮迭代过程中贪婪地用梯度方向更新对抗样本，这会导致对抗样本轻易地落入约束优化问题（2.1）中较差的局部极值点。这就意味着这些方法生成的对抗样本 “过拟合” 了当前采用的白盒模型，导致其迁移能力较差[69,104]。

C&W 方法[49]：该方法是加州大学伯克利分校的 Carlini 和 Wagner 提出的基于优化的攻击方法，其将约束优化问题（2.1）转化为无约束优化问题并进行求解，如下所示：

$$\arg\min_{x^*} \lambda \cdot \|x^* - x\|_p - \boldsymbol{J}(f(x^*), y), \tag{2.6}$$

其中 λ 为平衡两项损失的权重，其最优值通过二分查找进行搜索。与之前方法不同，C&W 方法采用的损失函数 $\boldsymbol{J}$ 定义为：

$$\boldsymbol{J}(f(x), y) = \max_{i \neq y} f(x)_i - f(x)_y. \tag{2.7}$$

当 $\boldsymbol{J}(f(x), y) > 0$ 时，输入数据 x 被模型分类为真实类别 y 以外的其他类别，即分

类错误。C&W 方法通过基于梯度的优化方式寻找优化问题（2.6）的最优解。与基于多步梯度迭代的攻击方法类似，C&W 方法同样也存在所生成对抗样本迁移能力较差的问题。

2.3　算法设计

通过以上分析可以发现，基于单步梯度更新的方法对白盒模型的攻击效果较差，而基于多步梯度迭代的方法和基于优化的方法虽然白盒攻击效果很好，但是迁移能力较差。因此，已有方法生成对抗样本的白盒攻击效果与其迁移能力相互制约，这就导致了黑盒迁移攻击成功率较低的问题。为了解决上述问题，本节详细介绍动量迭代法（MIM）并探讨在对抗攻击中对多个白盒模型进行融合的有效策略，以提高黑盒迁移攻击成功率。

2.3.1　动量迭代法

动量法（momentum method）[93]作为传统优化领域中一种典型的优化方法，具有以下三大优势：第一，在迭代过程中会对损失函数的梯度方向进行累积，因此可加快梯度下降的收敛速度[93]；第二，对历史梯度信息具有记忆能力，有助于避免落入优化问题中较差的局部极值点，帮助优化过程收敛到更优解[105]；第三，在随机梯度下降中可以使梯度的更新方向更加平稳[106]。受动量法的启发，动量迭代法在对抗样本生成的迭代优化过程中引入动量项，使优化过程中的更新方向更加平稳，避免对抗样本落入优化问题的局部极值点。该方法可以减轻所生成的对抗样本对白盒模型的依赖程度，防止对抗样本 “过拟合” 白盒模型，从而达到解决对抗样本的白盒攻击效果与其迁移能力相互制约的问题，提高黑盒迁移攻击成功率。

具体而言，在 ℓ_∞ 范数的限制下，动量迭代法生成对抗样本的过程如算法2.1所示。其中 g_t 代表第 t 轮迭代累积的梯度方向。在每一轮迭代中，算法首先计算模型损失函数对于输入的梯度 $\nabla_x \boldsymbol{J}(f(x_t^*), y)$，然后通过公式（2.8）更新累积的梯度方向，其中 μ 代表动量衰减系数。算法在每一轮迭代中利用 g_{t+1} 的符号方向更新对抗样本，如公式（2.9）所示。从算法中可以看到，在 $\mu = 0$ 的情况下，动量迭代法退化为基础迭代法，所以基础迭代法也可以被认为是动量迭代法的一个特例。值得注意的是，在每一轮迭代中，损失函数的梯度 $\nabla_x \boldsymbol{J}(f(x_t^*), y)$ 会除以其 ℓ_1 范数（或其他范数）进行归一化，这是因为不同轮迭代中梯度的幅度会出现巨大变化，所以采用归一化的方式可以使梯度更新更加平稳。动量迭代法可以被扩展到 ℓ_2 范数限制下攻击以及有目标攻击等场景，第2.3.3节会做进一步介绍。

算法 2.1 动量迭代法

输入：分类器 f，损失函数 $\boldsymbol{J}$，原始样本 x，真实类别 y，扰动规模 ϵ，迭代轮数 $\boldsymbol{T}$，动量衰减系数 μ；

输出：满足 $\|x^* - x\|_\infty \leqslant \epsilon$ 的对抗样本 x^*；

令 $\alpha = \frac{\epsilon}{T}$；

令 $g_0 = 0$，$x_0^* = x$；

对 $t = 0, \cdots, T-1$ **执行**

　　将 x_t^* 输入分类器 f 并计算损失函数的梯度 $\nabla_x \boldsymbol{J}(f(x_t^*), y)$；

　　通过梯度信息更新动量项：

$$g_{t+1} = \mu \cdot g_t + \frac{\nabla_x \boldsymbol{J}(f(x_t^*), y)}{\|\nabla_x \boldsymbol{J}(f(x_t^*), y)\|_1}; \tag{2.8}$$

　　更新对抗样本：

$$x_{t+1}^* = x_t^* + \alpha \cdot \mathrm{sign}(g_{t+1}); \tag{2.9}$$

返回：$x^* \leftarrow x_T^*$。

2.3.2　对抗攻击中多模型融合策略

本小节进一步研究如何针对多个白盒模型进行有效攻击。在实际场景中，攻击者往往可以针对多个白盒模型同时进行攻击以提高黑盒迁移攻击成功率[51]。此方法背后的原因是：对抗样本为了同时欺骗多个白盒模型，会寻找并利用对这些模型普遍有效的扰动，则这一扰动很有可能也可以欺骗未知的黑盒模型，达到更好的黑盒迁移攻击效果。

同时攻击多个白盒模型可以采用不同的模型融合策略。本小节提出一种针对多个模型 logits 层进行融合的策略。具体而言，假设有 $\boldsymbol{K}$ 个白盒模型，该策略将不同模型的 logits 层进行融合，如下所示：

$$l(x) = \sum_{k=1}^{K} w_k l_k(x), \tag{2.10}$$

其中 $l_k(x)$ 代表第 k 个模型的 logits 层输出，w_k 代表第 k 个模型所占的权重，满足 $w_k \geqslant 0$ 且 $\sum_{k=1}^{K} w_k = 1$。在此融合策略的基础上，攻击者可以将融合之后的模型 $l(x)$ 看作一个集成的白盒模型，并利用对抗攻击算法进行攻击。算法2.2给出了将此策略与动量迭代法相结合后的攻击过程。

除对多个模型 logits 层进行融合的策略外，本小节还介绍了另外两种模型融合策略作为对比。其一是对 $\boldsymbol{K}$ 个模型的预测概率分布进行融合，如下所示：

$$f(x) = \sum_{k=1}^{K} w_k f_k(x), \tag{2.11}$$

其中 $f_k(x)$ 代表第 k 个模型对输入数据 x 的预测概率分布。对多个模型预测概率分布进行融合的策略也被文献[51]所采用。其二是对 $\boldsymbol{K}$ 个模型的损失函数进行融合，

算法 2.2 针对多个模型 logits 层融合的动量迭代法

输入： K 个模型的 logits 层输出 $l_1, l_2, \cdots, l_K$，权重系数 $w_1, w_2, \cdots, w_K$，原始样本 x，真实类别 y，扰动规模 ϵ，迭代轮数 T，动量衰减系数 μ；
输出： 满足 $\|x^* - x\|_\infty \leqslant \epsilon$ 的对抗样本 x^*；

令 $\alpha = \frac{\epsilon}{T}$；
令 $g_0 = 0$，$x_0^* = x$；
对 $t = 0, \cdots, T-1$ **执行**
　对 $k = 1, 2, \cdots, K$ 计算第 k 个模型对于输入 x_t^* 的 logits 输出 $l_k(x_t^*)$；
　将 K 个模型的 logits 层融合：$l(x_t^*) = \sum_{k=1}^{K} w_k l_k(x_t^*)$；
　利用 softmax 函数将 $l(x_t^*)$ 转换为预测概率分布 $f(x_t^*)$ 并通过公式（2.2）计算损失 $J(f(x_t^*), y)$；
　计算损失函数的梯度 $\nabla_x J(f(x_t^*), y)$；
　通过公式（2.8）更新动量项 g_{t+1}；
　通过公式（2.9）更新对抗样本 x_{t+1}^*；
返回： $x^* \leftarrow x_T^*$。

如下所示：

$$J(f(x), y) = \sum_{k=1}^{K} w_k J(f_k(x), y). \tag{2.12}$$

由于上述三种策略对多个模型进行融合时选取的位置不同，会导致攻击效果出现较大的差异。第2.4.2节中的实验结果表明对 logits 层进行融合的策略比对预测概率分布融合和对损失函数融合的策略性能更好，黑盒迁移攻击成功率更高。

2.3.3　动量迭代法的扩展

本章所提出的动量迭代法在 ℓ_2 范数限制下攻击、有目标攻击等其他场景下也可以使用，并且可以提高各个场景下的黑盒迁移攻击成功率。本小节将描述动量迭代法在其他场景下的扩展。

首先考虑 ℓ_2 范数限制下的对抗攻击。在此情况下，为了使得生成的对抗样本满足 $\|x^* - x\|_2 \leqslant \epsilon$，动量迭代法更新对抗样本的方式为：

$$x_{t+1}^* = x_t^* + \alpha \cdot \frac{g_{t+1}}{\|g_{t+1}\|_2}, \tag{2.13}$$

其中 g_{t+1} 仍然通过公式（2.8）计算。

对于有目标攻击，攻击者希望生成的对抗样本 x^* 被分类为指定的目标类别 y^*。相比于无目标攻击的最大化优化问题（2.1），有目标攻击的优化问题变为在 ℓ_p 范数的限制下最小化 $J(f(x^*), y^*)$，即：

$$\arg\min_{x^*} J(f(x^*), y^*), \qquad \text{s.t. } \|x^* - x\|_p \leqslant \epsilon. \tag{2.14}$$

此时，动量迭代法中动量项 g_t 通过如下的的方式进行更新：

$$g_{t+1} = \mu \cdot g_t + \frac{\nabla_x J(f(x_t^*), y^*)}{\|\nabla_x J(f(x_t^*), y^*)\|_1}; \tag{2.15}$$

在 ℓ_∞ 和 ℓ_2 范数的限制下，有目标攻击动量迭代法生成对抗样本的更新方式为：

$$x_{t+1}^* = x_t^* - \alpha \cdot \text{sign}(g_{t+1}); \quad x_{t+1}^* = x_t^* - \alpha \cdot \frac{g_{t+1}}{\|g_{t+1}\|_2}. \tag{2.16}$$

因此，本章所提出的动量迭代法可以被扩展为不同攻击场景下的一系列攻击方法。实验结果也展示出该方法在不同场景下的有效性。

2.4　实验结果

本节在 ImageNet[87] 图像分类数据集上进行实验，并展示动量迭代法相比于已有攻击方法的有效性①。

对于被攻击的目标模型，本节选取在 ImageNet 数据集上预训练的七个模型进行实验。其中四个模型是正常训练得到的模型，包括 Inception v3（记为 Inc-v3）[5]、Inception v4（记为 Inc-v4）[107]、Inception ResNet v2（记为 IncRes-v2）[107] 和 ResNet v2 152（记为 Res-152）[108]。另外三个模型是通过集成对抗训练（ensemble adversarial training）[104] 得到的模型，分别记为 Inc-v3$_{ens3}$、Inc-v3$_{ens4}$ 和 IncRes-v2$_{ens}$。这些模型是 ImageNet 图像分类任务中非常典型且性能较好的模型，并由 TensorFlow[109] 算法库提供，因此本节的实验具有一定代表性。

对于测试数据的选取，本节在 ImageNet 验证集中针对共计 1000 个预测类别各随机选择 1 张图片，构成了包含 1000 张图片的数据集，并且这些图片可以被所研究的七个图像分类模型分类正确。这样设置的原因是如果原始图片不能被模型分类正确，则无目标攻击可以直接利用原始图片达到攻击目标，其效果好坏也就失去了意义。

本节实验将动量迭代法（MIM）与已有攻击方法进行对比。选取的基准方法包括基于单步梯度更新的快速梯度符号法（FGSM）[20] 和基于多步梯度迭代的基础迭代法（BIM）[48]。由于基于优化的方法无法明确控制对抗样本与原始样本之间的距离，本节没有将动量迭代法与此类方法进行比较。

本节首先针对 ℓ_∞ 范数限制下的无目标攻击进行实验，在第2.4.1节和第2.4.2节中分别展示了针对单模型进行攻击和针对多模型进行攻击的结果。第2.4.3节进一步展示了在其他场景下的攻击结果。

① 源代码参见 https://github.com/dongyp13/Non-Targeted-Adversarial-Attacks。

表 2.1　在 ℓ_∞ 范数的限制下不同攻击方法针对单模型的无目标攻击成功率（%）对比

	攻击方法	Inc-v3	Inc-v4	IncRes-v3	Res-152	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
Inc-v3	FGSM	72.3*	28.2	26.2	25.3	11.3	10.9	4.8
	BIM	**100.0***	22.8	19.9	16.2	7.5	6.4	4.1
	MIM	**100.0***	**48.8**	**48.0**	**35.6**	**15.1**	**15.2**	**7.8**
Inc-v4	FGSM	32.7	62.0*	26.6	27.2	13.7	11.9	6.2
	BIM	35.8	**99.9***	24.7	19.3	7.8	6.8	4.9
	MIM	**65.6**	**99.9***	**54.9**	**46.3**	**19.8**	**17.4**	**9.6**
IncRes-v2	FGSM	32.6	28.1	55.3*	25.8	13.1	12.1	7.5
	BIM	37.8	20.8	**99.6***	22.8	8.9	7.8	5.8
	MIM	**69.8**	**62.1**	99.5*	**50.6**	**26.1**	**20.9**	**15.7**
Res-152	FGSM	35.0	28.2	27.5	72.9*	14.6	13.2	7.5
	BIM	26.7	22.7	21.2	**98.6***	9.3	8.9	6.2
	MIM	**53.6**	**48.9**	**44.7**	98.5*	**22.1**	**21.7**	**12.9**

2.4.1　针对单模型的攻击结果

表2.1展示了不同攻击方法针对七个目标模型的攻击成功率。其中，对抗样本通过 FGSM、BIM 和 MIM 攻击方法分别针对 Inc-v3、Inc-v4、IncRes-v2 和 Res-152 生成。攻击成功率为对应模型以对抗样本作为输入数据的情况下计算的分类错误率。在所有实验中，最大扰动设置为 $\epsilon = 16$，图片像素值范围为 $[0, 255]$。BIM 和 MIM 的迭代轮数设置为 $T = 10$。MIM 的动量衰减系数设置为 $\mu = 1.0$，其不同取值将在第2.4.1.2节中进行研究。表2.1中有 * 注明的结果代表白盒攻击成功率，其余为黑盒迁移攻击成功率。

从表2.1展示的结果中可以得出以下结论。首先，MIM 的白盒攻击效果与 BIM 类似，都可以达到接近 100% 的攻击成功率。其次，与 FGSM 相比，BIM 降低了黑盒迁移攻击成功率。但是，通过加入动量项，MIM 的黑盒迁移攻击成功率明显高于 FGSM 和 BIM。在大多数情况下，MIM 的黑盒迁移攻击成功率是 BIM 的两倍以上，验证了所提出方法的有效性。

值得注意的是，尽管 MIM 大幅提高了黑盒迁移攻击成功率，但是对集成对抗训练得到的模型攻击成功率仍然很低。例如，在不同情况下针对 IncRes-v2$_{ens}$ 进行黑盒迁移攻击的成功率均低于 16%。针对这一问题，第2.4.2节展示了针对多个模型进行融合的对抗攻击可以进一步提高对集成对抗训练模型的黑盒迁移攻击成功率。下面将通过对比实验分析 MIM 与 BIM 之间的差异，以进一步解释 MIM 在黑

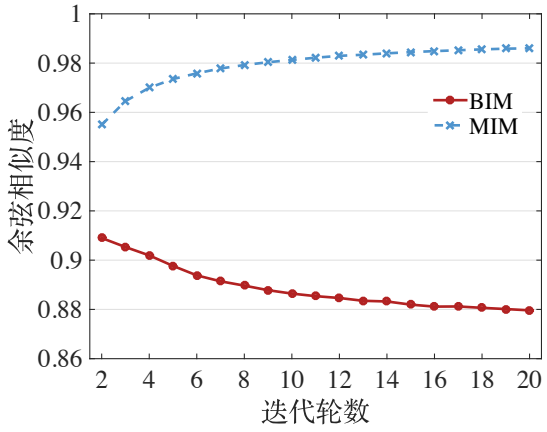


图 2.2　梯度更新方向变化趋势图

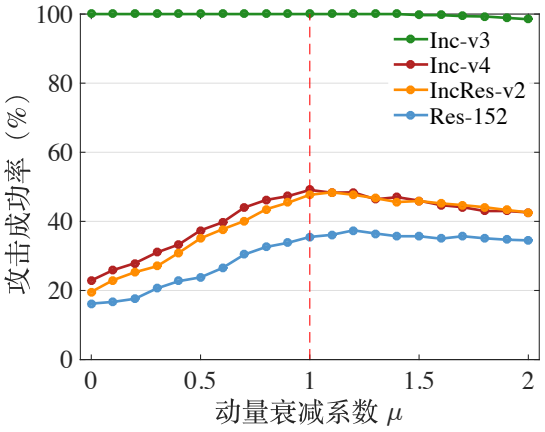


图 2.3　攻击成功率随动量衰减系数变化图

盒迁移攻击中性能更加优越的原因。

2.4.1.1　梯度更新方向

为了解释 MIM 相比于 BIM 黑盒迁移攻击能力更好的原因，本小节针对这两种攻击方法在迭代过程中的梯度更新方向进行分析。具体而言，实验计算不同方法在相邻两轮迭代中梯度更新方向的余弦相似度（cosine similarity）。图2.2展示了针对 Inc-v3 模型使用 BIM 和 MIM 进行攻击时梯度更新方向的余弦相似度。可以看到，MIM 的梯度更新方向呈现出更高的余弦相似度。因此，MIM 在生成对抗样本的多步梯度迭代过程中更新方向相比于 BIM 更加平稳，这也印证了动量法在优化过程中的优势。

对抗样本的迁移能力来源于不同深度学习模型在输入空间中学习到了相似的决策边界[51]。尽管不同模型的决策边界具有一定相似性，但是由于深度神经网络的非线性结构，不同模型的决策边界很难完全相同。一些模型在输入空间中可能存在一些异常的决策区域（如文献[51]中图 4 和图 5 所示的孔洞）。类似的异常区域在其他模型中可能不存在或存在于不同的位置。这些区域对应于优化问题中较差的局部极值点，因此 BIM 很容易陷入这些区域，导致生成的对抗样本迁移能力较差。另一方面，通过图2.2可以观察到 MIM 获得了更加平稳的梯度更新方向，这有助于 MIM 摆脱这些异常区域，从而使得生成的对抗样本具有更好的迁移能力。此外，平稳的更新方向会增大扰动的 ℓ_2 范数，可能使对抗样本具有更好的迁移能力，这也是 MIM 性能优于 BIM 的一个解释。

2.4.1.2　动量衰减系数

本小节分析动量衰减系数 μ 的适当取值。实验使用 MIM 针对 Inc-v3 模型进行攻击，并设置最大扰动规模为 $\epsilon = 16$，迭代轮数为 $T = 10$。动量衰减系数设置的变

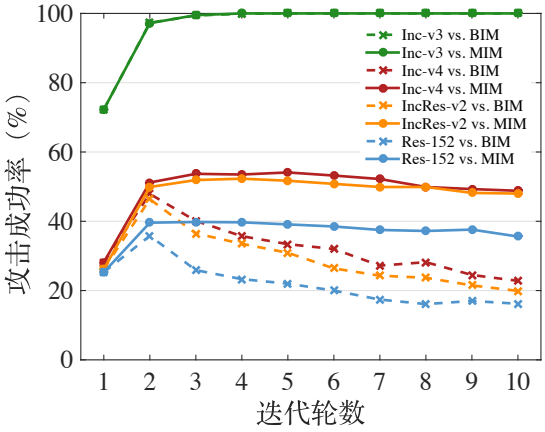


图 2.4　攻击成功率随迭代轮数变化图

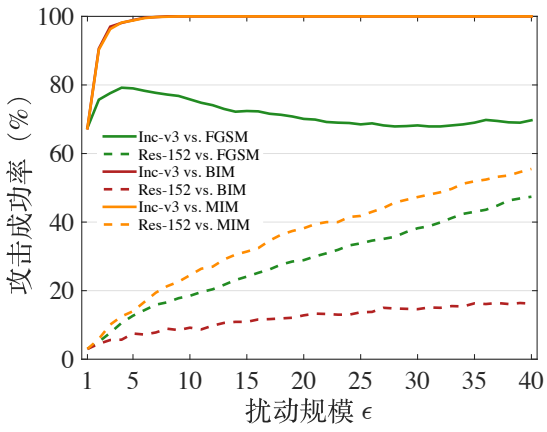


图 2.5　攻击成功率随扰动规模变化图

化范围为 [0.0, 2.0]，其变化粒度为 0.1。图2.3展示了在不同动量衰减系数 μ 的取值下生成的对抗样本对 Inc-v3、Inc-v4、IncRes-v2 和 Res-152 的攻击成功率。对于黑盒模型，攻击成功率通常在 $\mu = 1.0$ 左右取得最大值。当 $\mu = 1.0$ 时，公式（2.8）中 g_t 的另一种理解是其简单地将之前迭代过程中所有的梯度相加以更新对抗样本。

2.4.1.3　迭代轮数

迭代轮数在对抗攻击中也对攻击结果有着重要的影响，本小节对此进行进一步分析。实验使用 BIM 和 MIM 攻击方法针对 Inc-v3 模型进行攻击，其中的参数设置为 $\epsilon = 16$、$\mu = 1.0$。迭代轮数设置为 1 至 10。图2.4展示了在不同迭代轮数的情况下生成的对抗样本对 Inc-v3、Inc-v4、IncRes-v2 和 Res-152 的攻击成功率。

从图2.4中可以看到，BIM 对黑盒模型的迁移攻击成功率随着迭代轮数增多逐渐降低，而 MIM 的黑盒迁移攻击成功率随着迭代轮数增多不会出现明显下降。这一结果验证了 BIM 生成的对抗样本很容易“过拟合”白盒模型，展现出较差的迁移能力。但是 MIM 有助于减轻白盒攻击效果与迁移能力的相互制约，从而呈现出对白盒和黑盒模型更强的攻击能力。

2.4.1.4　扰动规模

最后，本小节研究对抗扰动规模 ϵ 对攻击成功率的影响。实验使用 FGSM、BIM 和 MIM 针对 Inc-v3 模型进行攻击，其中扰动规模 ϵ 设置为从 1 增长至 40。BIM 和 MIM 的步长 α 设置为 1，因此迭代轮数随扰动规模 ϵ 线性变化。MIM 的动量衰减系数设置为 $\mu = 1.0$。图2.5展示了在不同扰动规模的情况下生成的对抗样本对白盒 Inc-v3 模型和黑盒 Res-152 模型的攻击成功率。

对于白盒攻击，BIM 和 MIM 两种基于多步梯度迭代的方法很快就能够达到 100% 的攻击成功率，但是当扰动规模较大时，基于单步梯度更新的 FGSM 白盒攻

表 2.2　在 ℓ_∞ 范数的限制下不同攻击方法针对多模型的无目标攻击成功率（%）对比

	模型融合策略	FGSM		BIM		MIM	
		白盒模型	黑盒模型	白盒模型	黑盒模型	白盒模型	黑盒模型
-Inc-v3	logits 层	**55.7**	**45.7**	**99.7**	**72.1**	**99.6**	**87.9**
	概率分布层	52.3	42.7	95.1	62.7	97.1	83.3
	损失函数层	50.5	42.2	93.8	63.1	97.0	81.9
-Inc-v4	logits 层	**56.1**	**39.9**	**99.8**	**61.0**	**99.5**	**81.2**
	概率分布层	50.9	36.5	95.5	52.4	97.1	77.4
	损失函数层	49.3	36.2	93.9	50.2	96.1	72.5
-IncRes-v2	logits 层	**57.3**	**38.8**	**99.5**	**54.4**	**99.5**	**76.5**
	概率分布层	52.1	35.8	97.1	46.9	98.0	73.9
	损失函数层	50.7	35.2	96.2	45.9	97.4	70.8
-Res-152	logits 层	**53.5**	**35.9**	99.6	**43.5**	99.6	**69.6**
	概率分布层	51.9	34.6	**99.9**	41.0	**99.8**	67.0
	损失函数层	50.4	34.1	98.2	40.1	98.8	65.2

击成功率会降低。产生这一现象的主要原因是当扰动规模较大时，FGSM 中对模型损失函数为线性的假设不再成立[51]。对于黑盒迁移攻击，虽然这三种方法的攻击成功率随着扰动规模逐渐增长，但是 MIM 的黑盒迁移攻击成功率增长更快。换句话说，为了对黑盒模型达到同样的迁移攻击成功率，MIM 可以使用更小的扰动规模，这样生成的对抗样本更加难以分辨。

2.4.2　针对多模型的攻击结果

本小节首先对比第2.3.2节中介绍的在对抗攻击中对多个模型进行融合的不同策略。实验选取 Inc-v3、Inc-v4、IncRes-v2 和 Res-152 作为目标模型。每一组实验选取四个模型中的一个模型作为黑盒模型，并将其余三个模型作为白盒模型进行攻击。例如，“-Inc-v3” 代表选取 Inc-v4、IncRes-v2 和 Res-152 三个模型作为白盒模型，并将 Inc-v3 作为黑盒模型。针对三个白盒模型进行融合的策略包括对 logits 层进行融合、对预测概率分布进行融合和对损失函数进行融合。三个模型的融合权重均设置为 $\frac{1}{3}$。实验同样使用 FGSM、BIM 和 MIM 三种攻击方法进行攻击。扰动规模设置为 $\epsilon = 16$，BIM 和 MIM 的迭代轮数设置为 $T = 10$，MIM 中的动量衰减系数设置为 $\mu = 1.0$。

表2.2展示了在不同的模型融合策略下各个方法的攻击成功率。在每一组实验中，表中分别报告了对白盒模型和黑盒模型的攻击成功率。可以看到，在白盒和黑

表 2.3　在 ℓ_∞ 范数的限制下不同攻击方法针对集成对抗训练模型的无目标攻击成功率（%）对比

	-Inc-v3$_{ens3}$		-Inc-v3$_{ens4}$		-IncRes-v2$_{ens}$	
	白盒模型	黑盒模型	白盒模型	黑盒模型	白盒模型	黑盒模型
FGSM	36.1	15.4	33.0	15.0	36.2	6.4
BIM	**99.6**	18.6	99.2	18.7	99.5	9.9
MIM	**99.6**	**37.6**	**99.3**	**40.3**	**99.7**	**23.3**

表 2.4　在 ℓ_2 范数的限制下不同攻击方法的无目标攻击成功率（%）对比

攻击方法	Inc-v3	Inc-v4	IncRes-v3	Res-152	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
FGM	52.7	49.3	50.4	46.6	35.4	37.5	32.4
BIM	65.3	56.7	54.3	44.7	29.5	36.4	19.9
MIM	**89.7**	**88.0**	**86.1**	**81.4**	**59.8**	**66.5**	**56.4**

盒迁移攻击场景下，对 logits 层进行融合的策略相比于对预测概率分布进行融合和对损失函数进行融合的策略具有更好的攻击效果，取得了更高的攻击成功率。这样的实验现象在各个攻击方法中普遍存在。从表2.2中还可以观察到 MIM 生成的对抗样本具有非常强的迁移能力，展示出强大的黑盒迁移攻击效果。例如，通过对 Inc-v4、IncRes-v2 和 Res-152 三个模型进行融合并使用 MIM 进行攻击的情况下，所生成的对抗样本可以对黑盒的 Inc-v3 模型达到 87.9% 的迁移攻击成功率，验证了正常训练的深度神经网络在黑盒迁移攻击下的脆弱性。

本小节进一步针对具有防御机制的集成对抗训练模型进行对抗攻击实验。实验选取所有七个图像分类模型，为了对每一个集成对抗训练模型进行黑盒迁移攻击，选取其余六个模型作为白盒模型。对白盒模型进行融合的策略选取为对 logits 层进行融合，其中不同白盒模型所占权重均为 $\frac{1}{6}$。扰动规模设置为 $\epsilon = 16$，BIM 和 MIM 的迭代轮数设置为 $T = 20$，MIM 中的动量衰减系数设置为 $\mu = 1.0$。表2.3展示了对白盒模型和黑盒模型的攻击成功率。可以看到，在更强的攻击方式下，集成对抗训练这种防御机制对黑盒迁移攻击不能有效防御。例如，使用 MIM 对多个白盒模型进行攻击，所生成的对抗样本可以对黑盒的 Inc-v3$_{ens4}$ 模型达到 40% 以上的迁移攻击成功率，验证了集成对抗训练这种防御机制在黑盒迁移攻击下的脆弱性。

2.4.3　其他场景下的攻击结果

本小节展示动量迭代法在其他攻击场景下的结果。基于之前的实验分析，对多个模型进行攻击的效果显著优于对单个模型进行攻击的效果，所以本小节的实

表 2.5　在 ℓ_∞ 范数的限制下不同攻击方法的有目标攻击成功率（%）对比

攻击方法	Inc-v3	Inc-v4	IncRes-v3	Res-152	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
FGSM	0.5	0.4	0.2	0.5	0.1	0.1	0.1
BIM	9.0	7.0	7.3	3.3	0.4	0.1	0.1
MIM	**17.6**	**15.6**	**16.1**	**11.4**	**0.5**	**0.9**	**0.2**

表 2.6　在 ℓ_2 范数的限制下不同攻击方法的有目标攻击成功率（%）对比

攻击方法	Inc-v3	Inc-v4	IncRes-v3	Res-152	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
FGM	0.4	0.5	0.7	0.4	0.2	0.2	0.4
BIM	17.8	15.2	16.4	9.2	0.7	1.7	0.5
MIM	**21.0**	**21.8**	**21.7**	**17.4**	**1.6**	**2.0**	**1.9**

验仅包含在不同场景下对多个模型进行攻击的结果。

实验首先考虑在 ℓ_2 范数限制下的无目标攻击。由于 ℓ_2 距离与样本的维度相关，对抗扰动的规模设置为 $\epsilon = 16\sqrt{D}$，其中 D 代表对抗扰动的维度，D 同时也是模型输入图片的维度。实验选取 Inc-v3、Inc-v4、IncRes-v2、Res-152、Inc-v3$_{ens3}$、Inc-v3$_{ens4}$ 和 IncRes-v2$_{ens}$ 七个模型作为目标模型。每一组实验选取其中一个模型作为黑盒模型，并将其余六个模型作为白盒模型进行攻击。实验采用对六个模型的 logits 层进行融合的策略，其中不同模型的融合权重均为 $\frac{1}{6}$。表2.4展示了使用 FGM、BIM 和 MIM 三种攻击方法对不同模型的黑盒迁移攻击成功率。与在 ℓ_∞ 范数限制下的攻击结果类似，MIM 可以显著提高对不同模型的黑盒迁移攻击成功率，验证了所提出方法的有效性。同时，MIM 针对集成对抗训练模型可以达到 60% 左右的黑盒迁移攻击成功率，也说明了集成对抗训练在 ℓ_2 范数扰动下鲁棒性的不足。

有目标攻击需要模型将对抗样本错分为指定的目标类别，因此其相比于无目标攻击更加困难。尤其是对于总共包含 1000 个预测类别的 ImageNet 图像分类任务，有目标攻击很难在黑盒场景下成功欺骗目标模型。本小节进一步验证动量迭代法在有目标攻击场景下的效果。与之前的实验设置类似，有目标攻击实验仍然选取所有七个模型作为目标模型，并采用对 logits 层进行融合的策略对其中六个模型进行攻击，进而评估所生成的对抗样本对黑盒模型的攻击成功率。在有目标攻击场景下，攻击成功率为目标模型将对抗样本错分为指定目标类别的比例。每一张图片的攻击目标类别通过随机采样的方式选取。在 ℓ_∞ 范数的限制下，扰动规模设置为 $\epsilon = 48$。在 ℓ_2 范数的限制下，扰动规模设置为 $\epsilon = 48\sqrt{D}$，其中 D 仍然代表对抗扰动的维度。BIM 和 MIM 的迭代轮数设置为 $T = 20$。MIM 中的动量衰减系数设置为 $\mu = 1.0$。

表2.5展示了 FGSM、BIM 和 MIM 三种攻击方法在 ℓ_∞ 范数的限制下对不同模型进行有目标攻击的黑盒迁移攻击成功率。与之类似地，表2.6展示了 FGM、BIM 和 MIM 三种攻击方法在 ℓ_2 范数的限制下对不同模型进行有目标攻击的黑盒迁移攻击成功率。从实验结果中可以发现基于单步梯度更新的 FGSM 或 FGM 方法很难在有目标攻击场景下攻破黑盒模型，其黑盒迁移攻击成功率低于 1%。MIM 可以提高在此场景下的黑盒迁移攻击成功率，但是相比于无目标攻击其成功率仍然很低。尤其是对于集成对抗训练模型，有目标黑盒迁移攻击成功率不高于 2%。因此，如何在有目标攻击中提高黑盒迁移攻击成功率仍然是具有挑战的研究问题。

2.5　本章小结

本章提出了动量迭代法，在对抗样本生成的迭代优化过程中引入动量项，用于记录梯度更新的历史方向。动量迭代法可以使对抗样本生成过程中的梯度更新方向更加平稳，同时避免对抗样本落入优化问题的局部极值点，有效缓解已有方法生成对抗样本的白盒攻击效果与其迁移能力的相互制约，提高黑盒迁移攻击成功率。本章进一步提出针对多个模型 logits 层进行融合的策略。在 ImageNet 数据集上，针对多个正常训练和集成对抗训练模型的攻击结果验证了动量迭代法的有效性。相比于已有攻击方法，动量迭代法将黑盒迁移攻击成功率提高了一倍左右。本章的结果表明现有深度学习模型在黑盒迁移攻击场景下存在鲁棒性不足的问题，这也会对深度学习的实际应用带来新的安全挑战。

第 3 章　平移不变对抗攻击方法

为了解决深度学习模型在对抗攻击下的鲁棒性问题，不同类型的对抗防御方法被相继提出。一些典型的防御模型在黑盒场景下呈现出良好的防御能力，可以有效降低黑盒迁移攻击成功率。第2章提出的动量迭代法面对这些防御的攻击效果也受到了限制。为了针对对抗防御生成迁移能力更强的对抗样本，本章提出平移不变对抗攻击方法。该方法构建平移不变攻击目标函数，同时对一组经过平移变换的图片生成对抗样本。平移不变攻击方法可以有效减轻对抗样本对所采用白盒模型的依赖程度，从而提高对黑盒防御模型的迁移攻击成功率。为了更加高效地优化平移不变攻击目标函数，在卷积神经网络具有平移不变性的假设下，攻击者可以计算模型损失函数对于未平移图片的梯度，进而将该梯度与预设的核矩阵进行卷积，以近似平移不变攻击目标函数的梯度。平移不变攻击方法可以与所有基于梯度的攻击方法相结合。实验结果验证了所提出方法的有效性，其中最有效的攻击对八个典型对抗防御模型的黑盒迁移攻击成功率达到了 82%，说明了这些防御模型在黑盒迁移攻击下鲁棒性的不足。

3.1　本章引言

深度学习模型很容易被对抗样本欺骗[18-20]，这对深度学习的实际应用带来了巨大的安全威胁，同时引发了对构建更加鲁棒的深度学习模型的广泛研究。对抗防御技术通过在模型中加入不同的防御机制增强模型的鲁棒性。尽管很多防御对已有攻击方法生成的对抗样本表现出良好的防御能力，但是其对未知且更加强大的攻击方法是否仍然具有防御能力有待考究，这也成为了一个重要的研究问题。

Athalye 等人[36]发现大部分对抗防御通过引起混淆梯度（obfuscated gradients）虚假地提升模型在基于梯度的对抗攻击下的防御能力，而并没有真正增强模型的鲁棒性。以 JPEG 压缩[53-54]为例，对输入图片进行 JPEG 压缩可以使得模型损失函数对于输入数据的梯度不存在，这就导致了基于梯度的攻击方法无法生成对抗样本。然而，攻击者仍然可以通过其他攻击方式（例如黑盒攻击）攻破基于 JPEG 压缩的防御，说明其鲁棒性没有提升。一些工作[36,42,88]通过构造适应性攻击攻破了大部分防御模型，但这些攻击方法大多基于白盒场景，需要获取防御模型具体的防御机制。在黑盒迁移攻击场景下，一些典型的防御模型呈现出良好的防御能力。但是，目前这些防御对黑盒迁移攻击所展现出的防御能力很有可能是缺乏更加强大的攻击方法导致的，而不是其在黑盒迁移攻击场景下真正具有良好的鲁棒性。

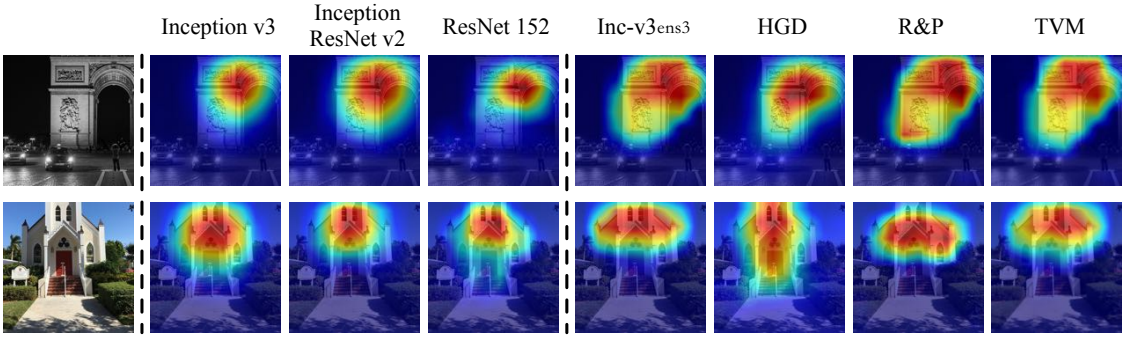


图 3.1　正常训练模型与对抗防御模型的判别区域示意图

通过对正常训练模型和对抗防御模型进行对比，本章发现了不同模型之间的区别。具体而言，对抗防御模型与正常训练模型对输入数据进行预测时所依赖的判别区域存在明显差别，在很大程度上这使得防御模型对黑盒迁移攻击具有防御能力。图3.1展示了三个正常训练模型和四个防御模型的注意力图（attention map）[110]，其中正常训练模型为 Inception v3[5]、Inception ResNet v2[107]和 ResNet 152[6]，对抗防御模型为集成对抗训练（Inc-v3$_{ens3}$）[104]、高层特征指导的去噪器（high-level representation guided denoiser，HGD）[38]、随机放缩与填充（random resize & padding，R&P）[77]和总方差最小化（total variance minimization，TVM）[54]。图中展示的注意力图可以表示不同模型对输入数据进行预测时所依赖的判别区域[110]。可以看到，正常训练模型具有比较相似的注意力图，而防御模型的注意力图表现出明显的不同。Tsipras 等人[45]也发现了类似的现象，即防御模型在输入空间中的梯度与图像的语义比较吻合，而正常训练模型在输入空间中的梯度与白噪声类似。产生这些现象的原因是防御模型使用不同的数据分布进行训练[104]或在分类前对输入数据进行预处理[38,54,77]，这就导致其判别区域出现差异。黑盒迁移攻击通常使用一个原始样本针对白盒模型生成对抗样本，使得对抗样本与白盒模型在输入空间中的判别区域或梯度高度相关，从而很难有效攻破其他依赖不同决策区域的防御模型。因此，一些典型的防御模型对黑盒迁移攻击呈现出良好的防御能力。

为了减轻对抗样本对所采用白盒模型的依赖程度，提高其对黑盒防御模型的迁移攻击成功率，本章提出*平移不变攻击方法*（translation-invariant attack method）。该方法构建平移不变攻击目标函数，针对原始图片及其经过平移变换后的图片组成的集合生成对抗样本。然而，优化平移不变攻击目标函数需要计算模型损失函数对于所有经过平移变换的图片的梯度，这会带来较高的计算复杂度。为了提升攻击效率，在卷积神经网络（Convolutional Neural Network，CNN）具有平移不变性（translation invariance）的假设下，攻击者可以计算模型损失函数对于未平移图片的梯度，进而将该梯度与预设的核矩阵进行卷积，以近似平移不变攻击目标函数的梯度。平移不变攻击方法可以与快速梯度符号法（FGSM）[20]、动量迭代法（MIM）

图 3.2　使用 FGSM 和 TI-FGSM 生成的对抗样本示例

在内的所有基于梯度的攻击方法相结合，并且在不增加计算复杂度的情况下提高对抗样本对防御模型的黑盒迁移攻击成功率。例如，将平移不变攻击方法与快速梯度符号法相结合可以得到平移不变快速梯度符号法（TI-FGSM）。图3.2展示了使用 FGSM 和 TI-FGSM 生成的对抗样本示例。

本章在 ImageNet[87] 数据集上针对八个典型的对抗防御模型进行实验。平移不变攻击方法可以大幅提高针对防御模型的黑盒迁移攻击成功率。实验中最有效的攻击对这八个防御模型的平均黑盒迁移攻击成功率达到了 82%，验证了所研究的对抗防御模型在黑盒迁移攻击场景下鲁棒性的不足。

3.2　相关工作

本节将介绍与本章研究内容相关的一些研究工作，包括针对对抗防御模型的适应性攻击以及针对一组图片生成对抗样本的攻击方法。

对抗防御旨在增强深度学习模型在对抗场景下的鲁棒性与安全性，防止攻击者对实际应用进行恶意攻击。尽管不同对抗防御方法对已有对抗攻击可以有效防御，但是大部分防御很快就会被新提出的攻击方法攻破[36,42,88]。攻击者通常可以针对不同的防御模型设计适应性攻击，以攻破特定的防御。例如，在防御方法引起混淆梯度的情况下，攻击者可以通过后向传递可微近似（BPDA）[36]等方式构造可以攻破防御模型的适应性攻击。尽管已有工作说明了很多防御方法在白盒场景下不具有防御能力，但是一些典型的防御[38,54,77,104]通过实验展现出其在黑盒迁移攻击场景下的防御能力。本章研究如何在黑盒迁移攻击场景下针对防御模型构造

更加有效的攻击方法，从而说明这些防御在此场景下的脆弱性。

本章提出的方法针对一组经过平移变换的图片生成对抗样本，类似的方法也在文献[55,111]中被讨论。Moosavi 等人[111]对给定数据集中的所有图片生成一个通用对抗扰动（universal adversarial perturbation），使其添加至大部分原始样本后均可以欺骗目标模型。Athalye 等人[55]提出期望变换（EoT）方法，针对一组图像变换生成对抗样本，该方法与本章所提出的方法主要有以下三个方面的区别。第一，本章目标是针对对抗防御模型生成迁移能力更好的对抗样本，而期望变换方法聚焦如何在真实世界中生成对抗样本。第二，本章的方法仅使用了平移变换，而期望变换方法使用了多种变换，包括旋转、平移、添加高斯噪声等。第三，本章针对平移不变攻击目标函数开发了高效的优化算法，只需要计算模型对于未平移图片的梯度，而期望变换方法通过随机采样的方式计算模型对于一批变换图片的梯度。

3.3　算法设计

本节将介绍平移不变攻击方法（translation-invariant attack method）的具体细节。与第2章的符号定义相同，本章使用 x 代表输入数据（即一张图片），y 代表 x 的真实类别。本章考虑在 ℓ_p 范数的限制下进行无目标对抗攻击，即生成对抗样本 x^* 使得分类器 f 分类错误，并且满足 $\|x^* - x\|_p \leqslant \epsilon$，其中 ϵ 代表最大扰动规模。特别地，本章仅考虑在 ℓ_∞ 范数限制下的攻击，所提出的方法可以被简单地扩展到 ℓ_2 范数限制下的攻击。已有对抗攻击方法为了在扰动范围内生成对抗样本，通常会最大化模型损失函数 $\boldsymbol{J}$，如公式（2.1）所示。

3.3.1　平移不变攻击目标函数

如图3.1所示，对抗防御模型与正常训练模型对输入数据进行预测时所依赖的判别区域存在明显差别。已有攻击方法在生成对抗样本的过程中仅对于原始样本 x 最大化模型的损失函数 $\boldsymbol{J}$，这会使得生成的对抗样本 x^* 与白盒模型的判别区域或梯度高度相关。例如，使用快速梯度符号法生成的对抗扰动即为模型损失函数对于原始样本 x 的梯度符号方向。因此，对于其他具有不同判别区域或梯度的黑盒防御模型，对抗样本很难具有迁移能力。从而使得一些典型的防御模型对黑盒迁移攻击呈现出良好的鲁棒性。

为了减轻对抗样本对白盒模型判别区域的依赖程度，提高其对黑盒模型的迁移攻击成功率，本章提出平移不变攻击方法。与已有攻击方法仅对原始图片最大化模型的损失函数不同，平移不变攻击方法计算模型对一组经过平移变换的图片

的损失并对其加权平均，得到平移不变攻击目标函数，如下所示：

$$\arg\max_{x^*} \sum_{i,j} w_{ij} J(f(T_{ij}(x^*)), y), \quad \text{s.t. } \|x^* - x\|_\infty \leqslant \epsilon, \tag{3.1}$$

其中 $J(\cdot,\cdot)$ 代表模型的损失函数（如交叉熵损失），T_{ij} 代表图像平移变换，其将图片 x 向两个维度分别平移 i 和 j 个像素，即平移图片 $T_{ij}(x)$ 的第 (a,b) 个像素为 $T_{ij}(x)_{a,b} = x_{a-i,b-j}$。此外，$w_{ij}$ 代表模型以平移图片 $T_{ij}(x)$ 作为输入计算得到的损失 $J(f(T_{ij}(x^*)), y)$ 所占权重，满足 $w_{ij} \geqslant 0$ 且 $\sum_{i,j} w_{ij} = 1$。平移不变攻击目标函数（3.1）设置平移量 i、j 的取值范围为 $i, j \in \{-k, \cdots, 0, \cdots, k\}$，其中 k 代表最大平移量。可以看到，如果设置 $w_{00} = 1$ 且其他权重为 0，平移不变攻击目标函数退化为一般的攻击目标函数，如公式（2.1）所示。通过优化平移不变攻击目标函数，生成的对抗样本不会过于依赖所采用白盒模型的判别区域，对防御模型的黑盒迁移攻击效果也会得到增强。本章仅采用平移变换构建攻击目标函数，而没有采用如旋转、放缩等其他变换，其背后的原因是：基于卷积神经网络的平移不变性，平移不变攻击目标函数可以被高效求解。

3.3.2　梯度计算

为了求解并优化平移不变攻击目标函数（3.1）以生成对抗样本，优化过程中的每一轮迭代需要计算模型损失函数对于共计 $(2k+1)^2$ 张不同平移图片的梯度，这会带来很高的计算复杂度。为了解决此问题，一个可行的方案是从所有的平移图片中随机选取一部分计算梯度[36]，但是这种方法得到的梯度具有随机性，且效率较低。本小节展示在卷积神经网络具有平移不变性的假设下，平移不变攻击目标函数的梯度可以被高效计算。

卷积神经网络在设计之初就考虑到了平移不变性[112]，即输入图片中的目标物体无论出现在什么位置都可以被正确识别。然而，由于汇聚（pooling）等结构的存在，卷积神经网络并不能完全保持平移不变性[113-114]。本章提出的平移不变攻击目标函数通常会选取较小的平移变换，在一般情况下图片向每个维度平移的距离不会超过 10 个像素，即 $k \leqslant 10$。在这种情况下，卷积神经网络的平移不变性基本可以得到保证，同时第3.4.1节中的实验会对其进行验证。因此，本小节假设卷积神经网络具有平移不变性，即模型对平移后的图片 $T_{ij}(x)$ 与原始图片 x 计算的损失相同，并且假设模型对于它们的梯度也相同，可以表示为：

$$\nabla_x J(f(x), y)\big|_{x=T_{ij}(\hat{x})} \approx \nabla_x J(f(x), y)\big|_{x=\hat{x}}, \tag{3.2}$$

其中 $\hat{x}$ 代表给定的输入图片。基于以上假设，平移不变攻击目标函数（3.1）对于

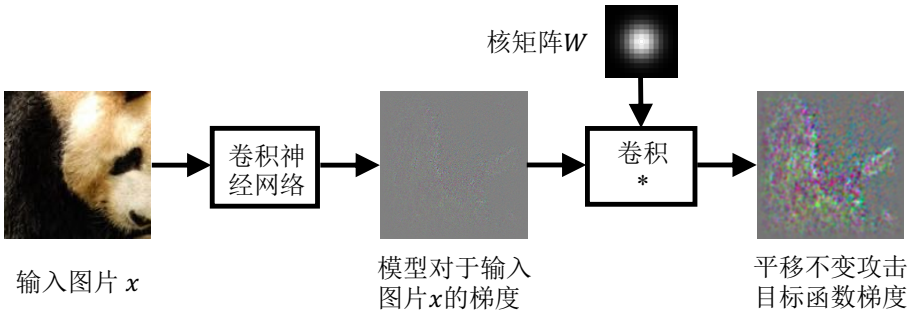


图 3.3　计算平移不变攻击目标函数梯度的示意图

输入数据 $\hat{x}$ 的梯度可以进行如下近似：

$$
\begin{aligned}
&\nabla_x \Big(\sum_{i,j} w_{ij} J(f(T_{ij}(x)), y) \Big) \Big|_{x=\hat{x}} \\
&= \sum_{i,j} w_{ij} \nabla_x J(f(T_{ij}(x)), y) \big|_{x=\hat{x}} \\
&= \sum_{i,j} w_{ij} \left(\nabla_{T_{ij}(x)} J(f(T_{ij}(x)), y) \cdot \frac{\partial T_{ij}(x)}{\partial x} \right) \Big|_{x=\hat{x}} \\
&= \sum_{i,j} w_{ij} T_{-i-j} \left(\nabla_x J(f(x), y) \big|_{x=T_{ij}(\hat{x})} \right) \\
&\approx \sum_{i,j} w_{ij} T_{-i-j} \left(\nabla_x J(f(x), y) \big|_{x=\hat{x}} \right).
\end{aligned}
\tag{3.3}
$$

从公式（3.3）中可以看到，平移不变攻击不需要计算模型对于 $(2k+1)^2$ 张图片的梯度，而可以仅计算模型对于未平移的原始图片 $\hat{x}$ 的梯度，进而将此梯度进行平移并加权平均。这一过程和卷积操作等价，如下所示：

$$
\sum_{i,j} w_{ij} T_{-i-j} \left(\nabla_x J(f(x), y) \big|_{x=\hat{x}} \right) \Leftrightarrow W * \nabla_x J(f(x), y) \big|_{x=\hat{x}}, \tag{3.4}
$$

其中 W 代表核矩阵（kernel matrix），其维度为 $(2k+1) \times (2k+1)$，具体取值为 $W_{i,j} = w_{-i-j}$。图3.3展示了计算平移不变攻击目标函数梯度的示意图。通过此方式，平移不变攻击目标函数的梯度可以通过一次梯度计算得到，计算复杂度相比于已有方法并没有增加。

此外，平移不变攻击方法并不限制核矩阵的具体选取方式。本小节进一步提出三种不同的核矩阵，包括：

- 均匀核矩阵（uniform kernel matrix）：其取值为 $W_{i,j} = \frac{1}{(2k+1)^2}$，即模型对于所有平移变换的图片计算的损失所占权重相同；
- 线性核矩阵（linear kernel matrix）：其取值为 $W_{i,j} = \frac{\tilde{W}_{i,j}}{\sum_{i,j} \tilde{W}_{i,j}}$，其中 $\tilde{W}_{i,j} = (1 - \frac{|i|}{k+1}) \cdot (1 - \frac{|j|}{k+1})$，即模型对于平移变换的图片计算的损失所占权重随平移距离线性变化；

算法 3.1 平移不变动量迭代法

输入：分类器 f，损失函数 J，原始样本 x，真实类别 y，扰动规模 ϵ，迭代轮数 T，动量衰减系数 μ，核矩阵 W；

输出：满足 $\|x^* - x\|_\infty \leqslant \epsilon$ 的对抗样本 x^*；

令 $\alpha = \frac{\epsilon}{T}$；

令 $g_0 = 0$，$x_0^* = x$；

对 $t = 0, \cdots, T-1$ **执行**

　将 x_t^* 输入分类器 f 并计算损失函数的梯度 $\nabla_x J(f(x_t^*), y)$；

　通过平移不变攻击目标函数的梯度更新动量项：

$$g_{t+1} = \mu \cdot g_t + \frac{W * \nabla_x J(f(x_t^*), y)}{\|W * \nabla_x J(f(x_t^*), y)\|_1}; \tag{3.5}$$

　更新对抗样本：

$$x_{t+1}^* = x_t^* + \alpha \cdot \text{sign}(g_{t+1}); \tag{3.6}$$

返回：$x^* \leftarrow x_T^*$。

- 高斯核矩阵（Gaussian kernel matrix）：其取值为 $W_{i,j} = \frac{\tilde{W}_{i,j}}{\sum_{i,j} \tilde{W}_{i,j}}$，其中 $\tilde{W}_{i,j} = \frac{1}{2\pi\sigma^2} \exp(-\frac{i^2+j^2}{2\sigma^2})$，且 $\sigma = \frac{k}{\sqrt{3}}$，即模型对于平移变换的图片计算的损失所占权重随平移距离呈高斯概率密度函数变化。

虽然本小节介绍的三种核矩阵均通过人工设计得到，但是本章提出的平移不变攻击方法并不局限于所介绍的核矩阵。攻击者可以设计更好的核矩阵进一步增强平移不变攻击方法的有效性。

3.3.3　攻击算法

上一节仅介绍了如何计算平移不变攻击目标函数的梯度，但并没有具体指出如何利用计算得到的梯度生成对抗样本。这意味着本章提出的平移不变攻击方法可以与所有基于梯度的攻击方法相结合。基于梯度的攻击方法需要计算模型损失函数对于输入数据的梯度，并基于此梯度通过不同的攻击方式生成对抗样本。将平移不变攻击方法与基于梯度的攻击方法相结合，可以将原本方法中的梯度替换为公式（3.4）中推导出的平移不变攻击目标函数的梯度，从而得到不同的攻击算法。

具体而言，将平移不变攻击方法与快速梯度符号法（FGSM）相结合可以得到平移不变快速梯度符号法（TI-FGSM），其生成对抗样本的过程可以被描述为：

$$x^* = x + \epsilon \cdot \text{sign}(W * \nabla_x J(f(x), y)). \tag{3.7}$$

将平移不变攻击方法与动量迭代法（MIM）相结合可以得到平移不变动量迭代法（TI-MIM），算法3.1描述了平移不变动量迭代法的具体攻击过程。此外，平移不变攻击方法在 ℓ_2 范数限制下攻击和有目标攻击等场景下均可以应用。攻击过程与上

述方法类似，均通过将原本方法中的梯度替换为平移不变攻击目标函数的梯度实现，本小节不再赘述。

3.4　实验结果

本节在 NeurIPS 2017 对抗攻防竞赛使用的数据集①上进行实验②。该数据集总共包含 1000 张图片，且与 ImageNet[87]数据集分布一致。

本节总共选取八个在 ImageNet 数据集上对黑盒迁移攻击比较有效的防御模型，包括：

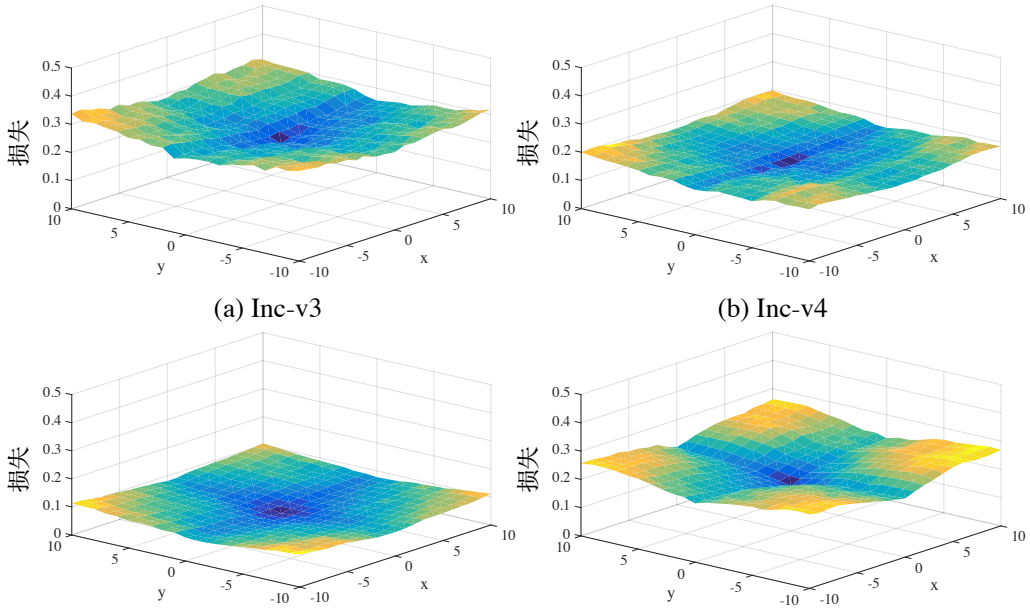
- 集成对抗训练模型 Inc-v3$_{ens3}$、Inc-v3$_{ens4}$ 和 IncRes-v2$_{ens}$[104]；
- 高层特征指导的去噪器（HGD）[38]，该防御获得了 NeurIPS 2017 对抗攻防竞赛防御赛道的第一名；
- 图像随机放缩与填充（R&P）[77]，该防御获得了 NeurIPS 2017 对抗攻防竞赛防御赛道的第二名；
- 基于 JPEG 压缩和总方差最小化（TVM）的图像变换防御[54]；
- NeurIPS 2017 对抗攻防竞赛防御赛道的第三名（记为 NeurIPS-r3）③。

为了针对这些防御进行黑盒迁移攻击，本节选取四个正常训练的模型作为白盒模型生成对抗样本，包括 Inception v3（记为 Inc-v3）[5]、Inception v4（记为 Inc-v4)[107]、Inception ResNet v2（记为 IncRes-v2）[107]和 ResNet v2 152（记为 Res-152）[108]。

实验将平移不变攻击方法与快速梯度符号法（FGSM）[20]、动量迭代法（MIM）和多样输入法（Diverse Inputs Method，DIM）[56]相结合，其中多样输入法在对抗样本生成的每一轮梯度迭代中对输入数据进行随机变换提升对抗样本的迁移能力。结合后的方法分别记为平移不变快速梯度符号法（TI-FGSM）、平移不变动量迭代法（TI-MIM）和平移不变多样输入法（TI-DIM）。

本节主要在 ℓ_∞ 范数的限制下进行无目标攻击实验。在所有实验中，最大扰动规模均设置为 $\epsilon = 16$，图片像素取值范围为 $[0, 255]$。对于基于多步梯度迭代的方法，迭代轮数设置为 $T = 10$，迭代步长设置为 $\alpha = 1.6$。MIM 和 TI-MIM 中的动量衰减系数设置为 $\mu = 1.0$。值得注意的是，这些参数的设置仅与原始攻击方法相关，本章提出的平移不变攻击方法可以适用于上述参数的不同取值。

① 数据集参见https://github.com/cleverhans-lab/cleverhans/tree/master/cleverhans_v3.1.0/examples/nips17_adversarial_competition。
② 源代码参见https://github.com/dongyp13/Translation-Invariant-Attacks。
③ 代码参见https://github.com/anlthms/nips-2017/tree/master/mmd。



(a) Inc-v3　(b) Inc-v4

(c) IncRes-v2　(d) Res-152

图 3.4　卷积神经网络在平移变换下的损失平面

3.4.1　卷积神经网络的平移不变性

本小节首先验证卷积神经网络的平移不变性。实验使用数据集中原始的 1000 张图片，并将它们向每个维度平移 −10 到 10 个像素。将原始图片以及经过平移变换后的图片分别输入 Inc-v3、Inc-v4、IncRes-v2 和 Res-152 四个模型并计算模型的损失。图3.4展示了模型在平移变换下的损失平面，其中模型在每一个平移变换下的损失为所有平移图片损失的平均值。

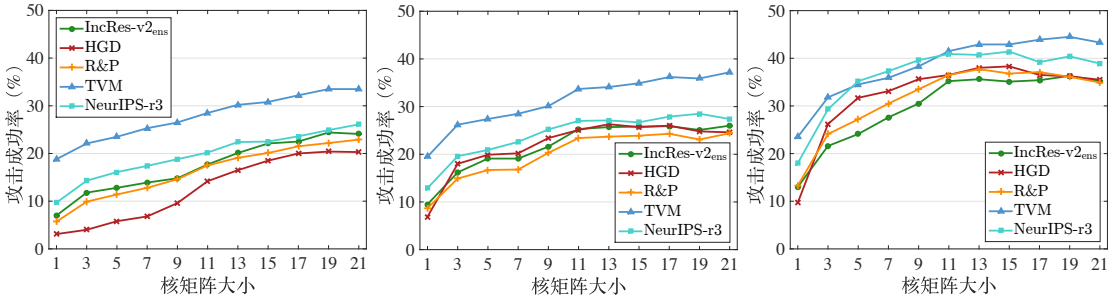
从图3.4中可以看到，卷积神经网络在每个维度平移 −10 到 10 个像素内的损失平面是比较平滑的。这可以说明在平移变换比较小的情况下，卷积神经网络具有平移不变性。在平移不变攻击目标函数中，图片向每个维度的平移距离不会超过 10 个像素，所以原始图片和平移图片的损失非常接近。因此，平移不变攻击方法可以假设模型对平移后的图片与原始图片计算的损失相同，并基于此假设推导平移不变攻击目标函数的梯度。

3.4.2　不同核矩阵的攻击结果

本小节针对平移不变攻击方法中不同核矩阵的选取进行实验。实验使用 TI-FGSM、TI-MIM 和 TI-DIM 三种方法攻击 Inc-v3 模型，其中每种方法分别使用第3.3.2节中介绍的三种不同的核矩阵，即均匀核矩阵、线性核矩阵和高斯核矩阵。表3.1展示了使用不同核矩阵的攻击方法对八个对抗防御模型的黑盒迁移攻击成功率，其中攻击成功率为相应模型将不同方法生成的对抗样本作为输入计算的分类错误率。

表 3.1　平移不变攻击方法采用不同核矩阵的攻击成功率（%）对比

攻击方法	核矩阵	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$	HGD	R&P	JPEG	TVM	NeurIPS-r3
TI-FGSM	均匀	25.0	27.9	21.1	15.7	19.1	24.8	32.3	21.9
	线性	**30.7**	**32.4**	**24.2**	**20.9**	**23.3**	**28.1**	**34.6**	**25.8**
	高斯	28.2	28.9	22.3	18.4	19.8	25.5	30.7	24.5
TI-MIM	均匀	30.0	32.2	22.8	21.7	22.8	26.4	32.7	25.9
	线性	**35.8**	35.0	**26.8**	25.5	23.4	**29.0**	**35.8**	**27.5**
	高斯	**35.8**	**35.1**	25.8	**25.7**	**23.9**	28.2	34.9	26.7
TI-DIM	均匀	32.6	34.6	25.6	24.1	27.2	30.2	34.9	28.8
	线性	45.2	47.0	34.9	35.6	35.2	**38.5**	43.6	39.7
	高斯	**46.9**	**47.1**	**37.4**	**38.3**	**36.8**	37.0	**44.2**	**41.4**



(a) TI-FGSM　(b) TI-MIM　(c) TI-DIM

图 3.5　攻击成功率随核矩阵大小变化图

从表3.1展示的结果中可以看到，对于 TI-FGSM，线性核矩阵相比于均匀核矩阵和高斯核矩阵取得了更高的攻击成功率。而对于 TI-MIM 和 TI-DIM，高斯核矩阵取得了与线性核矩阵相似的攻击效果。在所有情况下，线性核矩阵和高斯核矩阵都比均匀核矩阵取得的攻击效果更好。这表明在核矩阵的设计中，应该对于平移距离较小的变换设置更高的权重，而对于平移距离较大的变换设置更低的权重。基于本小节的实验分析，以下实验均采用高斯核矩阵。

3.4.3　核矩阵大小对攻击结果的影响

平移不变攻击方法中核矩阵的大小也对黑盒迁移攻击成功率起着关键的作用。如果核矩阵大小为 1×1，则平移不变攻击目标函数退化为原始攻击目标函数，基于平移不变的攻击方法也退化为其原始版本。因此，本小节通过实验验证核矩阵大小对攻击结果的影响。

实验使用 TI-FGSM、TI-MIM 和 TI-DIM 攻击 Inc-v3 模型，其中核矩阵选取为高斯核矩阵，其大小从 1×1 增长至 21×21，变化粒度为 2。由于核矩阵的长度和

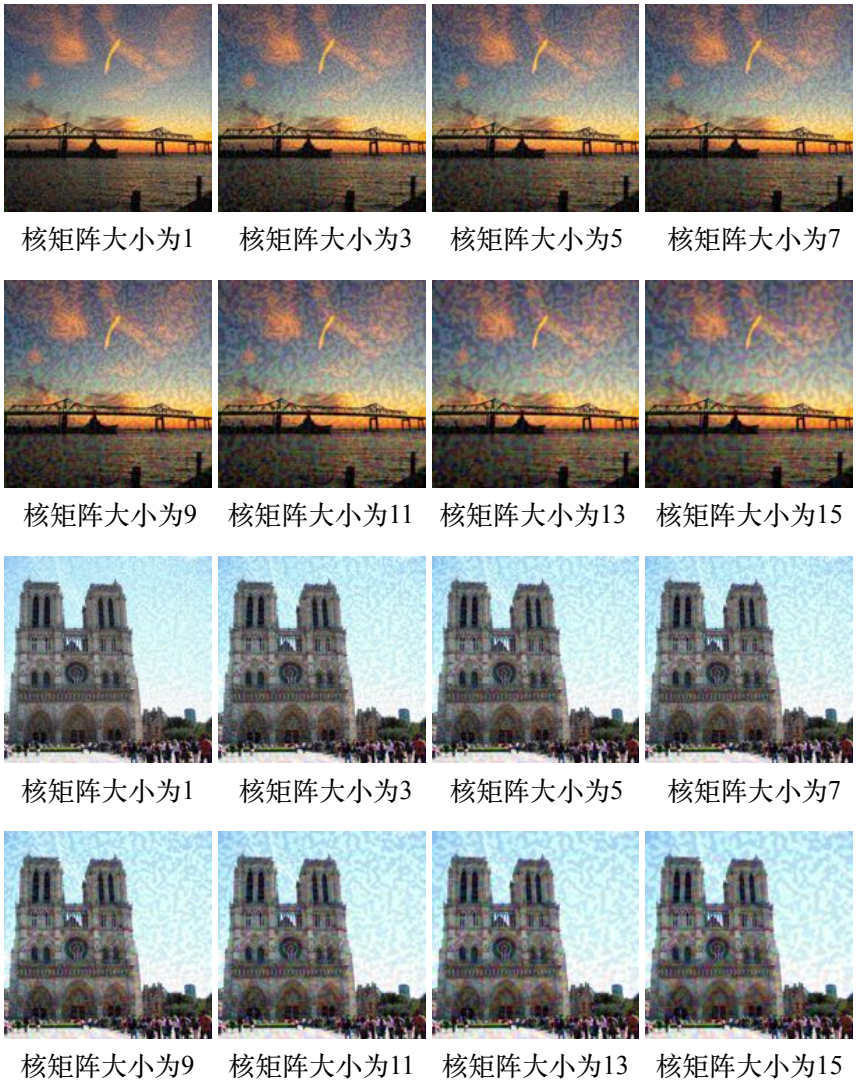


图 3.6　在不同核矩阵大小下 TI-FGSM 生成的对抗样本示例

宽度相等，为了叙述简便本小节使用其长度代表核矩阵大小，即核矩阵大小的变化范围为 1 至 21。图3.5展示了在不同核矩阵大小下 TI-FGSM、TI-MIM 和 TI-DIM 生成的对抗样本对 IncRes-v2$_{ens}$、HGD、R&P、TVM 和 NeurIPS-r3 五个防御模型的黑盒迁移攻击成功率变化曲线。可以看到，黑盒迁移攻击成功率随着核矩阵大小的增长不断提高，但在核矩阵大小超过 15 后趋于稳定。以下实验将核矩阵大小设置为 15。

图3.6展示了在不同核矩阵大小下使用 TI-FGSM 攻击 Inc-v3 模型生成的对抗样本示例。在平移不变攻击方法中，更新对抗样本的梯度通过卷积操作得到，所以呈现出更加平滑的效果。平移不变攻击方法生成的对抗样本中添加的扰动也随着核矩阵大小的增长变得更加平滑。

表 3.2　FGSM 与 TI-FGSM 的攻击成功率（%）对比

	攻击方法	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$	HGD	R&P	JPEG	TVM	NeurIPS-r3
Inc-v3	FGSM	15.6	14.7	7.0	2.1	6.5	19.9	18.8	9.8
	TI-FGSM	**28.2**	**28.9**	**22.3**	**18.4**	**19.8**	**25.5**	**30.7**	**24.5**
Inc-v4	FGSM	16.2	16.1	9.0	2.6	7.9	21.8	19.9	11.5
	TI-FGSM	**28.2**	**28.3**	**21.4**	**18.1**	**21.6**	**27.9**	**31.8**	**24.6**
IncRes-v2	FGSM	18.0	17.2	10.2	3.9	9.9	24.7	23.4	13.3
	TI-FGSM	**32.8**	**33.6**	**28.1**	**25.4**	**28.1**	**32.4**	**38.5**	**31.4**
Res-152	FGSM	20.2	17.7	9.9	3.6	8.6	24.0	22.0	12.5
	TI-FGSM	**34.6**	**34.5**	**27.8**	**24.4**	**27.4**	**32.7**	**38.1**	**30.1**

表 3.3　MIM 与 TI-MIM 的攻击成功率（%）对比

	攻击方法	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$	HGD	R&P	JPEG	TVM	NeurIPS-r3
Inc-v3	MIM	20.5	17.4	9.5	6.9	8.7	20.3	19.4	12.9
	TI-MIM	**35.8**	**35.1**	**25.8**	**25.7**	**23.9**	**28.2**	**34.9**	**26.7**
Inc-v4	MIM	22.1	20.1	12.1	9.6	12.1	26.0	24.8	15.6
	TI-MIM	**36.7**	**39.2**	**28.7**	**27.8**	**28.0**	**31.6**	**38.4**	**29.5**
IncRes-v2	MIM	31.3	27.2	19.7	19.6	18.6	31.6	34.4	22.7
	TI-MIM	**50.7**	**51.7**	**49.3**	**45.1**	**45.2**	**45.9**	**55.4**	**46.2**
Res-152	MIM	25.1	23.7	13.3	15.1	14.6	31.2	24.5	18.0
	TI-MIM	**39.9**	**37.7**	**32.8**	**31.8**	**31.1**	**38.3**	**41.2**	**34.4**

3.4.4　攻击结果比较

本小节对比平移不变攻击 TI-FGSM、TI-MIM 和 TI-DIM 与基准方法 FGSM、MIM 和 DIM 的黑盒迁移攻击成功率。实验首先使用以上攻击方法针对 Inc-v3、Inc-v4、IncRes-v2 和 Res-152 四个白盒模型进行攻击，然后使用生成的对抗样本作为输入数据测试本节所研究的八个对抗防御模型，得到黑盒迁移攻击成功率。在实验中，平移不变攻击 TI-FGSM、TI-MIM 和 TI-DIM 均采用高斯核矩阵，其大小设置为 15。表3.2展示了 FGSM 与 TI-FGSM 对所有防御模型的黑盒迁移攻击成功率。表3.3展示了 MIM 与 TI-MIM 对所有防御模型的黑盒迁移攻击成功率。表3.4展示了 DIM 和 TI-DIM 对所有防御模型的黑盒迁移攻击成功率。

从三个表格中的实验结果可以看到，相比于基准攻击方法，平移不变攻击方

表 3.4　DIM 与 TI-DIM 的攻击成功率（%）对比

	攻击方法	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$	HGD	R&P	JPEG	TVM	NeurIPS-r3
Inc-v3	DIM	24.2	24.3	13.0	9.7	13.3	30.7	24.4	18.0
	TI-DIM	**46.9**	**47.1**	**37.4**	**38.3**	**36.8**	**37.0**	**44.2**	**41.4**
Inc-v4	DIM	28.3	27.5	15.6	14.6	17.2	38.6	29.1	14.1
	TI-DIM	**48.6**	**47.5**	**38.7**	**40.3**	**39.3**	**43.5**	**45.6**	**41.9**
IncRes-v2	DIM	41.2	40.0	27.9	32.4	30.2	47.2	41.7	37.6
	TI-DIM	**61.3**	**60.1**	**59.5**	**58.7**	**61.4**	**55.7**	**66.2**	**61.5**
Res-152	DIM	40.5	36.0	24.1	32.6	26.4	42.4	36.8	34.4
	TI-DIM	**56.1**	**55.5**	**49.5**	**51.8**	**50.4**	**50.8**	**55.7**	**52.9**

表 3.5　多模型融合下不同方法的攻击成功率（%）对比

攻击方法	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$	HGD	R&P	JPEG	TVM	NeurIPS-r3
FGSM	27.5	23.7	13.4	4.9	13.8	38.1	30.0	19.8
TI-FGSM	**39.1**	**38.8**	**31.6**	**29.9**	**31.2**	**43.3**	**39.8**	**33.9**
MIM	50.5	48.3	32.8	38.6	32.8	67.7	50.1	43.9
TI-MIM	**76.4**	**74.4**	**69.6**	**73.3**	**68.3**	**77.2**	**72.1**	**71.4**
DIM	66.0	63.3	45.9	57.7	51.7	82.5	64.1	63.7
TI-DIM	**84.8**	**82.7**	**78.0**	**82.6**	**81.4**	**83.4**	**79.8**	**83.1**

法可以在选取不同白盒模型的情况下大幅提高对防御模型的黑盒迁移攻击成功率。在通常情况下，平移不变攻击方法可以将攻击成功率提升 5% ～ 30%。值得强调的是，从表3.4中可以看到，通过将平移不变攻击方法与 DIM 相结合得到 TI-DIM 攻击，在采用 IncRes-v2 作为白盒模型的情况下生成的对抗样本对防御模型的黑盒迁移攻击成功率达到了 60% 左右。实验结果说明了这些防御模型对黑盒迁移攻击的脆弱性，同时也验证了本章所提出的平移不变攻击方法的有效性。图3.2展示了两个 FGSM 和 TI-FGSM 针对 Inc-v3 模型生成的对抗样本示例。使用 TI-FGSM 生成的对抗样本中添加的扰动更加平滑。其他平移不变攻击方法生成的对抗样本也能观察到类似的现象。

本章提出的平移不变攻击方法也可应用于针对多个白盒模型进行攻击的场景，本小节进一步展示在此场景下的实验结果。实验使用 FGSM、TI-FGSM、MIM、TI-MIM、DIM 和 TI-DIM 针对 Inc-v3、Inc-v4、IncRes-v2 和 Res-152 四个模型的融合进行攻击。对这四个白盒模型进行融合的策略选取为第2.3.2节中介绍的对多个模

型 logits 层融合的策略，其融合权重均为 $\frac{1}{4}$。表3.5展示了不同攻击方法生成的对抗样本对防御模型的黑盒迁移攻击成功率。与上述结果类似，平移不变攻击方法可以提高不同基准方法对防御模型的黑盒迁移攻击成功率。值得注意的是，TI-DIM 生成的对抗样本可以对这八个典型的对抗防御模型达到平均 82% 的黑盒迁移攻击成功率，同时所采用的白盒模型仅为四个正常训练得到的模型。实验结果进一步说明了这些防御模型的脆弱性，对构建更加有效的防御方法带来了新的挑战。

3.5　本章小结

本章提出了平移不变对抗攻击方法，可以针对对抗防御模型生成黑盒迁移攻击效果更好的对抗样本。该方法构建平移不变攻击目标函数，同时对一组经过平移变换的图片生成对抗样本，可以有效减轻对抗样本对所采用白盒模型的依赖程度。基于卷积神经网络的平移不变性，攻击者可以计算模型损失函数对于未平移图片的梯度，进而将该梯度与预设的核矩阵进行卷积，以近似平移不变攻击目标函数的梯度。在不增加计算复杂度的情况下，平移不变攻击方法可以与所有基于梯度的攻击方法相结合。实验结果验证了平移不变攻击方法的有效性，其可以大幅提高针对典型防御模型的黑盒迁移攻击成功率。通过将平移不变攻击方法与多样输入法相结合得到 TI-DIM 攻击方法，对八个典型防御模型的平均黑盒迁移攻击成功率达到了 82%。本章的结果表明了所研究的对抗防御在黑盒迁移攻击场景下的脆弱性，对构建更加鲁棒的深度学习模型带来了新的挑战。

第 4 章　面向人脸识别的高效黑盒决策攻击

人脸识别作为计算机视觉乃至人工智能领域应用最广的一项任务，近年来随着深度学习的发展取得了巨大的进步。然而，深度学习模型在对抗攻击下鲁棒性存在不足，这很可能会对真实世界中的人脸识别应用带来严重的安全威胁。本章研究在黑盒决策攻击场景下人脸识别模型的脆弱性，旨在发现人脸识别模型面临的安全问题。在此场景下，攻击者无法获取模型的梯度信息，只能通过查询的方式获取黑盒模型对输入数据的预测类别。对于真实世界中的人脸识别系统，黑盒决策攻击相比于白盒攻击更加现实。为了提升黑盒决策攻击的查询效率，本章提出进化攻击方法，在黑盒决策攻击的过程中对搜索方向的局部几何结构进行建模，并降低搜索空间的维度。实验结果表明进化攻击方法相比于已有方法可以通过更少的模型查询次数得到对人脸图片更小的扰动。本章还应用所提出的方法成功地攻破了商用人脸识别系统。

4.1　本章引言

人脸识别（face recognition）包含两个子任务[115]：人脸验证（face verification）与人脸鉴别（face identification）。其中人脸验证的目标是判断一对人脸图片是否代表同一个身份，而人脸鉴别的目标是将一张人脸图片分类为原型图像集（gallery set）中的一个身份。目前最有效的人脸识别模型[10-12,116-119]通常使用深度卷积神经网络提取人脸特征，并设计损失函数使人脸特征满足类内方差（intra-class variance）最小化和类间方差（inter-class variance）最大化。随着深度学习的快速发展，人脸识别模型取得了优越的性能，也被广泛应用于金融、支付、公共访问等众多领域的身份认证。

然而，深度学习模型很容易受到对抗样本的干扰产生预测错误，基于深度卷积神经网络的人脸识别模型也呈现出在对抗攻击下的脆弱性。例如，攻击者可以将对抗扰动添加至眼镜上，并佩戴对抗眼镜欺骗人脸识别模型[65,120]。人脸识别模型在对抗攻击下的脆弱性会对真实世界中的人脸识别应用带来严重的安全威胁。

对抗攻击作为发现模型脆弱性，评估模型鲁棒性的重要工具，可以对真实世界中的人脸识别系统进行鲁棒性测评。已有工作[65,120]针对人脸识别模型进行对抗攻击时主要基于白盒攻击场景，即攻击者需要获取模型的结构和参数信息，进而通过基于梯度的方法优化攻击目标函数生成对抗样本。但是对于真实世界中的人脸识别系统，攻击者无法获取模型的具体信息，因此白盒攻击方法难以应用。本章

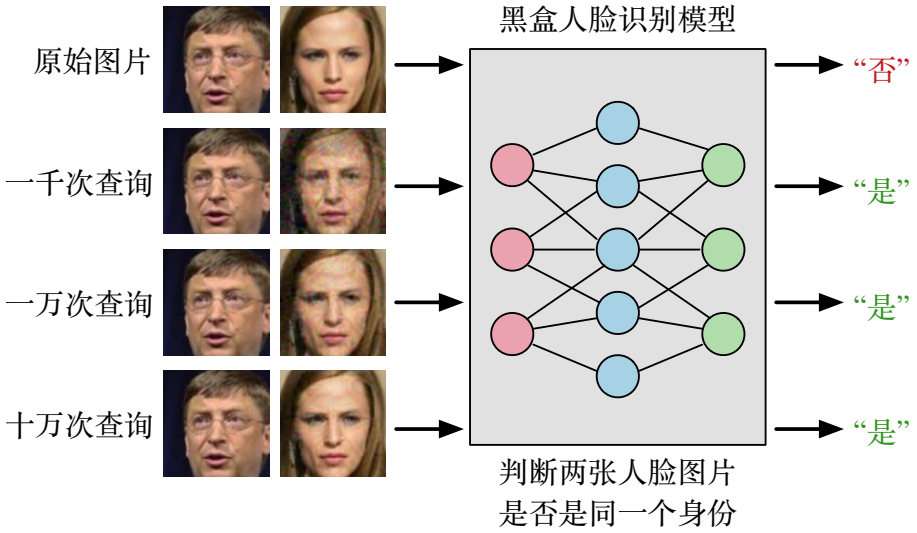


图 4.1　对人脸识别模型的黑盒决策攻击示意图

关注更加现实且通用的黑盒决策攻击场景[63]，其中攻击者无法获取模型的结构和参数等具体信息，只能通过查询的方式获取黑盒模型对输入数据的预测类别信息。黑盒决策攻击的目标是通过有限查询次数生成扰动最小的对抗样本。由于无法计算模型的梯度，黑盒决策攻击相比于白盒攻击更具挑战，同时也更加现实，这是因为真实世界中的大多数人脸识别系统都是黑盒的，且只提供查询接口返回对输入数据的预测结果。图4.1展示了对人脸识别模型进行黑盒决策攻击的示意图，攻击者通过对模型的查询不断减小对抗样本中添加的扰动。

目前已有一些方法[60,63-64]可以在黑盒决策攻击场景下生成对抗样本，包括边界攻击（Boundary attack）[63]、基于优化的攻击（Optimization-based attack）[64]等。这些方法的主要问题是对黑盒模型的查询效率较低，通常需要大量查询次数才能生成扰动较小的对抗样本，或者在有限查询次数下生成扰动较大的对抗样本。因此，本章研究如何通过更少的模型查询次数生成添加扰动更小的对抗样本，从而提升黑盒决策攻击的效率。

为了解决上述问题，本章提出进化攻击（Evolutionary attack）方法。基于给定的攻击目标函数，进化攻击通过查询黑盒模型对攻击目标函数进行高效优化。该方法在黑盒决策攻击的过程中对搜索方向的局部几何结构（local geometry）进行建模，并降低搜索空间的维度，以提升对黑盒模型的查询效率。

本章在广泛使用的 Labeled Face in the Wild（LFW）[115]人脸识别数据集上进行实验。在黑盒决策攻击场景下，实验应用所提出的进化攻击测评 SphereFace[117]、CosFace[118]和 ArcFace[119]的鲁棒性。实验结果验证了进化攻击方法的有效性，其相比于基准方法可以通过更少的模型查询次数生成扰动更小的对抗样本。本章进一步应用进化攻击方法攻破了商用人脸识别系统，表明其具有很好的实用性。

4.2　相关工作

本节将介绍与本章研究内容相关的研究工作，包括人脸识别方法、针对人脸识别模型的对抗攻击以及黑盒决策攻击方法。

随着深度学习的快速发展，人脸识别任务也取得了巨大的突破。在早期工作中，DeepFace[10]和 DeepID[11]将人脸识别视为一个多类别分类问题，并使用深度卷积神经网络学习人脸特征。文献[12,116]分别提出三元组损失（triplet loss）和中心损失（center loss）增加人脸特征在欧几里德空间中的类间间距。在近期工作中，SphereFace[117]提出角度 softmax 损失（angular softmax loss）学习在角度空间中更具判别能力的特征。CosFace[118]提出大间距余弦损失（large margin cosine loss）最大化类间余弦间距。ArcFace[119]提出加性角度间距损失（additive angular margin loss）进一步提升人脸特征的判别能力。

深度学习模型很容易受到对抗样本的干扰产生错误预测[18-20]，基于深度卷积神经网络的人脸识别模型也显示出在对抗攻击下的脆弱性。Sharif 等人[65]将对抗扰动限制在眼镜区域内并采用基于梯度的方法生成对抗眼镜，在真实世界中可以成功欺骗人脸识别系统。Sharif 等人[120]进一步采用生成模型（generative model）生成对抗眼镜，可以取得更好的攻击效果。然而，已有方法均依赖白盒攻击场景，需要获取人脸识别模型的结构和参数等具体信息，这在实际应用中难以实现。与之不同，本章关注在黑盒决策攻击场景下人脸识别模型的鲁棒性。

在黑盒决策攻击场景下，攻击者可以查询黑盒模型并获取模型对输入数据的预测类别生成对抗样本。Brendel 等人[63]提出边界攻击方法，在模型的决策边界上进行随机游走不断减小对抗样本中添加的扰动。Cheng 等人[64]提出基于优化的攻击方法，构建连续的攻击目标函数，然后对攻击目标函数的梯度进行估计生成对抗样本。Ilyas 等人[60]通过模型的预测类别估计其预测概率分布，然后使用自然进化策略（NES）[62]最大化目标类别预测概率或最小化真实类别预测概率，实现有目标和无目标攻击。这些方法存在的问题是攻击效率较低，通常需要对黑盒模型进行大量查询才能生成扰动较小的对抗样本，或者在查询次数有限的情况下生成扰动较大的对抗样本。

4.3　面向人脸识别的攻击场景

本节将首先介绍面向人脸识别模型的对抗攻击的符号表示与问题定义，然后介绍人脸识别任务中的威胁模型。

4.3.1　符号表示与问题定义

本章使用 $f(x):\mathcal{X}\to\mathcal{Y}$ 代表人脸识别模型，其中 $\mathcal{X}\subset\mathbb{R}^n$ 代表输入空间，n 为输入空间维度，$\mathcal{Y}$ 代表预测类别集合。人脸识别模型 $f(x)$ 对输入的人脸图片 $x\in\mathcal{X}$ 进行预测并输出其类别。对于人脸验证（face verification），模型通常会接收另外一张人脸图片作为输入，并判断这一对人脸图片是否属于同一个身份，其预测类别为 $\mathcal{Y}=\{0,1\}$。对于人脸鉴别（face identification），模型会将输入图片 x 与原型图像集（gallery set）中的所有图片进行比对，并将 x 分类为其中的一个身份。人脸鉴别也可以被认为是一个多类别分类任务，其预测类别为 $\mathcal{Y}=\{1,2,\cdots,L\}$，其中 L 代表所有身份数量。尽管这两个人脸识别子任务会使用额外的一张图片或原型图像集对输入数据 x 进行身份识别，但是本章为了叙述简洁，将人脸识别模型记为与额外数据无关的 $f(x)$。

给定一张原始人脸图片 x，攻击者的目标是在 x 的邻域内寻找可以被人脸识别模型预测错误的对抗样本 x^*。对抗样本可以通过求解约束优化问题生成，如下所示：

$$\arg\min_{x^*}\mathcal{D}(x^*,x),\qquad \text{s.t. } \mathcal{C}(f(x^*))=1, \tag{4.1}$$

其中 $\mathcal{D}(\cdot,\cdot)$ 代表距离度量函数，$\mathcal{C}(\cdot)$ 代表对抗判别准则（adversarial criterion），其当攻击成功时取值为 1，否则为 0。本章采用 ℓ_2 范数作为距离度量 $\mathcal{D}$。约束优化问题（4.1）的目标是在保证对抗样本 x^* 成功攻破模型 f 的情况下，最小化添加扰动的大小，即对抗样本与原始样本之间的距离 $\mathcal{D}(x^*,x)$。约束优化问题（4.1）可以被等价地转换为一个无约束优化问题，如下所示：

$$\arg\min_{x^*}\mathcal{L}(x^*)=\mathcal{D}(x^*,x)+\delta\big(\mathcal{C}(f(x^*))=1\big), \tag{4.2}$$

其中 $\mathcal{L}(\cdot)$ 代表攻击目标函数，$\delta(\cdot)$ 函数以对抗判别准则是否成立作为输入，当输入（记为 a）为真时其取值 $\delta(a)=0$，否则 $\delta(a)=+\infty$。可以看到，攻击目标函数 $\mathcal{L}(x^*)$ 在 x^* 不满足对抗判别准则时取值为 $+\infty$，而在 x^* 满足对抗判别准则时取值为 x^* 与 x 之间的距离。通过求解优化问题（4.2），攻击者可以得到添加最小扰动的对抗样本 x^*。值得注意的是，攻击目标函数 $\mathcal{L}(x^*)$ 为不连续函数。这是因为在黑盒决策攻击场景下，模型只提供对输入数据的预测类别，所以攻击者只能判断对抗样本是否攻击成功，而不能定义类似于交叉熵损失的连续函数作为攻击目标函数，这也使得黑盒决策攻击更具挑战。

4.3.2　人脸识别中的威胁模型

在人脸识别任务中，根据攻击者的不同目标，对抗攻击可以被分为*躲避攻击*（dodging attack）和*伪装攻击*（impersonation attack）。

躲避攻击的目标是使人脸识别模型将对抗样本识别错误或不能识别出其身份。在人脸信息被过度采集并滥用的情况下，躲避攻击有助于保护用户的人脸隐私。对于人脸验证，给定属于同一个身份的一对人脸图片，躲避攻击向其中一张图片添加微小的扰动，使模型将这一对图片识别为不同的身份，其攻击判别准则可以被描述为 $C(f(x^*)) = \mathbb{I}(f(x^*) = 0)$，其中 $\mathbb{I}$ 代表指示函数。对于人脸鉴别，躲避攻击的目标是使人脸识别模型将对抗样本错分为其他身份，其攻击判别准则可以被描述为 $C(f(x^*)) = \mathbb{I}(f(x^*) \neq y)$，其中 y 代表原始样本 x 的真实身份类别。

伪装攻击的目标是使人脸识别模型将对抗样本识别为给定的目标身份。在与安全密切相关的人脸识别应用（如刷脸支付）中，伪装攻击可以攻破人脸认证系统，带来严重的安全威胁。对于人脸验证，给定属于不同身份的一对人脸图片，伪装攻击对其中一张图片生成对抗样本，使模型将这一对图片识别为相同的身份，其攻击判别准则可以被描述为 $C(f(x^*)) = \mathbb{I}(f(x^*) = 1)$。对于人脸鉴别，伪装攻击生成的对抗样本需要被人脸识别模型分类为指定的目标身份 y^*，其攻击判别准则可以被描述为 $C(f(x^*)) = \mathbb{I}(f(x^*) = y^*)$。

4.4　进化攻击

在黑盒决策攻击场景下，攻击者无法通过基于梯度的方法优化攻击目标函数 $\mathcal{L}(x^*)$，只能查询黑盒模型并获取其对输入数据的预测类别，然后采用黑盒优化方法最小化 $\mathcal{L}(x^*)$。在黑盒优化中，一些典型方法采用有限差分（finite difference）[59]等方式估计目标函数的梯度，并通过梯度下降对目标函数进行优化。这些方法通常需要获取模型对输入数据的预测概率分布，更适用于黑盒得分攻击。在模型只提供预测类别的情况下，攻击目标函数 $\mathcal{L}(x^*)$ 是不连续的，从而不可导，所以不能直接使用梯度估计方法。文献[60,64]将攻击目标函数 $\mathcal{L}(x^*)$ 转换为其他的连续函数，然后使用梯度估计方法进行优化。但是这些方法需要计算输入数据到决策边界的距离或通过模型的预测类别估计其预测概率分布，导致这些方法的查询效率较低。因此，本节考虑如何利用模型的预测类别更加高效地优化攻击目标函数 $\mathcal{L}(x^*)$。

本章提出进化攻击（Evolutionary attack）方法优化攻击目标函数 $\mathcal{L}(x^*)$ 生成对抗样本。该方法基于协方差矩阵自适应进化策略（Covariance Matrix Adaptation Evolution Strategy，CMA-ES）[121]的有效变种 (1+1)-CMA-ES[122]。(1+1)-CMA-ES 通过启发式搜索求解黑盒优化问题，在每轮迭代中的主要步骤包括：1）向父代解

算法 4.1 进化攻击算法流程

输入：攻击目标函数 $\mathcal{L}(x^*)$，原始人脸图片 x，输入空间维度 n，搜索空间维度 m，随机坐标选取数量 k，对模型的总查询次数 T；
输出：对抗样本 x^*；
1: 初始化 $\mathbf{C}=\mathbf{I}_m$，$p_c=\mathbf{0}$，$\sigma,\mu,c_c,c_{cov}\in\mathbb{R}_+$，$\tilde{x}^*\in\mathbb{R}^n$；
2: **对** $t=1,\cdots,T$ **执行**
3:　　采样 $z\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{C})$;
4:　　从 m 个坐标中以 $\mathbf{C}$ 的对角元素作为概率采样 k 个坐标;
5:　　将 z 中未被选取的坐标设为 0;
6:　　通过双线性差值将 z 映射为 $\mathbb{R}^n$ 空间中的 $\tilde{z}$;
7:　　令 $\tilde{z}\leftarrow\tilde{z}+\mu(x-\tilde{x}^*)$;
8:　　**如果** $\mathcal{L}(\tilde{x}^*+\tilde{z})<\mathcal{L}(\tilde{x}^*)$ **则**
9:　　　$\tilde{x}^*\leftarrow\tilde{x}^*+\tilde{z}$;
10:　　　通过公式（4.3）和公式（4.4）分别更新 p_c 和 $\mathbf{C}$；
11: **返回：**$x^*\leftarrow\tilde{x}^*$。

（即当前解）添加随机噪声生成一个子代解（即候选解）；2）计算目标函数在父代解和子代解下的取值；3）从中选取目标函数更优的解作为下一轮迭代的父代解。虽然 (1+1)-CMA-ES 是求解黑盒优化问题的典型方法，但是直接应用 (1+1)-CMA-ES 优化攻击目标函数 $\mathcal{L}(x^*)$ 仍然存在效率较低的问题。这是因为攻击目标函数具有一些独特的性质，且输入空间 $\mathcal{X}$ 维度较高。本节在 (1+1)-CMA-ES 的基础上设计采样随机噪声的合理分布用于建模搜索方向的局部几何结构，并且进一步提出了几种降低搜索空间维度的技术。

算法4.1描述了进化攻击的整体流程。该算法在较低维的搜索空间 $\mathbb{R}^m$（$m<n$）中进行搜索，以提升优化效率。在每轮迭代中，算法依次执行以下几个步骤：1）从高斯分布 $\mathcal{N}(\mathbf{0},\sigma^2\mathbf{C})$ 中采样一个随机向量 $z\in\mathbb{R}^m$，其中 $\mathbf{C}$ 是一个对角协方差矩阵，用于建模搜索方向的局部几何结构；2）随机选取 z 中的 k 个坐标进行搜索，并将 z 中未被选取的坐标设置为 0；3）通过双线性插值将 z 映射到输入空间得到 $\tilde{z}\in\mathbb{R}^n$；4）向 $\tilde{z}$ 添加偏置减小对抗样本与原始样本之间的距离；5）测试子代解 $\tilde{x}^*+\tilde{z}$ 是否比父代解 $\tilde{x}^*$ 更优，并根据结果更新父代解与其他变量。下面将对进化攻击方法中一些具体步骤进行详细介绍。

4.4.1　初始化

算法4.1首先初始化对抗样本 $\tilde{x}^*$（见步骤 1）。在一般情况下，初始的 $\tilde{x}^*$ 需要满足对抗判别准则（即成功攻破黑盒模型）。如果初始的 $\tilde{x}^*$ 不满足对抗判别准则，那么攻击目标函数 $\mathcal{L}(\tilde{x}^*)$ 的取值为 $+\infty$。后续的迭代过程会向 $\tilde{x}^*$ 中添加随机噪声，但是深度神经网络通常对随机噪声具有一定的鲁棒性[19]，对抗判别准则仍然难以成立，损失函数的取值会保持为 $+\infty$。因此，初始化的对抗样本 $\tilde{x}^*$ 需要满足对抗判别准则，且在后续的迭代过程中 $\tilde{x}^*$ 也需要使对抗判别准则保持成立。对于躲避

攻击，初始的 $\tilde{x}^*$ 可以简单地设置为一个随机噪声；对于伪装攻击，初始的 $\tilde{x}^*$ 可以设置为目标身份的一个图片，如图4.4所示。可以看到，初始化的对抗样本 $\tilde{x}^*$ 与原始样本 x 距离很远，所以算法后续的迭代过程中会逐渐减小 $\tilde{x}^*$ 与 x 之间的距离。

4.4.2　高斯分布的均值

本小节解释算法4.1中向随机向量 $\tilde{z}$ 添加偏置的原因（见步骤 7）。为了叙述简洁，本小节假设搜索空间维度与输入空间维度相同，并且选择所有坐标进行搜索（即 $k = m = n$）。在此情况下，每轮迭代会从高斯分布中采样随机向量 z，且 $\tilde{z} = z$。在一般情况下，高斯分布应该是无偏的，其均值为零，以便在搜索空间中更加有效地进行探索。但是，对于本章研究的攻击目标函数 $\mathcal{L}(x^*)$，从均值为零的高斯分布中采样随机向量 z 会导致在 $n \to \infty$ 时对抗样本更新的概率趋近于 0。定理4.1描述了此结论。

定理 4.1： 若协方差矩阵 $\mathbf{C}$ 是正定矩阵，λ_{max} 和 $\lambda_{min}(> 0)$ 分别代表 $\mathbf{C}$ 的最大和最小特征值，那么：

$$P_{z\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{C})}\big(\mathcal{L}(\tilde{x}^* + z) < \mathcal{L}(\tilde{x}^*)\big) \leqslant \frac{4\lambda_{max}\|\tilde{x}^* - x\|^2}{\sigma^2\lambda_{min}^2 n^2}.$$

定理的详细证明参见[96]，本小节不再赘述。从定理4.1中可以看到，算法需要从均值为零的高斯分布中随机采样 $\mathcal{O}(n^2)$ 个向量才能产生一次成功的更新，这在 n 比较大时十分低效。产生这一现象的原因是：在高维搜索空间中，随机采样的向量 z 几乎与 $\tilde{x}^* - x$ 正交，因此 $\mathcal{D}(\tilde{x}^* + z, x)$ 很难比 $\mathcal{D}(\tilde{x}^*, x)$ 更小。为了解决此问题，随机向量 z 需要从有偏的高斯分布中采样，以减小对抗样本 $\tilde{x}^*$ 与原始样本 x 之间的距离。因此，算法在步骤 7 中向随机向量 $\tilde{z}$ 添加偏置 $\mu(x - \tilde{x}^*)$，其中 μ 是一个超参数，用于控制对抗样本向原始样本 x 移动的速度，第4.4.6节会详细介绍 μ 的选取和更新方式。

4.4.3　协方差矩阵自适应

协方差矩阵 $\mathbf{C}$ 可以建模搜索方向的局部几何结构，在优化过程中通过不断更新对搜索方向进行自适应。如果对优化变量所有维度间的相关关系进行建模，协方差矩阵的存储和计算复杂度至少为 $\mathcal{O}(m^2)$，这在 m 比较大时代价很高。对于黑盒决策攻击，搜索空间的维度通常比较高。例如，实验中设置 $m = 45 \times 45 \times 3$。因此，进化攻击方法采用对角协方差矩阵提升计算效率。受文献[123]启发，本小节设计了如下的规则更新对角协方差矩阵：

$$p_c = (1 - c_c)p_c + \sqrt{c_c(2 - c_c)}\frac{z}{\sigma}, \tag{4.3}$$

$$c_{ii} = (1 - c_{cov})c_{ii} + c_{cov}(p_c)_i^2, \tag{4.4}$$

其中 $p_c \in \mathbb{R}^m$ 被称为进化路径（evolution path），用于存储搜索过程中成功的方向信息。对于 $i = 1, 2, \cdots, m$，c_{ii} 代表协方差矩阵的第 i 个对角元素，$(p_c)_i$ 代表 p_c 的第 i 个元素。c_c 和 c_{cov} 是两个超参数。对公式（4.3）和公式（4.4）更新协方差矩阵的直观理解是其增加成功的搜索方向上的方差，以提升在未来搜索中对这些方向的采样频率。

4.4.4　随机坐标选取

对抗攻击可以修改输入图片中的一小部分像素生成对抗样本，欺骗深度学习模型[124]。如果能够找到关键的像素，黑盒决策攻击就可以只修改这些像素以降低搜索空间的维度，提升攻击效率。尽管在黑盒决策攻击场景下寻找关键的像素十分困难，进化攻击方法为找到关键搜索坐标提供了一种自然的方式（在搜索空间与输入空间维度相同的情况下，坐标即为像素）。在进化攻击方法中，对角协方差矩阵 $\mathbf{C}$ 中的元素代表成功的搜索方向，即较大的 c_{ii} 表明沿第 i 个坐标进行搜索可能会使对抗样本更新的概率更高。因此，在每轮迭代中，算法随机选取 k（$k \ll m$）个坐标生成随机向量 z，其中第 i 个坐标被选取的概率与 c_{ii} 成正比（见步骤 4-5）。

4.4.5　搜索空间降维

已有工作发现对搜索空间降维可以加速黑盒得分攻击[59]。基于类似的想法，进化攻击方法在较低维空间 $\mathbb{R}^m$（$m < n$）中采样随机向量 z（见算法中步骤 3）。算法此后采用双线性插值将 z 映射到原始空间 $\mathbb{R}^n$ 中（见步骤 6）。值得注意的是，进化攻击方法并不会改变输入图片的维度，而只会降低搜索空间的维度。

4.4.6　超参数设置

进化攻击方法中有几个比较关键的超参数，包括 σ，μ，c_c 和 c_{cov}。其中，c_c 被统一设置为 0.01，c_{cov} 被统一设置为 0.001，σ 被设置为 $0.01 \cdot D(\tilde{x}^*, x)$。$\mu$ 是一个需要被精细调节的关键超参数。如果 μ 太大，子代解 $\tilde{x}^* + \tilde{z}$ 很可能不满足对抗判别准则，导致对抗样本更新的概率较低。如果 μ 太小，尽管对抗样本更新的概率很高，但算法无法有效减小对抗样本 $\tilde{x}^*$ 与原始样本 x 之间的距离。进化攻击方法基于进化策略中超参数控制的传统方法[125]将 μ 更新规则设置为 $\mu = \mu \cdot \exp(P_{\text{update}} - \frac{1}{5})$，其中 P_{update} 是之前迭代过程中对抗样本更新的概率。

表 4.1　不同攻击方法在一千、一万和十万次查询下对人脸验证模型生成的对抗扰动大小

攻击目标	攻击方法	SphereFace			CosFace			ArcFace		
		一千	一万	十万	一千	一万	十万	一千	一万	十万
躲避攻击	Boundary	2.3e-2	7.0e-4	1.9e-5	2.0e-2	7.7e-4	1.6e-5	2.4e-2	1.5e-3	2.3e-5
	Optimization	1.2e-2	1.3e-3	7.1e-5	1.1e-2	1.3e-3	6.6e-5	1.5e-2	2.6e-3	9.9e-5
	NES-LO	1.4e-1	2.4e-2	7.4e-3	1.4e-1	2.0e-2	6.5e-3	1.4e-1	2.3e-2	1.5e-2
	Evolutionary	**1.6e-3**	**3.4e-5**	**1.3e-5**	**1.7e-3**	**3.3e-5**	**1.1e-5**	**2.8e-3**	**5.2e-5**	**1.6e-5**
伪装攻击	Boundary	1.5e-2	5.7e-4	1.6e-5	1.1e-2	2.8e-4	7.4e-6	2.0e-2	1.2e-3	1.7e-5
	Optimization	1.1e-2	1.3e-3	6.1e-5	7.7e-3	7.1e-4	2.8e-5	1.6e-2	3.3e-3	7.7e-5
	NES-LO	8.4e-2	1.7e-2	5.5e-3	9.3e-2	1.2e-2	3.1e-3	9.3e-2	1.9e-2	8.1e-3
	Evolutionary	**1.2e-3**	**2.9e-5**	**1.2e-5**	**6.5e-4**	**1.5e-5**	**5.3e-6**	**2.3e-3**	**3.9e-5**	**1.2e-5**

4.5　实验结果

本节通过实验验证进化攻击方法的有效性①。本节选取三个典型的人脸识别模型进行实验，包括 SphereFace[117]、CosFace[118]和 ArcFace[119]。在实验中，这些模型首先提取人脸图片的特征表示，然后计算不同图片的特征表示之间的余弦相似度，最后根据余弦相似度是否超过阈值或通过寻找余弦相似度最高的身份分别进行人脸验证和人脸鉴别。

本节选取 Labeled Face in the Wild（LFW）[115]数据集进行实验。对于人脸验证，实验选取 500 对人脸图片进行躲避攻击，其中每一对图片代表相同的身份；并选取另外 500 对人脸图片进行伪装攻击，其中每一对图片代表不同的身份。对于人脸鉴别，实验首先选取 500 个不同身份的 500 张图片组成原型图像集，然后相应地选取 500 张额外的图片作为原始样本进行躲避攻击和伪装攻击。对于伪装攻击，每一张图片的攻击目标身份通过随机选取得到。模型的输入图片大小（即输入空间维度 n）为 $112\times112\times3$。本节选取的所有图片均可以被三个人脸识别模型正确识别。尽管本节仅在 LFW 数据集上进行实验，所提出的进化攻击方法在人脸识别数据集 MegaFace[126]和自然图像数据集 ImageNet[87]上均取得了类似的结果，详细实验可参见[96]，本节不再赘述。

本节实验将进化攻击方法（记为 Evolutionary）与黑盒决策攻击基准方法进行比较，基准方法包括边界攻击（记为 Boundary）[63]、基于优化的攻击（记为 Optimization）[64]和自然进化策略在黑盒决策攻击场景下的扩展（记为 NES-LO）[60]。所有方法在迭代搜索的过程中均可以保证每一轮的对抗样本都满足对抗判别准则。

① 进化攻击方法代码参见https://github.com/thu-ml/ares/blob/main/ares/attack/evolutionary.py。

表 4.2　不同攻击方法在一千、一万和十万次查询下对人脸鉴别模型生成的对抗扰动大小

攻击目标	攻击方法	SphereFace			CosFace			ArcFace		
		一千	一万	十万	一千	一万	十万	一千	一万	十万
躲避攻击	Boundary	2.4e-2	4.7e-4	1.4e-5	2.0e-2	5.4e-4	1.2e-5	3.1e-2	1.6e-3	2.3e-5
	Optimization	1.1e-2	8.3e-4	4.6e-5	1.0e-2	8.2e-4	4.0e-5	2.0e-2	2.7e-3	9.8e-5
	NES-LO	1.4e-1	2.5e-2	5.5e-3	1.5e-1	2.2e-2	4.7e-3	1.5e-1	3.1e-2	1.3e-2
	Evolutionary	**1.3e-3**	**2.5e-5**	**9.9e-6**	**1.2e-3**	**2.3e-5**	**7.5e-6**	**3.2e-3**	**5.4e-5**	**1.6e-5**
伪装攻击	Boundary	2.4e-2	1.7e-3	3.6e-5	2.5e-2	1.3e-3	2.3e-5	2.5e-2	2.5e-3	3.8e-5
	Optimization	1.9e-2	3.7e-3	1.6e-4	1.9e-2	3.3e-3	1.1e-4	2.0e-2	6.0e-3	3.5e-4
	NES-LO	7.9e-2	2.8e-2	1.0e-2	8.8e-2	2.7e-2	8.8e-3	8.8e-2	2.3e-2	1.1e-2
	Evolutionary	**2.5e-3**	**6.3e-5**	**2.3e-5**	**2.2e-3**	**4.6e-5**	**1.5e-5**	**3.7e-3**	**8.8e-5**	**2.6e-5**

因此，实验通过均方误差（Mean Square Error，MSE）度量对抗样本与原始样本之间的距离（即对抗扰动大小），评估不同攻击方法的性能。实验将每张图片对黑盒模型的最大查询次数设置为十万。

4.5.1　实验结果对比

本小节展示在 LFW 数据集上进行黑盒决策攻击的实验结果。实验使用 Boundary、Optimization、NES-LO 和 Evolutionary 针对 SphereFace、CosFace 和 ArcFace 分别进行躲避攻击和伪装攻击。在 Evolutionary 方法中，搜索空间的维度设置为 $m = 45 \times 45 \times 3$，随机坐标选取的数量设置为 $k = \frac{m}{20}$。基准方法中的参数采用默认设置。对于人脸验证和人脸鉴别任务，图4.2和图4.3分别展示了不同攻击方法生成的对抗扰动大小随查询次数的变化曲线，其中扰动大小为共计 500 张图片生成的扰动的平均 MSE 值。此外，表4.1和表4.2分别展示了在这两个任务中不同攻击方法在一千、一万和十万次查询下生成的对抗扰动大小。图4.4展示了使用进化攻击方法针对 ArcFace 模型进行躲避攻击和伪装攻击的示例。

实验结果表明，相比于已有的黑盒决策攻击方法，本章提出的进化攻击在不同攻击场景（包括躲避攻击和伪装攻击）下对两个人脸识别任务（包括人脸验证和人脸鉴别）中所有的人脸识别模型均取得了更好的效果，即其收敛速度更快，可以在相同的模型查询次数下生成扰动更小的对抗样本。例如，如表4.1和表4.2所示，在给定一千次查询的情况下，进化攻击方法生成的对抗扰动比其他方法生成的对抗扰动小 10 倍左右，验证了所提出方法的有效性。从图4.4中可以看到，对模型进行两千次查询足以生成扰动很小的对抗样本，在视觉上难以分辨对抗样本与原始样本之间的区别。NES-LO 效率较低的原因是：其首先使用模型的预测类别估计其

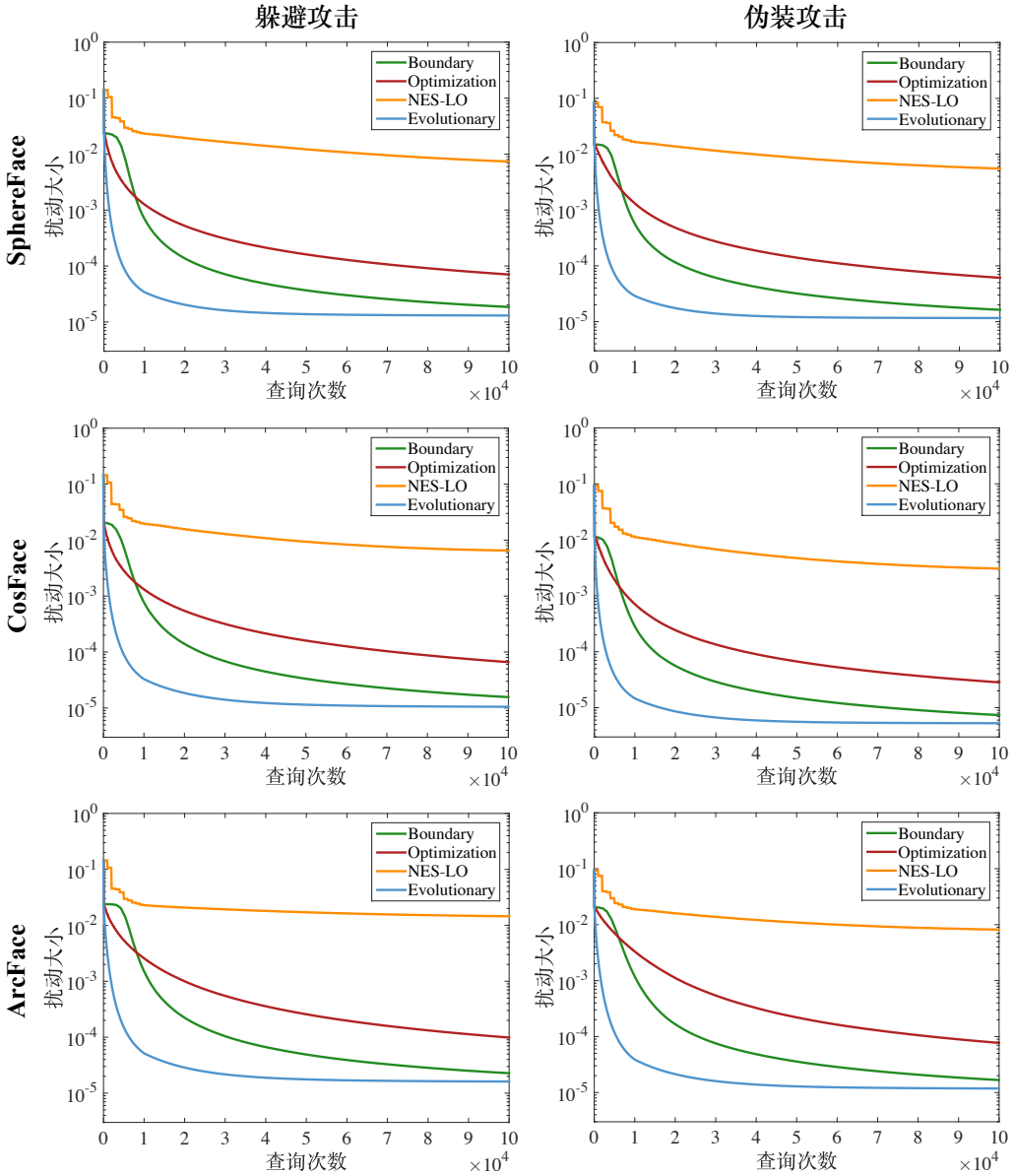


图 4.2　不同攻击方法对人脸验证模型生成的对抗扰动大小随查询次数变化图

预测概率分布，然后通过 NES 近似攻击目标函数的梯度，该方法更新一次对抗样本需要超过 1000 次查询，导致其攻击效率很低。

此外，本小节的实验还说明了人脸识别模型在黑盒决策攻击下的脆弱性。在 MSE 指标下通过加入大约 $1e^{-5}$ 量级的扰动就足以欺骗黑盒人脸识别模型，而这些扰动在视觉上很难察觉出来，如图4.4所示。人脸识别模型在对抗攻击下的脆弱性很可能对其实际应用带来严重的安全威胁。

4.5.2　消融实验

本小节进一步通过消融实验验证进化攻击方法中各个组成部分的效果。实验分别验证协方差矩阵自适应、随机坐标选取和搜索空间降维的效果。

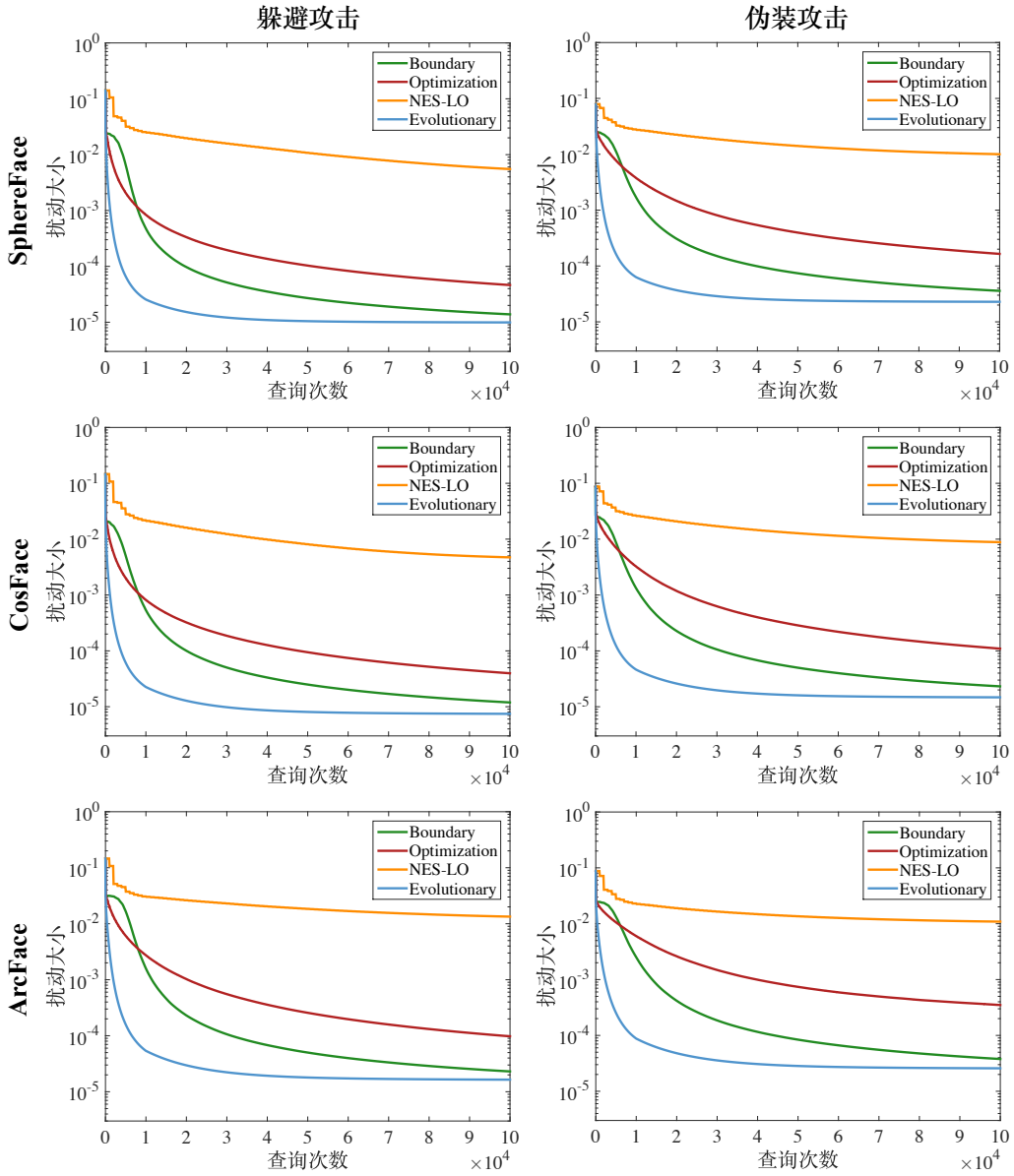


图 4.3　不同攻击方法对人脸鉴别模型生成的对抗扰动大小随查询次数变化图

首先，本小节验证协方差矩阵自适应（Covariance Matrix Adaptation，CMA）的有效性。实验选取将协方差矩阵设置为 $\mathbf{I}_n$ 而不对其进行更新的方法作为对比方法，并与 CMA 进行比较。为了仅验证 CMA 的效果，此实验不使用随机坐标选取和搜索空间降维的技术。表4.3中的第 2-3 和第 6-7 行展示了给定一万次查询下生成的对抗扰动的大小。可以看到，CMA 可以生成扰动更小的对抗样本，验证了其在搜索过程中对协方差矩阵 $\mathbf{C}$ 进行更新的优势。

其次，本小节验证随机坐标选取（Stochastic Coordinate Selection，SCS）的有效性，并验证是否应该将坐标选取的概率设置为与协方差矩阵 $\mathbf{C}$ 中的对角线元素成正比。对比方法将所有坐标被选取的概率设为相同，即坐标选取的概率与单位矩阵 $\mathbf{I}_n$ 中的对角元素成正比。表4.3展示了不同方法的实验结果。可以看到，随机

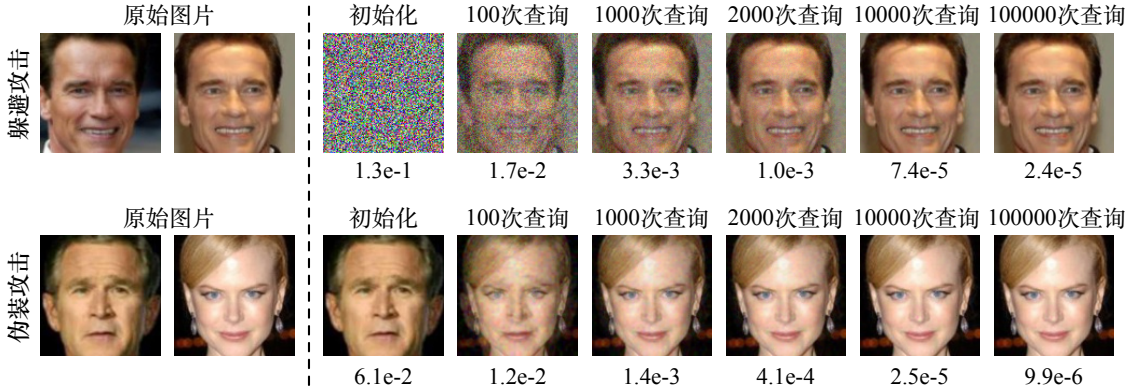


图 4.4　使用进化攻击方法进行躲避攻击和伪装攻击示意图

表 4.3　进化攻击方法中协方差矩阵自适应和随机坐标选取的消融实验结果

攻击目标	实验设置	SphereFace	CosFace	ArcFace
躲避攻击	无 CMA，无 SCS	2.6e-4	2.5e-4	4.2e-4
	有 CMA，无 SCS	2.4e-4	2.3e-4	3.8e-4
	有 CMA，有 SCS（$\mathbf{C}$）	**1.7e-4**	**1.6e-4**	**2.6e-4**
	有 CMA，有 SCS（$\mathbf{I}_n$）	2.0e-4	1.9e-4	3.0e-4
伪装攻击	无 CMA，无 SCS	1.9e-4	9.2e-5	2.6e-4
	有 CMA，无 SCS	1.8e-4	8.5e-5	2.5e-4
	有 CMA，有 SCS（$\mathbf{C}$）	**1.3e-4**	**6.4e-5**	**1.7e-4**
	有 CMA，有 SCS（$\mathbf{I}_n$）	1.5e-4	7.5e-5	2.0e-4

坐标选取有助于取得更好的结果，并且坐标选取的概率与协方差矩阵 $\mathbf{C}$ 中的对角线元素成正比要比与 $\mathbf{I}_n$ 中的对角元素成正比结果更好。

最后，本小节验证搜索空间降维的有效性。实验将搜索空间的维度 m 分别设置为 $15\times15\times3$、$30\times30\times3$、$45\times45\times3$、$60\times60\times3$ 和 $112\times112\times3$。在人脸验证任务中，实验使用进化攻击方法在不同的搜索空间维度下针对 SphereFace、CosFace 和 ArcFace 进行躲避攻击和伪装攻击。图4.5展示了扰动大小随查询次数的变化曲线。可以看到，进化攻击方法在搜索空间维度较低的情况下收敛速度更快。但是，如果搜索空间维度太低（例如，$m = 15\times15\times3$），会导致最终的扰动较大。因此，进化攻击方法需要选取适当的搜索空间维度（例如，$m = 45\times45\times3$），以平衡搜索效率和最终生成的对抗扰动的大小。

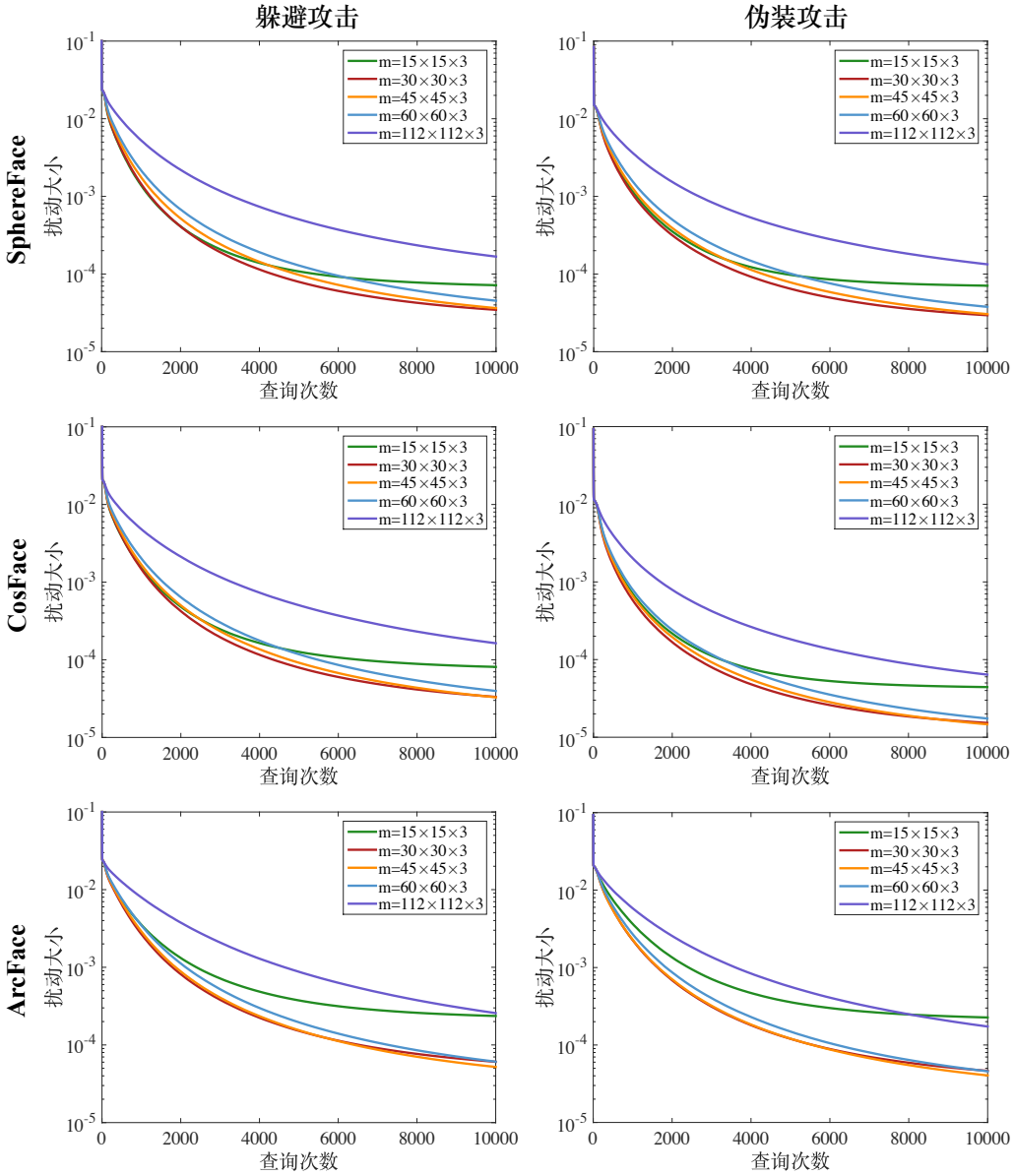


图 4.5　对人脸识别模型的黑盒决策攻击示例

4.5.3　攻击商用人脸识别系统

本小节针对腾讯人工智能开放平台中的人脸验证 API[①]进行黑盒决策攻击。该人脸验证 API 允许用户上传两张人脸图片，并输出其相似度分数。本小节将人脸相似度阈值设置为 90，即人脸相似度分数大于 90 时模型将其识别为同一身份，否则识别为不同身份。在 LFW 数据集中选取 10 对图片进行伪装攻击，其中每对原始图片代表不同身份。攻击方法对每一对图片中的一张图片生成对抗扰动，目标是使人脸验证 API 将这一对图片识别为同一身份。对人脸验证 API 的最大查询次数设置为一万次。实验使用本章提出的进化攻击方法 Evolutionary 攻击腾讯人脸验

① https://ai.qq.com/#compare。

表 4.4　对腾讯人脸验证 API 的攻击结果

攻击算法	扰动大小（MSE）
Boundary	1.63e-2
Optimization	1.71e-2
Evolutionary	**2.54e-3**

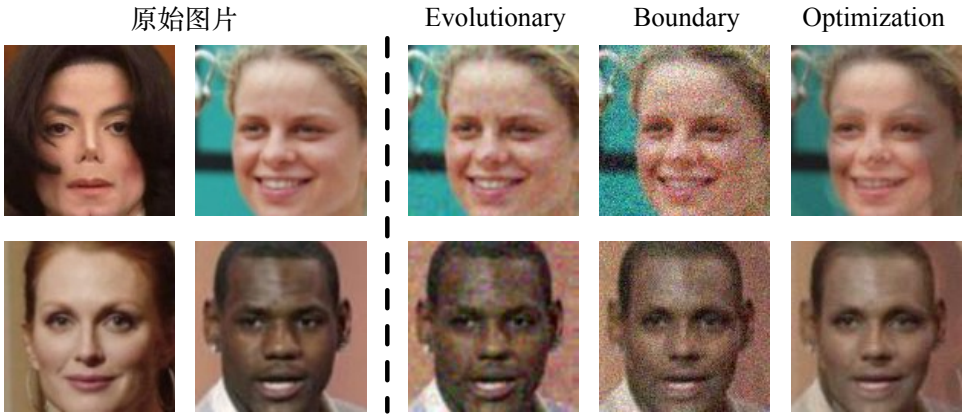


图 4.6　对腾讯人脸验证 API 进行攻击的对抗样本示例

证 API，并与 Boundary 和 Optimization 方法进行对比。表4.4展示了不同方法在一万次查询后生成的对抗扰动的平均大小。进化攻击相比于基准方法得到了更小的对抗扰动，说明其有效性。

图4.6展示了不同方法生成的对抗样本示例。可以看到，进化攻击生成的对抗样本与原始样本更加相似，而其他方法生成的对抗样本明显改变了原始图像的语义，很容易被人眼区分出来。本小节的实验结果展示出进化攻击方法可以攻破商用人脸识别系统，对真实世界中安全性要求较高的人脸识别应用（例如刷脸支付）带来了严重的威胁。

4.6　本章小结

本章面向人脸识别任务提出了进化攻击方法。在黑盒决策攻击场景下，该方法不需要获取模型的结构和参数信息，可以通过黑盒优化的方式仅利用对黑盒模型的查询结果生成对抗样本。进化攻击方法通过对搜索方向的局部几何结构进行建模，并降低搜索空间的维度，以提升黑盒决策攻击的查询效率。本章应用进化攻击方法研究典型的人脸识别模型的鲁棒性。实验结果验证了进化攻击方法的有效性，其相比于已有方法可以通过更少的模型查询次数生成扰动更小的对抗样本。实验进一步应用进化攻击方法攻破了商用人脸识别系统，验证了所提出方法的实

用性。本章的结果表明现有人脸识别模型极易被黑盒决策攻击攻破，这很可能会对真实世界中的人脸识别系统带来严重的安全威胁。

为了解决人脸识别模型在对抗攻击下的脆弱性问题，保障人脸识别系统在一些高风险任务上的应用，需要构建有效的对抗防御算法。一些典型的防御技术可以被扩展到人脸识别任务中，例如可以通过对抗训练或第5章提出的对抗分布训练增强人脸识别模型的鲁棒性，但是这些防御方法会导致模型的识别精度降低，影响人脸识别系统的实际使用。因此，如何在不影响模型精度的情况下提升人脸识别的鲁棒性仍然是亟待解决的重要研究问题。

第 5 章　面向对抗样本多样化的对抗分布训练

对抗攻击除了可用于发现深度学习模型的脆弱性以及比较不同模型的鲁棒性，还可以生成额外的训练数据增强模型的鲁棒性。对抗训练采用对抗攻击生成的对抗样本作为训练数据，是目前最有效的防御方式之一。然而，大多数对抗训练方法会采用特定的攻击生成对抗样本，导致训练得到的模型在测试时对未知的攻击无法有效防御。一些对抗攻击方法对输入空间中不同对抗扰动的探索也不够充分，导致模型的鲁棒性不足。本章提出对抗分布训练，其建模每一个输入样本邻域内的对抗（样本）分布，以生成更加多样化的对抗样本，增强模型在不同攻击下的泛化能力以及在测试数据上的鲁棒性。对抗分布训练可以被描述为一个最小最大优化问题，其中内层最大化旨在学习对抗分布，通过加入熵正则项更好地表征原始样本周围多样的对抗样本；而外层最小化旨在优化模型在对抗分布下的期望损失，以训练更加鲁棒的分类器。本章通过理论分析推导出求解对抗分布训练的通用算法，并提出三种对抗攻击方式建模对抗分布，包括显式分布建模、均摊显式分布建模和均摊隐式分布建模。实验结果表明对抗分布训练相比于多个对抗训练方法可以增强模型的鲁棒性。此外，本章进一步说明对抗训练可以提升深度学习模型的可解释性。

5.1　本章引言

为了解决对抗攻击所带来的安全威胁，大量研究工作致力于增强深度学习模型的鲁棒性。对抗训练（Adversarial Training，AT）[20,37,69]是目前最有效的防御方式之一[36,99]。对抗训练可以被描述为一个最小最大优化问题（minimax optimization problem）[37]，其中内层最大化通过最大化模型的损失函数生成对抗样本；外层最小化利用生成的对抗样本训练鲁棒的分类器。由于内层最大化问题是非凹的（non-concave）且难以在可承受的时间内精确求解（intractable），对抗训练通常采用对抗攻击近似求解内层最大化问题，例如快速梯度符号法（FGSM）[20]和投影梯度下降法（PGD）[37]。因此，对抗攻击的一个重要作用是在对抗训练中提供额外的训练数据增强模型的鲁棒性。

然而，已有对抗训练方法通常选取特定的攻击求解内层最大化问题，一些方法导致训练得到的模型对测试阶段未知的攻击无法有效防御，即在不同攻击下的泛化能力较差[127]。例如，通过快速梯度符号法进行对抗训练[20]，在不使用随机初始化和早停（early stopping）等技术[128]的情况下，训练得到的模型很容易被基

于多步梯度迭代的攻击方法攻破[69,104]。尽管最近的对抗训练方法[70,129-130]对常见的基于多步梯度迭代的攻击方法（如投影梯度下降法）展现出优越的鲁棒性，但是仍然可以被更强大的攻击或适应性攻击攻破[42,131]。

采用单一的对抗攻击方法也难以充分探索输入空间中多样的对抗扰动，导致对抗训练模型的鲁棒性存在不足。基于投影梯度下降法的对抗训练[37]试图使用随机初始化解决此问题，但是投影梯度下降法在不同随机初始化下生成的对抗样本仍然十分聚集，缺乏多样性[132]。还有一些方法采用多种攻击生成的对抗样本进行训练[104,133]，可以更好地近似求解内层最大化问题[37]。尽管如此，目前对抗训练中仍然缺乏对多样化对抗样本进行建模的合理方式。

为解决上述问题，本章提出*对抗分布训练*（Adversarial Distributional Training，ADT），用于增强深度学习模型在不同攻击下的泛化能力及鲁棒性。相比于对抗训练，对抗分布训练提供了一种新颖的学习框架，其通过对抗分布建模每一个原始样本邻域内的对抗样本。对抗分布训练同样可以被描述为一个最小最大优化问题，其中内层最大化旨在通过最大化模型的期望损失学习每个原始样本的对抗分布；外层最小化旨在通过最小化模型在对抗分布下的期望损失训练鲁棒的分类器。为了防止对抗分布退化为 Delta 分布，即对抗分布训练退化为对抗训练，训练的损失函数中加入了熵正则项，使学习到的对抗分布可以更好地表征多样的对抗样本。

通过理论分析，本章推导出求解对抗分布训练中最小最大优化问题的通用算法，其与对抗训练的求解方式类似，即首先求解内层最大化问题，进而通过得到的内层最优解求解外层最小化问题。本章进一步提出三种不同的对抗攻击方式参数化建模对抗分布，实现对抗分布训练，其中包括*显式分布建模*、*均摊显式分布建模*和*均摊隐式分布建模*。本章在广泛使用的 CIFAR-10[134]、CIFAR-100[134]和 SVHN[135] 数据集上进行大量实验，第5.5节中的实验结果表明对抗分布训练可以有效增强模型的鲁棒性，其取得了比多个对抗训练方法更加优异的性能。

本章进一步研究对抗训练方法对深度学习模型可解释性的影响。实验发现正常训练的模型可解释性存在不足，其内部神经元（neuron）学习到的特征与人类所理解的语义概念（semantic concept）不一致。本章面向深度学习的可解释性提出*加入特征表示一致性损失的对抗训练*方法。第5.6节中的实验结果表明该方法可以有效提升深度学习模型的可解释性，使其内部神经元学习到的特征与人类所理解的语义概念更加一致。

5.2　背景知识

本章将包含 n 个训练样本的数据集记为 $\boldsymbol{D}=\{x_i,y_i\}_{i=1}^n$，其中 $x_i\in\mathbb{R}^d$ 代表输入数据，$y_i\in\{1,2,\cdots,L\}$ 代表 x_i 的真实类别，d 为输入空间维度，L 为总共的类别数量。对抗训练可以被描述为一个最小最大优化问题[37]，如下所示：

$$\min_{\theta}\frac{1}{n}\sum_{i=1}^{n}\max_{\delta_i\in S}\boldsymbol{J}(f_\theta(x_i+\delta_i),y_i), \tag{5.1}$$

其中 f_θ 代表参数为 θ 的深度神经网络，输出为所有类别上的预测概率分布。$\boldsymbol{J}(\cdot,\cdot)$ 代表损失函数，通常选取为交叉熵损失，如公式（2.2）所示。$\boldsymbol{S}=\{\delta:\|\delta\|_\infty\leqslant\epsilon\}$ 代表对抗扰动的范围，其中 ϵ 为最大扰动规模。本章所介绍的方法使用 ℓ_∞ 范数，这也被已有的对抗训练工作广泛采用。

对抗训练通常会依次求解优化问题（5.1）中的内层最大化问题和外层最小化问题，即首先求解内层最大化问题生成对抗样本，然后利用生成的对抗样本求解外层最小化问题优化模型参数。一些典型的对抗攻击方法可近似求解内层最大化问题，包括快速梯度符号法（FGSM）[20]和投影梯度下降法（PGD）[37]。投影梯度下降法生成对抗样本的过程可以被描述为：

$$\delta_i^{t+1}=\boldsymbol{\Pi}_S\left(\delta_i^t+\alpha\cdot\operatorname{sign}(\nabla_x\boldsymbol{J}(f_\theta(x_i+\delta_i^t),y_i))\right), \tag{5.2}$$

其中 δ_i^t 代表第 t 轮迭代生成的对抗扰动，$\boldsymbol{\Pi}(\cdot)$ 代表投影函数，α 代表迭代步长。投影梯度下降法将初始化的对抗扰动 δ_i^0 设置为 $\boldsymbol{S}$ 中均匀采样的随机噪声，这也是其与基础迭代法[69]唯一的区别。

尽管对抗训练取得了不错的鲁棒性性能，但是已有方法也存在一些问题。首先，对抗训练采用特定的攻击方法生成对抗样本进行训练，导致模型对未知的攻击防御能力不足。其次，单一的攻击方式也难以有效探索输入空间中多样的对抗扰动，导致模型在测试数据上的鲁棒性不及预期。

5.3　对抗分布训练

为解决对抗训练中存在的上述问题，本章提出对抗分布训练（Adversarial Distributional Training，ADT）框架，其通过建模对抗样本的分布（即对抗分布）训练更加鲁棒的分类器。具体而言，对抗分布训练通过概率分布 $p(\delta_i)$ 对每个原始数据 x_i 周围的对抗扰动进行建模，其中 $p(\delta_i)$ 的支撑集（support set）包含在 $\boldsymbol{S}$ 中，保证从 $p(\delta_i)$ 中采样的对抗扰动不会超越 ℓ_∞ 范数的限制。对抗分布训练可以被描述为

基于概率分布的最小最大优化问题，如下所示：

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \max_{p(\delta_i) \in P} \mathbb{E}_{p(\delta_i)} \left[J(f_\theta(x_i + \delta_i), y_i) \right], \tag{5.3}$$

其中 $P = \{p : \text{supp}(p) \subseteq S\}$ 代表所有支撑集包含在 S 中的分布族。从公式（5.3）中可以看到，对抗分布训练的内层最大化问题旨在学习对抗分布 $p(\delta_i)$ 以最大化模型在 $p(\delta_i)$ 下的期望损失；外层最小化问题的目标是通过最小化模型在对抗分布下的期望损失训练模型的参数。值得注意的是，当指定的分布族 P 仅包含 Delta 分布时，即对抗分布仅在某个点存在概率密度，对抗分布训练退化为对抗训练。因此，对抗训练也可以被认为是对抗分布训练的一个特例。

5.3.1　正则化对抗分布

对于对抗分布训练的内层最大化问题，可以看到其具有以下的性质：

$$\max_{p(\delta_i) \in P} \mathbb{E}_{p(\delta_i)} \left[J(f_\theta(x_i + \delta_i), y_i) \right] \leqslant \max_{\delta_i \in S} J(f_\theta(x_i + \delta_i), y_i). \tag{5.4}$$

这意味着对抗分布训练的内层最优解会退化为 Delta 分布。在此情况下，对抗分布无法建模多样的对抗扰动，对抗分布训练也退化为普通的对抗训练。为解决此问题，对抗分布训练在损失函数中加入熵正则项（entropic regularization），可以被表示为：

$$\begin{aligned} &\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \max_{p(\delta_i) \in P} F(p(\delta_i), \theta), \\ &F(p(\delta_i), \theta) = \mathbb{E}_{p(\delta_i)} \left[J(f_\theta(x_i + \delta_i), y_i) \right] + \lambda \cdot H(p(\delta_i)), \end{aligned} \tag{5.5}$$

其中 $H(p(\delta_i)) = -\mathbb{E}_{p(\delta_i)}[\log p(\delta_i)]$ 代表对抗分布 $p(\delta_i)$ 的熵（entropy），λ 代表控制熵正则项的超参数。为了方便描述，公式（5.5）中使用 $F(p(\delta_i), \theta)$ 代表总体的训练损失函数。

5.3.2　对抗分布训练的优势分析

从公式（5.1）和公式（5.5）中可以看到，对抗训练与对抗分布训练之间的主要区别在于：对每个输入数据 x_i，对抗训练寻找最大化模型损失的对抗样本，而对抗分布训练学习最大化模型期望损失的对抗分布。对抗分布具有覆盖不同攻击方法生成的对抗扰动的能力。通过最小化模型在对抗分布下的期望损失，训练得到的分类器在不同攻击下具有更好的鲁棒性。

由于对抗分布训练在损失函数中加入熵正则项，其可以更好地探索并表征输入空间中可能存在的多样化的对抗扰动。本小节通过实验验证这一结论。实验选

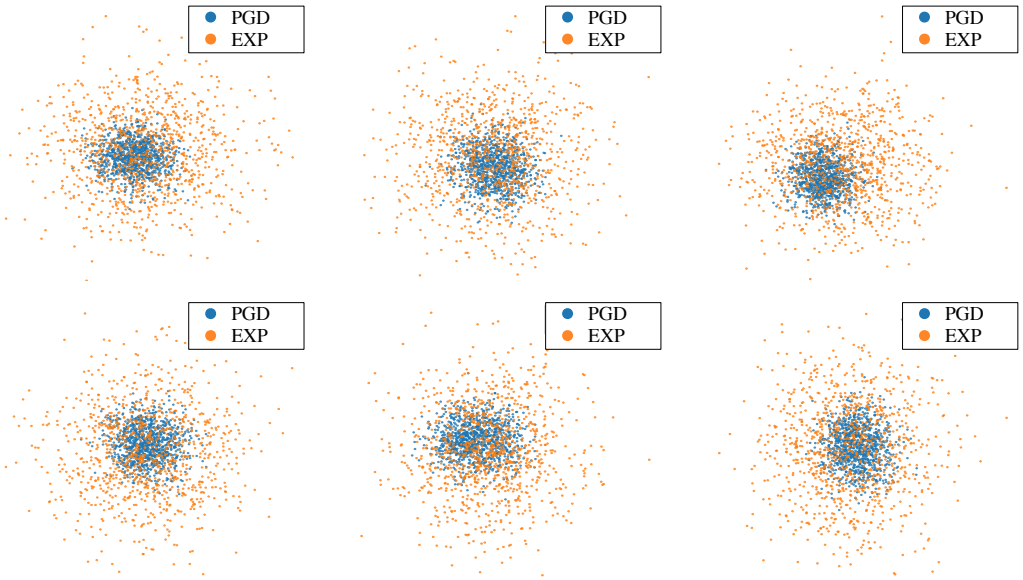


图 5.1　投影梯度下降法生成的对抗样本与对抗分布中采样的对抗样本可视化

取正常训练的模型，对于每个数据通过随机初始化的投影梯度下降法（PGD）生成一组不同的对抗样本，并从第5.4.1节介绍的显式对抗分布（EXP）中采样另一组对抗样本。此后，实验通过主成分分析（Principal Component Analysis，PCA）[136]将这两组对抗样本投影到二维子空间上进行可视化。图5.1展示了一些数据的可视化结果。从图中可以看到，对抗分布中采样的对抗样本更加分散，而由投影梯度下降法生成的对抗样本十分聚集，缺乏多样性。实验进一步定量测量每组对抗样本两两之间距离的平均值，用于估计对抗样本的多样性。对于 100 个随机选取的数据，对抗分布中采样的对抗样本的平均 ℓ_2 距离为 1.95，而投影梯度下降法生成的对抗样本的平均 ℓ_2 距离为 1.56。虽然对抗分布表征了更加多样化的对抗样本，但具有与投影梯度下降法相似的攻击能力，如表5.4所示。因此，通过最小化模型在对抗分布下的期望损失有助于学习更加平滑的损失平面，如图5.5所示。与对抗训练相比，对抗分布训练可以增强模型的鲁棒性，第5.5节中的实验结果也将验证这一结论。

5.3.3　对抗分布训练的通用算法

为了求解最小最大优化问题，Danskin 定理[137] 指出如何使用内层最大化问题的最优解得到外层最小化问题的梯度，同时这也是对抗训练的理论基础[37]。然而，对抗分布训练的最小最大优化问题无法直接应用 Danskin 定理进行求解，这是因为搜索空间 P 不一定是紧的（compact），此要求为 Danskin 定理的一个必要假设。这导致从理论上分析对抗分布训练的求解方式非常困难。因此，本小节首先提出以下假设。

算法 5.1 对抗分布训练的通用算法

输入：训练数据集 $\mathcal{D}$，损失函数 $F(p(\delta_i),\theta)$，分布族 P，训练轮数 N，学习率 η；

初始化模型参数 θ；
对 $j=1,\cdots,N$ **执行**
　采样小批量数据 $\mathcal{B}\subset\mathcal{D}$；
　对每个数据 $(x_i,y_i)\in\mathcal{B}$，求解 $p^*(\delta_i)=\arg\max_{p(\delta_i)\in P}F(p(\delta_i),\theta)$ 得到对抗分布 $p^*(\delta_i)$；
　通过随机梯度下降更新模型参数 θ：$\theta\leftarrow\theta-\eta\cdot\mathbb{E}_{(x_i,y_i)\in\mathcal{B}}\left[\nabla_\theta F(p^*(\delta_i),\theta)\right]$。

假设 5.1： 对抗分布训练的损失函数 $F(p(\delta_i),\theta)$ 关于模型参数 θ 是连续可微的。

文献[37]也对对抗训练进行了同样的假设。由于深度神经网络中存在修正线性单元（ReLU）层，损失函数不是完全连续可微的，但是其不连续的点集度量为零，在实际中可以假设损失函数具有连续可微的性质。

假设 5.2： 分布族 P 中分布的概率密度函数是有界的（bounded）且等度连续的（equicontinuous）。

假设5.2对分布族 P 进行了限制。第5.4.1节介绍的显式对抗分布满足此假设。基于假设5.1和假设5.2，本小节给出求解对抗分布训练中最小最大优化问题的理论依据。

定理 5.1： 在假设5.1和假设5.2成立的情况下，定义 $\rho(\theta)=\max_{p(\delta_i)\in P}F(p(\delta_i),\theta)$，$P^*(\theta)=\{p(\delta_i)\in P:F(p(\delta_i),\theta)=\rho(\theta)\}$，则 $\rho(\theta)$ 沿向量 v 的方向导数满足：

$$\rho'(\theta;v)=\sup_{p(\delta_i)\in P^*(\theta)}v^\top\nabla_\theta F(p(\delta_i),\theta). \tag{5.6}$$

特别地，当 $P^*(\theta)=\{p^*(\delta_i)\}$ 仅包含一个最优解时，$\rho(\theta)$ 在 θ 处可微且满足：

$$\nabla_\theta\rho(\theta)=\nabla_\theta F(p^*(\delta_i),\theta). \tag{5.7}$$

定理的详细证明参见[97]，本小节不再赘述。定理5.1提供了求解对抗分布训练的通用算法，即首先求解内层最大化问题，然后使用内层最大化问题的全局最优解计算损失函数的梯度更新模型参数。算法5.1描述了对抗分布训练的通用算法。与对抗训练类似，对抗分布训练中内层最大化问题的全局最优解难以精确计算，因此本章提出三种不同的攻击方法近似求解内层最大化问题。实验结果表明通过设计合理的对抗分布形式及攻击方法可以有效求解对抗分布训练中的最小最大优化问题，增强模型的鲁棒性。

5.3.4　相关工作

本章所提出的对抗分布训练的核心思想是建模对抗样本的分布，与之类似的想法在文献[60-61]中被讨论，但是其研究目标是黑盒得分攻击。文献[60-61]在高斯分布下搜索对抗样本，其形式与公式（5.3）中的内层最大化问题类似。但是这些方

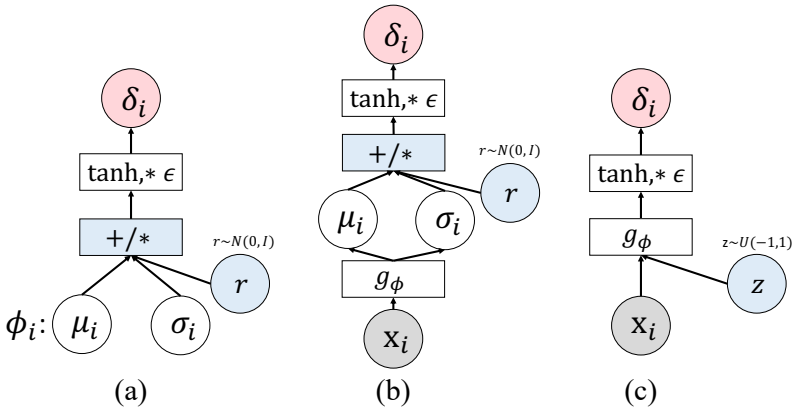


图 5.2　参数化对抗分布的三种方式

法采用自然进化策略（NES）[62]估计模型的梯度以优化分布的参数，呈现出很高的方差[138]。本章介绍的攻击方法基于损失函数的梯度优化对抗分布的参数，可以减小方差。

模型的对抗鲁棒性与其在某些噪声分布下的鲁棒性直接相关[139]。例如，对高斯噪声鲁棒的分类器可以被转换为一个平滑分类器，其对 ℓ_2 范数下的对抗样本具有可证实鲁棒性（certified robustness）[85]。文献[140]进一步采用对抗训练增强随机平滑（randomized smoothing）方法[85]的可证实鲁棒性。与之不同，本章所提出的方法属于以实验为依据的防御（empirical defense），其目标是学习依赖于输入数据的对抗分布以训练鲁棒的分类器。

本章所提出的对抗分布训练框架与分布式鲁棒优化（Distributionally Robust Optimization，DRO）[82,141-142]有一定相似性，但其存在本质区别。分布式鲁棒优化通过在整体数据分布的变换下进行训练，得到对数据分布改变具有鲁棒性的模型。通过 Wasserstein 距离定义数据分布的变化，分布式鲁棒优化会与对抗训练十分类似[82,143]。但是，对抗分布训练不会对数据分布的变化进行建模，而是对每个输入数据周围的对抗样本的分布进行建模。

5.4　参数化对抗分布

从第5.3.3节中的讨论可以看到，对抗分布训练的关键是求解优化问题（5.5）中的内层最大化问题。本节提出三种不同的求解方式，其基本思想均为参数化（parameterize）对抗分布。通过将对抗分布 $p_{\phi_i}(\delta_i)$ 表示为依赖于可训练参数 ϕ_i 的形式，对抗分布训练中的内层最大化问题变为对于 ϕ_i 最大化模型的期望损失。本节将分别介绍这三种不同的参数化建模方式及其学习过程。图5.2展示了显式分布建模、均摊显式分布建模和均摊隐式分布建模的示意图。

5.4.1　显式分布建模

对输入数据周围的对抗扰动进行参数化建模的一种非常自然的方式是使用具有概率密度函数的显式分布，被称为*显式对抗分布*（EXPlicit adversarial distribution，EXP）。为了使对抗分布的支撑集包含在扰动范围 S 内，可以采用随机变量变换的方式定义显式对抗分布，如下所示：

$$\delta_i = \epsilon \cdot \tanh(u_i), \quad u_i \sim \mathcal{N}(\mu_i, \mathrm{diag}(\sigma_i^2)), \tag{5.8}$$

其中 u_i 从对角高斯分布（diagonal Gaussian distribution）中采样得到，μ_i 和 σ_i 分别为对角高斯分布的均值和标准差。对抗扰动 δ_i 通过 tanh 函数将 u_i 进行变换并乘以扰动规模 ϵ 得到。本小节使用 $\phi_i = (\mu_i, \sigma_i)$ 代表显式对抗分布中需要学习的参数。

通过公式（5.8）中定义的显式对抗分布，公式（5.5）中内层最大化问题可以被描述为：

$$\max_{\phi_i} \left\{ \mathbb{E}_{p_{\phi_i}(\delta_i)} \left[J(f_\theta(x_i + \delta_i), y_i) \right] + \lambda \cdot H(p_{\phi_i}(\delta_i)) \right\}. \tag{5.9}$$

为了求解此问题，优化的过程需要计算期望损失对于分布参数 ϕ_i 的梯度。一种常用的方法是重参数化技巧（reparameterization trick）[138,144]，其利用与随机变量相关的可微变换代替采样过程。通过使用此技术，损失函数的梯度可以直接通过样本反向传播到分布的参数上，更加高效地优化损失函数。对于公式（5.9）中的损失函数，该技巧可以将 δ_i 重参数化为 $\delta_i = \epsilon \cdot \tanh(u_i) = \epsilon \cdot \tanh(\mu_i + \sigma_i r)$，其中 r 服从标准高斯分布 $\mathcal{N}(0, I)$。因此，公式（5.9）中损失函数的梯度可以被估计为：

$$\mathbb{E}_{r\sim\mathcal{N}(0,I)} \nabla_{\phi_i} \left[J\left(f_\theta(x_i + \epsilon \cdot \tanh(\mu_i + \sigma_i r)), y_i \right) - \lambda \cdot \log p_{\phi_i}(\epsilon \cdot \tanh(\mu_i + \sigma_i r)) \right]. \tag{5.10}$$

其中第一项为分类器在随机采样噪声下的损失，第二项为显式对抗分布的熵，其可以被解析地计算，如下所示：

$$\begin{aligned} &- \log p_{\phi_i}(\epsilon \cdot \tanh(\mu_i + \sigma_i r)) \\ &= \sum_{j=1}^{d} \left(\frac{1}{2}(r^{(j)})^2 + \frac{\log 2\pi}{2} + \log \sigma_i^{(j)} + \log\left(1 - \tanh(\mu_i^{(j)} + \sigma_i^{(j)} r^{(j)})^2\right) + \log \epsilon \right), \end{aligned} \tag{5.11}$$

其中上标 j 代表向量的第 j 维取值。

在实际中，公式（5.10）中的期望可以通过 k 次蒙特卡洛（Monte Carlo，MC）采样近似。优化过程使用 T 轮梯度上升优化显式对抗分布的参数 ϕ_i。在优化结束后可以得到分布的参数 ϕ_i^*，对抗分布训练采用显式对抗分布 $p_{\phi_i^*}(\delta_i)$ 更新模型的参数 θ。

5.4.2　均摊显式分布建模

尽管显式对抗分布是实现对抗分布训练的一种简单方式，但是需要为每个输入数据学习对抗分布的参数，这会带来较高的计算代价。相比于使用 T 轮迭代的投影梯度下降法进行对抗训练[37]，基于显式分布建模的对抗分布训练会慢 k 倍左右，这是因为每轮迭代中 ϕ_i 的梯度由 k 次蒙特卡洛采样进行估计。为了更加高效地学习对抗分布，本小节通过均摊的方式学习显式对抗分布，被称为均摊显式对抗分布（AMortized EXPlicit adversarial distribution，EXP-AM）。

具体而言，均摊显式分布建模方式不会为每一个输入数据优化对抗分布的参数，而是学习映射 $g_\phi : \mathbb{R}^d \to \boldsymbol{P}$ 以条件概率分布 $p_\phi(\delta_i|x_i)$ 的方式建模每个样本的对抗分布。映射函数 g_ϕ 选取为条件生成网络（conditional generative network），其以原始样本 x_i 作为输入，并输出 x_i 对应的显式对抗分布的参数 (μ_i, σ_i)，该分布的具体形式同样由公式（5.8）定义。使用生成模型定义 $p_\phi(\delta_i|x_i)$ 的优势在于：生成网络可以潜在地学习对抗扰动中的通用结构，使其可以泛化到具有相似特征的其他训练样本上[145-146]。这就意味着该方法可以将优化生成网络参数 ϕ 的代价均摊到不同样本上，加速训练过程。

通过利用均摊显式对抗分布，对抗分布训练中最小最大优化问题（5.5）可以被描述为：

$$\min_\theta \max_\phi \frac{1}{n}\sum_{i=1}^{n}\left\{\mathbb{E}_{p_\phi(\delta_i|x_i)}\left[J(f_\theta(x_i+\delta_i), y_i)\right] + \lambda \cdot H(p_\phi(\delta_i|x_i))\right\}, \tag{5.12}$$

其中 θ 和 ϕ 分别代表分类器和生成网络的参数。对抗分布训练的训练过程同时对 θ 和 ϕ 进行随机梯度下降与上升，其中损失函数对于 ϕ 的梯度同样可以通过重参数化技巧得到。

5.4.3　均摊隐式分布建模

由于对抗扰动的真实分布尚不清楚，并且不同数据周围的对抗分布可能存在差异，很难找到适当的显式分布建模对抗扰动，采用第5.4.1节中介绍的分布形式可能会造成欠拟合的问题。针对此问题，本小节提出利用隐式分布建模对抗扰动，被称为均摊隐式对抗分布（AMortized IMPlicit adversarial distribution，IMP-AM）。隐式分布的特点是无法被明确表示其概率密度函数，但是可以从中进行采样。最近的研究显示出隐式分布对于建模复杂高维数据的灵活性[147-148]。

具体而言，均摊隐式分布建模方式同样使用生成网络 $g_\phi : \mathbb{R}^{d_z} \times \mathbb{R}^d \to \mathbb{R}^d$ 建模对抗分布，通过 $\delta_i = g_\phi(z; x_i)$ 隐式地定义对抗扰动的条件分布 $p_\phi(\delta_i|x_i)$，其中 x_i 是原始样本，$z \in \mathbb{R}^{d_z}$ 是一个随机噪声向量。在此方法中，z 从均匀分布 $\mathrm{U}(-1, 1)$

中采样得到。

由于隐式分布的概率密度函数无法表示，均摊隐式对抗分布的熵无法直接优化。一个解决方案是最大化熵的变分下界（variational lower bound）[149]，该方法的有效性也在生成对抗网络中被验证[150]。在基于均摊隐式对抗分布的对抗分布训练中，可以通过对抗扰动 δ_i 与随机噪声 z 之间的互信息（mutual information）推导出如下所示的变分下界：

$$H(p_\phi(\delta_i|x_i)) \geqslant \mathcal{U}(q) = \mathbb{E}_{p(z)} \log q(z|g_\phi(z;x_i)) + c, \tag{5.13}$$

其中 c 是一个常数，$q(\cdot|\cdot)$ 是所引入的变分分布。通过优化 $\mathcal{U}(q)$ 可以有效地优化 $H(p_\phi(\delta_i|x_i))$。在实际中，变分分布 q 通过对角高斯分布定义，其均值和方差通过参数为 ψ 的网络输出。因此，对抗分布训练中最小最大化优化问题（5.5）变为：

$$\min_\theta \max_{\phi,\psi} \frac{1}{n} \sum_{i=1}^{n} \left\{ \mathbb{E}_{p(z)} \left[J(f_\theta(x_i + g_\phi(z;x_i)), y_i) + \lambda \cdot \log q_\psi(z|g_\phi(z;x_i)) \right] \right\}. \tag{5.14}$$

该损失函数同样可以通过对于模型参数 θ 的随机梯度下降和对于生成网络及变分网络参数 (ϕ, ψ) 的随机梯度上升进行求解。

5.5　实验结果

本节在 CIFAR-10[134]，CIFAR-100[134]和 SVHN[135]数据集上进行实验①，其中图片像素值的取值范围为 $[0, 1]$。在 CIFAR-10 和 CIFAR-100 数据集上的最大扰动规模设置为 $\epsilon = \frac{8}{255}$，在 SVHN 数据集上的最大扰动规模设置为 $\epsilon = \frac{4}{255}$。

本节选取 Wide ResNet（WRN-28-10）模型[151]作为分类器。均摊显式分布建模和均摊隐式分布建模中的生成网络选取为基于残差连接的图像到图像（image-to-image）结构[152-153]。均摊隐式分布建模中的变分网络 q_ψ 选取为五层卷积神经网络结构。

对抗分布训练中最小最大优化问题（5.5）的分类损失 J 选取为交叉熵损失，超参数设置为 $\lambda = 0.01$，第5.5.3节进一步研究 λ 的不同取值对于结果的影响。显式分布建模采用 Adam 优化器[154]优化显式对抗分布的参数 ϕ_i，其中学习率设置为 0.3，迭代轮数设置为 $T = 7$，蒙特卡洛采样次数设置为 $k = 5$。均摊显式分布建模和均摊隐式分布建模同样采用 Adam 优化器，但是将蒙特卡洛采样次数设置为 $k = 1$ 以加速训练过程。

本节将采用三种对抗分布建模方式的对抗分布训练记为 ADT_{EXP}，$\text{ADT}_{\text{EXP-AM}}$ 和 $\text{ADT}_{\text{IMP-AM}}$，并与正常训练（记为 Standard）、基于投影梯度下降法的对抗训练

① 源代码参见https://github.com/dongyp13/Adversarial-Distributional-Training。

表 5.1　在 CIFAR-10 数据集上不同模型针对白盒对抗攻击的鲁棒性对比

	$\mathcal{A}_{\text{nat}}$	FGSM	PGD-20	PGD-100	MIM	C&W	FeaAttack	$\mathcal{A}_{\text{rob}}$
Standard	**94.81%**	12.05%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
AT_{FGSM}	93.80%	**79.86%**	0.12%	0.04%	0.06%	0.13%	0.01%	0.01%
$\text{AT}_{\text{PGD}}{}^{\dagger}$	87.25%	56.04%	45.88%	45.33%	47.15%	46.67%	46.01%	44.89%
AT_{PGD}	86.91%	58.30%	50.03%	49.40%	51.40%	50.23%	50.46%	48.26%
ALP	86.81%	56.83%	48.97%	48.60%	50.13%	49.10%	48.51%	47.90%
FeaScatter	89.98%	**77.40%**	**70.85%**	**68.81%**	**72.74%**	**58.46%**	37.45%	37.40%
ADT_{EXP}	86.89%	60.41%	52.18%	**51.69%**	**53.27%**	52.49%	**52.38%**	**50.56%**
$\text{ADT}_{\text{EXP-AM}}$	87.82%	62.42%	51.95%	51.26%	52.99%	51.75%	52.04%	50.04%
$\text{ADT}_{\text{IMP-AM}}$	88.00%	**64.89%**	**52.28%**	51.23%	52.64%	**52.65%**	51.89%	49.81%

表中每一行代表一个防御模型，每一列代表一种攻击算法，橙色标记的结果代表模型存在针对不同攻击泛化能力不足的问题。

（记为 AT_{PGD}）[37]、预训练的 AT_{PGD} 模型（记为 $\text{AT}_{\text{PGD}}{}^{\dagger}$）[37]、基于快速梯度符号法的对抗训练（记为 AT_{FGSM}）[69]、对抗 logit 匹配（adversarial logit pairing，记为 ALP）[155]和基于特征分散的对抗训练（记为 FeaScatter）[129]进行对比。

为了更好地测评这些防御模型的鲁棒性，本节选取大量攻击方法，并使用逐样本准确率（per-example accuracy）[39]作为评估指标，其计算公式如下：

$$\mathcal{A}_{\text{rob}} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \min_{a \in \mathbf{A}} \mathbb{I}\big(\arg\max\{ f_\theta(a(x_i)) \} = y_i \big), \tag{5.15}$$

其中 n_{test} 代表测试样本数量，$\mathbf{A}$ 代表一组攻击方法，$a(x_i)$ 代表使用攻击方法 a 生成的对抗样本，$\mathbb{I}(\cdot)$ 代表指示函数。$\mathcal{A}_{\text{rob}}$ 可以更加准确地评估模型的整体鲁棒性。

5.5.1　白盒攻击实验

本小节首先对比对抗分布训练与基准方法在各种白盒攻击下的鲁棒性。实验采用快速梯度符号法（FGSM）[20]、投影梯度下降法（PGD）[37]、动量迭代法（MIM）、C&W 方法[49]和特征攻击（记为 FeaAttack）①作为白盒攻击方法。本小节使用的 C&W 方法采用公式（2.7）中的目标函数并利用 PGD 进行优化。实验设置 PGD 的迭代轮数为 20 和 100 步，MIM 的迭代轮数为 20 步，C&W 的迭代轮数为 30 步。在这些攻击中，迭代步长设置为 $\alpha = \frac{\epsilon}{4}$。FeaAttack 采用默认设置。

表5.1展示了在 CIFAR-10 数据集上不同模型在白盒攻击下的鲁棒性，其中 $\mathcal{A}_{\text{nat}}$ 代表模型在真实测试数据上的准确率，橙色标记的结果代表模型存在针对不同攻

① 特征攻击代码参见https://github.com/Line290/FeatureAttack。

表 5.2　在 CIFAR-100 和 SVHN 数据集上不同模型针对白盒对抗攻击的鲁棒性对比

	$\mathcal{A}_{\text{nat}}$	FGSM	PGD-20	PGD-100	MIM	C&W	FeaAttack	$\mathcal{A}_{\text{rob}}$
(a) CIFAR-100 数据集								
Standard	**78.59%**	8.73%	0.02%	0.01%	0.02%	0.00%	0.00%	0.00%
AT_{PGD}	61.45%	30.78%	25.71%	25.40%	26.60%	25.80%	33.95%	24.49%
ADT_{EXP}	62.70%	34.22%	28.96%	**28.60%**	**29.83%**	**28.99%**	**35.07%**	**27.13%**
$\text{ADT}_{\text{EXP-AM}}$	62.84%	36.28%	29.01%	28.46%	29.68%	28.78%	34.91%	26.87%
$\text{ADT}_{\text{IMP-AM}}$	64.07%	**39.39%**	**29.40%**	28.43%	29.64%	28.76%	35.00%	26.80%
(b) SVHN 数据集								
Standard	**96.12%**	39.05%	3.64%	2.95%	4.08%	3.91%	2.14%	2.14%
AT_{PGD}	95.07%	82.19%	74.22%	73.79%	74.56%	74.77%	73.51%	73.38%
ADT_{EXP}	95.70%	86.72%	**77.01%**	**76.62%**	**77.18%**	**77.50%**	**75.64%**	**75.55%**
$\text{ADT}_{\text{EXP-AM}}$	95.67%	85.24%	76.12%	75.58%	76.63%	76.70%	75.20%	75.00%
$\text{ADT}_{\text{IMP-AM}}$	95.62%	**86.73%**	75.61%	74.85%	75.91%	76.12%	74.24%	74.13%

击泛化能力不足的问题。从实验结果中可以看到，AT_{FGSM} 和 FeaScatter 存在此问题，即其虽然在某些攻击下防御能力更好，但是会被更强大的攻击攻破，导致整体鲁棒性较差（见最后一列）。基于对抗分布训练的模型在所有攻击下均表现出良好的鲁棒性，可以有效解决此问题。虽然 AT_{PGD} 也不存在这一问题，并且在对抗训练的模型中取得了最好的鲁棒性，但是对抗分布训练相比于 AT_{PGD} 显著提高了模型的鲁棒性，验证了对抗分布训练的有效性。表5.2展示了在 CIFAR-100 和 SVHN 数据集上的结果。实验结果一致地表明在白盒攻击下基于对抗分布训练的模型可以取得比对抗训练更好的鲁棒性。

从实验结果中还可以看到，基于显式分布建模的 ADT_{EXP} 在大多数情况下优于基于生成网络均摊建模的 $\text{ADT}_{\text{EXP-AM}}$ 和 $\text{ADT}_{\text{IMP-AM}}$。产生此现象的原因可能是生成网络的容量有限，通过生成网络学习对抗分布的方式很难为每个输入数据学习适当的对抗分布。尽管如此，$\text{ADT}_{\text{EXP-AM}}$ 和 $\text{ADT}_{\text{IMP-AM}}$ 可以加速训练过程。值得注意的是，$\text{ADT}_{\text{IMP-AM}}$ 并没有取得比 $\text{ADT}_{\text{EXP-AM}}$ 更好的鲁棒性。这说明虽然采用隐式分布可以更加灵活地建模对抗扰动，但是实验结果表明隐式分布并没有帮助训练出更鲁棒的模型。

5.5.2　黑盒攻击实验

本小节在 CIFAR-10 数据集上测评不同防御模型在黑盒攻击下的鲁棒性，以更加全面地验证所提出方法的有效性。首先，实验对比模型在黑盒迁移攻击下的效

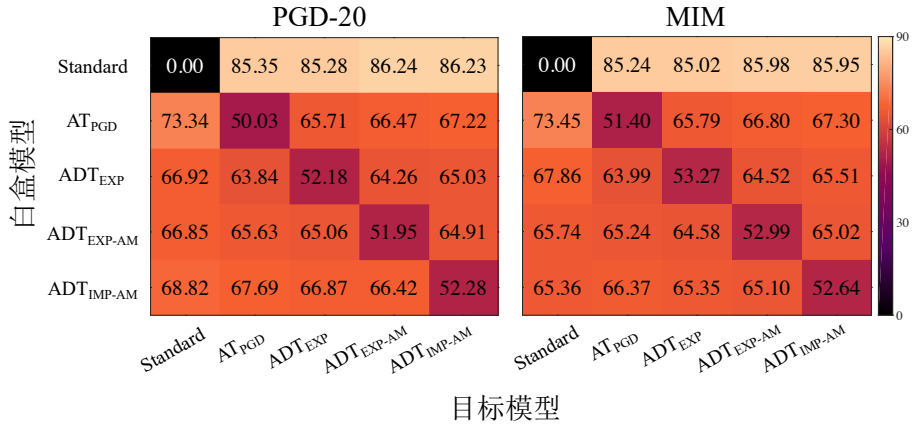


图 5.3　在 CIFAR-10 数据集上不同模型针对黑盒迁移攻击的准确率（%）

表 5.3　在 CIFAR-10 数据集上不同模型针对 SPSA 攻击的准确率

	$SPSA_{256}$	$SPSA_{512}$	$SPSA_{1024}$	$SPSA_{2048}$
Standard	0.00%	0.00%	0.00%	0.00%
AT_{PGD}	60.67%	58.10%	55.82%	54.37%
ADT_{EXP}	62.22%	59.94%	**57.97%**	**56.27%**
$ADT_{EXP\text{-}AM}$	**62.58%**	**60.12%**	57.62%	55.84%
$ADT_{IMP\text{-}AM}$	62.49%	59.77%	57.34%	55.67%

果。图5.3展示了 Standard、AT_{PGD}、ADT_{EXP}、$ADT_{EXP\text{-}AM}$ 和 $ADT_{IMP\text{-}AM}$ 在 PGD-20 和 MIM 迁移攻击下的分类准确率。实验结果表明防御模型在黑盒迁移攻击下的准确率比在白盒攻击下的准确率更高。其次，实验采用黑盒得分攻击 SPSA[88]测评模型的鲁棒性。SPSA 中的采样次数分别设为 256、512、1024 和 2048 以估计模型的梯度。表5.3展示了不同模型在 SPSA 攻击下的准确率。可以看到，模型在 SPSA 攻击下的准确率也高于其在白盒攻击下的准确率，并且对抗分布训练取得了比 AT_{PGD} 更好的鲁棒性。因此，在黑盒攻击下的实验结果说明了对抗分布训练可靠地增强了模型的鲁棒性。

5.5.3　消融实验

首先，本小节测试第5.4节中提出的三种对抗分布建模方式（即 EXP、EXP-AM 和 IMP-AM）的攻击性能。EXP 的迭代轮数设置为 $T = 20$，蒙特卡洛采样次数设置为 $k = 10$，以构建更加强大的攻击。EXP-AM 和 IMP-AM 对于每个防御模型重新训练生成网络。在使用不同方式学习到对抗分布之后，从中采样对抗扰动并计算模型的分类准确率。表5.4展示了 PGD-20、EXP、EXP-AM 和 IMP-AM 针对 Standard、AT_{PGD}、ADT_{EXP}、$ADT_{EXP\text{-}AM}$ 和 $ADT_{IMP\text{-}AM}$ 五个模型的攻击结果。从

表 5.4　对抗分布的攻击结果对比

	PGD-20	EXP	EXP-AM	IMP-AM
Standard	0.00%	0.00%	9.24%	9.83%
AT_{PGD}	50.03%	49.97%	50.46%	50.36%
ADT_{EXP}	52.18%	51.96%	52.71%	52.82%
$ADT_{EXP\text{-}AM}$	51.95%	51.62%	52.85%	52.72%
$ADT_{IMP\text{-}AM}$	52.28%	51.46%	52.76%	52.48%

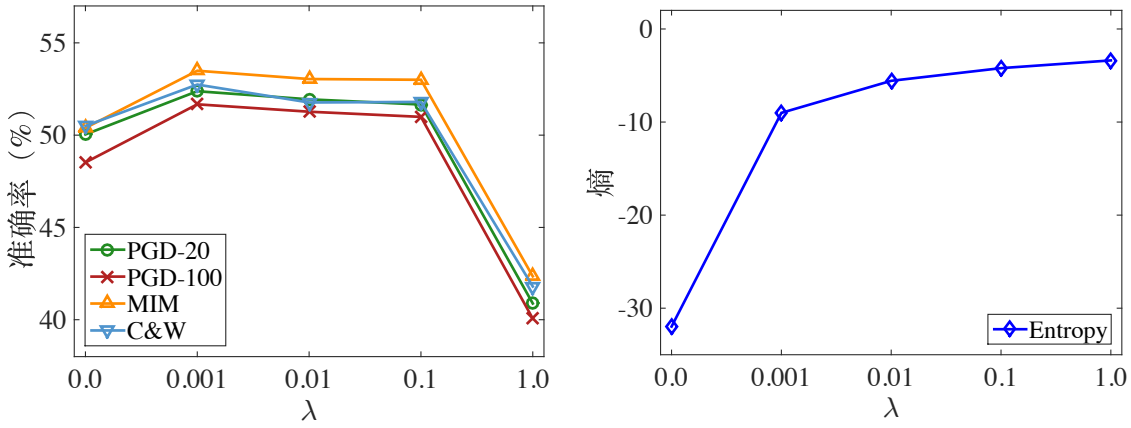
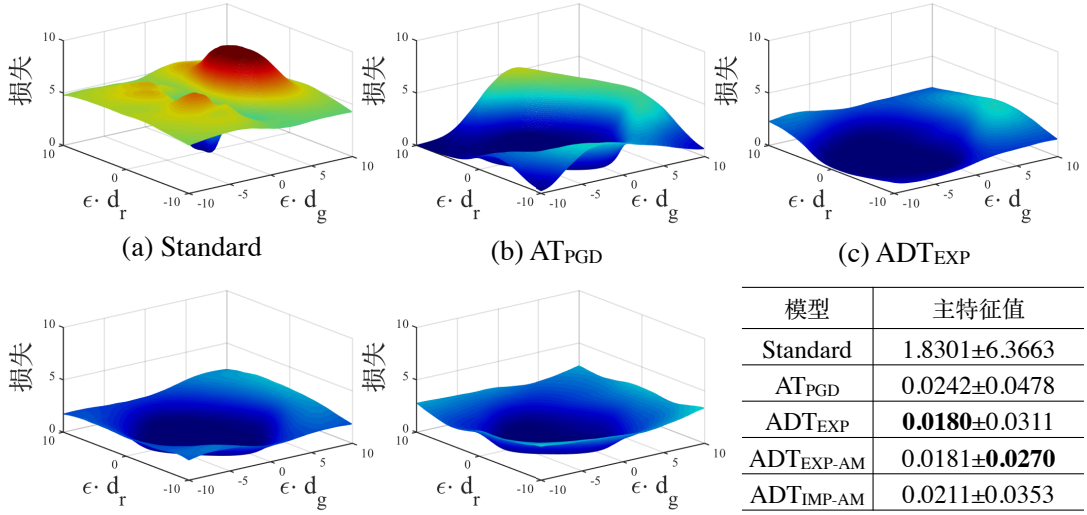


图 5.4　在不同 λ 取值下模型的鲁棒性与对抗分布的熵

中可以看到，EXP 取得了比 PGD-20 更好的攻击效果，而 EXP-AM 和 IMP-AM 也表现出类似的攻击能力。

其次，本小节研究对抗分布训练中最小最大优化问题（5.5）的超参数 λ 的取值对于结果的影响。由于 $ADT_{EXP\text{-}AM}$ 训练速度较快，同时显式对抗分布的熵可以被解析计算，实验选取其作为研究对象。图5.4展示了在 $\lambda = 0.0$、0.001、0.01、0.1 和 1.0 下训练得到的 $ADT_{EXP\text{-}AM}$ 模型的鲁棒性以及对抗分布的熵。可以看到，采用更大的 λ 可以使对抗分布的熵更大，并得到更好的模型鲁棒性。这是因为对抗分布可以学习到多样化的对抗样本，增强模型的鲁棒性。但是，如果 λ 的取值过大，对抗分布对模型的攻击效果变差，导致模型的鲁棒性出现下降。

最后，本小节分析不同方法训练得到模型的损失平面。具体而言，通过计算模型在原始样本周围沿梯度方向 d_g 和随机方向 d_r 上的交叉熵损失得到损失平面。图5.5展示了 Standard、AT_{PGD}、ADT_{EXP}、$ADT_{EXP\text{-}AM}$ 和 $ADT_{IMP\text{-}AM}$ 五个模型在某一个数据周围的损失平面。从图中可以看到，基于对抗分布训练的模型相比于 AT_{PGD} 学习到了更加平滑的损失表面，说明模型对于输入空间中的微小扰动不会过于敏感，从而取得更好的鲁棒性。图5.5 (f) 还展示了不同模型分类损失对于输入的黑塞矩阵（Hessian matrix）的主特征值（dominant eigenvalue），用于定量地衡量



(a) Standard　(b) AT_{PGD}　(c) ADT_{EXP}

模型	主特征值
Standard	1.8301±6.3663
AT_{PGD}	0.0242±0.0478
ADT_{EXP}	**0.0180**±0.0311
$ADT_{EXP\text{-}AM}$	0.0181±**0.0270**
$ADT_{IMP\text{-}AM}$	0.0211±0.0353

(d) $ADT_{EXP\text{-}AM}$　(e) $ADT_{IMP\text{-}AM}$　(f) 黑塞矩阵对比

图 5.5　模型损失平面可视化与模型对于输入的黑塞矩阵主特征值

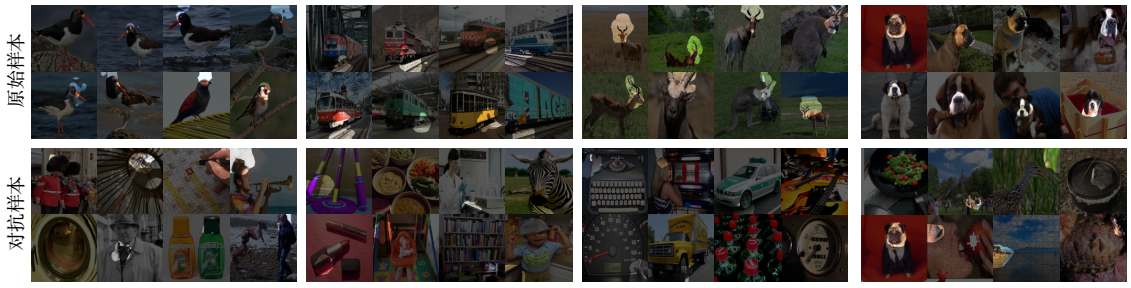
损失平面的平滑程度。图中的结果通过 CIFAR-10 测试集中的 1000 张图片计算得到。此结果表明对抗分布训练得到模型的黑塞矩阵主特征值更小，反映出模型的损失平面更加平滑，与可视化结果相吻合。

5.6　可解释性分析

对抗训练除了可以增强模型的鲁棒性，还具备提升深度学习模型的可解释性（interpretability）的作用[45]。这是因为对抗训练模型会学习到输入数据中更加鲁棒的特征[45]，与人类所理解的语义概念更加一致。本节在 ImageNet[87] 数据集上进一步分析对抗训练对于深度学习模型可解释性的作用。

一些研究发现深度学习模型中的神经元（neuron）可以检测人类所理解的语义概念[156-157]。然而，已有工作仅在原始样本上分析模型的可解释性，而没有在导致模型发生错误的对抗样本上验证模型的可解释性。针对这一问题，本节首先通过基于优化的攻击方法对 ImageNet 数据集中的每一张原始图片生成对抗样本，并将原始样本与对抗样本同时输入深度学习模型检查其特征表示的可解释性。具体而言，本节针对深度学习模型内部神经元学习到的特征进行分析，采用响应最大化（activation maximization）算法[157]检查其特征表示。对于每个神经元，该方法寻找对其产生响应最强的一组图片代表该神经元学习到的特征。

图5.6展示了在 ImageNet 数据集上正常训练的 VGG-16[4] 模型中某些神经元学习到的特征的可视化结果。对于每个选取的神经元，图中展示了对其产生响应最强的八张原始图片和八张对抗图片作为代表，其中每张图片的显著区域通过差异图（discrepancy map）[157]可视化。从图5.6中可以看到，对每个神经元产生响应较



(a) 神经元 #147：鸟　(b) 神经元 #20：铁轨　(c) 神经元 #102：羊角　(d) 神经元 #38：狗

图 5.6　VGG-16 模型中神经元特征可视化

(a) 神经元 #38：狗　(b) 神经元 #20：铁轨　(c) 神经元 #12：塔　(d) 神经元 #135：鸟

图 5.7　对抗训练后 VGG-16 模型中神经元特征可视化

强的原始图片中具有较为一致的语义概念。例如，在图5.6 (a) 中，对第 147 个神经元产生响应较强的图片均为“鸟”的图片，所以此神经元学习到的特征可以解释为“鸟”。然而，对每个神经元产生响应较强的对抗图片中的语义信息与原始图片存在不一致，且不具有任何关联。这说明在对抗样本存在的情况下，深度学习模型中的神经元并不能一致地检测图像中的某些语义概念，说明其可解释性不足。

针对这一问题，本节提出加入特征表示一致性损失的对抗训练方式，在训练的过程中提升深度学习模型内部特征表示与人类所理解的语义概念之间的一致性。该方法使模型学习到对于对抗样本与原始样本更加接近的特征表示，减轻噪声对模型内部特征表示的干扰，使其内部神经元在相关的语义概念出现时才会产生响应，以提升模型的可解释性。该方法构建以下损失函数进行对抗训练：

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \max_{\delta_i \in S} \left[J(f_\theta(x_i + \delta_i), y_i) + w \cdot \| f'_\theta(x_i + \delta_i) - f'_\theta(x_i) \|_2^2, \right], \tag{5.16}$$

其中 $f'_\theta(\cdot)$ 代表模型 f_θ 中的特征表示部分，w 为控制两项损失所占权重的超参数。

图5.7展示了通过本节所提出的方法训练得到的 VGG-16 模型中神经元学习到的特征的可视化结果。从图中可以看到，通过加入特征表示一致性损失的对抗训练方式，模型内部神经元在原始样本和对抗样本作为输入的情况下均可以一致地检测图像中的语义概念，验证其学习到的特征的可解释性更好，与人类所理解的语义概念更加一致。

5.7　本章小结

本章提出了对抗分布训练框架，建模每个原始样本周围对抗扰动的分布。对抗分布训练被描述为一个最小最大优化问题，其中内层最大化旨在学习对抗分布，外层最小化旨在通过优化模型在对抗分布下的期望损失训练鲁棒的分类器。通过加入熵正则项，对抗分布能够生成多样化的对抗样本，增强模型在不同攻击下的泛化能力和鲁棒性。本章通过理论分析推导出求解对抗分布训练的通用算法，并提出三种对抗攻击方式建模并学习参数化的对抗分布。实验结果验证了对抗分布训练的有效性，其相比于多种对抗训练方法可以有效增强模型的鲁棒性。本章进一步分析对抗训练对深度学习模型可解释性的影响，并提出了加入特征表示一致性损失的对抗训练方式，以提升模型的可解释性。

第 6 章　面向图像分类的对抗鲁棒性测评基准

随着深度学习的鲁棒性问题受到广泛关注，越来越多的对抗攻击与防御算法被相继提出，如何公平合理地测评对抗攻防算法的有效性成为重要研究问题。本章面向图像分类任务构建了对抗鲁棒性测评基准，旨在公平、全面地评估图像分类任务中对抗攻防模型及算法的有效性。本章构建的鲁棒性基准采用两条鲁棒性曲线作为公正的评估指标，并在多种威胁模型下针对典型的对抗攻防算法进行大规模实验，以充分评估这些算法的性能。基于实验分析，本章得出一些重要结论：1）防御模型在同样攻击的不同参数（如扰动规模、迭代轮数等）下鲁棒性的相对好坏存在差异；2）对抗训练是增强深度学习模型鲁棒性最有效的方式，其鲁棒性可以泛化到其他威胁模型下；3）随机化防御在黑盒查询攻击下防御能力较好。本章开发了对抗攻防平台 ARES 用于鲁棒性测评，也为后续研究工作提供了便捷的测评工具。

6.1　本章引言

近年来，深度学习在对抗攻击下的鲁棒性与安全性问题引起广泛关注，研究者们提出了许多对抗攻击与防御算法，旨在发现深度学习模型的脆弱性以及构建更加安全可靠的深度学习模型。随着相关研究不断增多，如何对已有对抗攻防算法进行公平合理的鲁棒性测评成为重要研究问题，此方面的研究有助于了解不同对抗攻防算法的优缺点，比较其在多种威胁模型下的性能，并为构建更加鲁棒的深度学习模型提供新的思路。

深度学习鲁棒性的研究随着攻击与防御的博弈而不断发展[36,42,49,88,158]，如图6.1所示。这意味着某些防御方法在提出时虽然可以有效防御已有攻击，但是很快会被新的攻击方法攻破，反之亦然。例如，早期的一种防御方式采用快速梯度符号法进行对抗训练[69]，可以有效防御快速梯度符号法[20]生成的对抗样本，但是该防御会被其他攻击方法攻破，说明其并没有真正增强模型的鲁棒性。后续工作[38,77]设计不同的对抗防御机制，通过引起混淆梯度提升模型在典型白盒攻击下的防御能力，但是很容易被适应性攻击[36,42]攻破，也说明了这些防御模型的鲁棒性没有得到提升。因此，进行公平、全面、合理的鲁棒性测评对于了解不同攻防算法的效果及评估此领域的进展具有重要意义，同时此项工作也极具挑战。

目前关于深度学习鲁棒性测评的研究工作存在很多不足。首先，虽然一些开源对抗攻防平台（包括 CleverHans[89]、Foolbox[90]等）实现了典型的对抗攻防算

图 6.1　对抗攻击与对抗防御算法的发展过程

法，但是其涵盖的算法不够全面，也没有对已有攻防算法进行测评。其次，大部分对抗攻击与防御算法在提出时往往没有进行全面的鲁棒性测评。比如，大多数防御方法仅针对某些威胁模型下的一小部分攻击进行鲁棒性测试，没有覆盖到更加全面的对抗攻击方法。最后，目前使用较多的鲁棒性评估指标过于简单，无法全面分析对抗攻防算法的性能。在一般情况下，防御模型对具有某个扰动规模的对抗样本的识别准确率或对抗扰动的最小范数被用作鲁棒性评估指标，其不足以全面测评攻防算法的有效性。综上所述，目前鲁棒性测评机制不够完善，难以对比不同对抗攻防算法的优劣。

本章面向图像分类任务构建公平、全面的对抗鲁棒性测评基准（adversarial robustness benchmark），简称为鲁棒性基准，旨在对已有对抗攻防算法进行全面的测评与分析。鲁棒性基准涵盖多种典型的对抗攻击与防御，包括 15 种攻击方法和 16 个防御模型，这些防御模型是在广泛使用的 CIFAR-10[134] 和 ImageNet[87] 数据集上训练得到的。为了全面展示攻防算法的性能，鲁棒性基准采用两条不同的鲁棒性曲线作为公正的评估指标。本章在多种威胁模型下针对所选取的攻防算法进行大规模实验，以得到测评结果。鲁棒性基准所研究的威胁模型根据攻击者的目标包含有目标攻击和无目标攻击；根据攻击者的能力包含 ℓ_∞ 和 ℓ_2 范数限制下的攻击；根据攻击者的知识包含白盒攻击、黑盒迁移攻击、黑盒得分攻击和黑盒决策攻击。

本章基于实验分析得出一些重要结论。第一，防御模型在同样攻击的不同参数（如扰动规模、迭代轮数等）下鲁棒性的相对好坏存在差异，因此很难通过采用特定参数的攻击方法比较不同防御的鲁棒性，然而这种测评方式在当前工作中十分常见。第二，虽然存在多种类型的防御技术，但是最有效的防御仍然是对抗训练得到的模型，其鲁棒性可以泛化到其他威胁模型下。第三，随机化防御在黑盒

查询攻击（包括黑盒得分攻击与黑盒决策攻击）下防御能力较好。第6.3.3节会进一步介绍更多的实验结论。

本章开发了对抗攻防平台 ARES（Adversarial Robustness Evaluation for Safety）用于鲁棒性测评，同时其也支持了本章所进行的所有实验。ARES 平台包含了前面章节中所提出的对抗攻防算法，目前已开源①，可以大幅降低相关模型及算法的研发和测评门槛，并可以通过通用算法模块的研制降低新模型的开发成本，为后续攻防算法的研发提供了便捷的测评工具。

6.2　鲁棒性基准

本节将分别介绍鲁棒性基准中所包含的威胁模型、数据集、攻击算法、防御模型和评估指标，并简要介绍本章开发的对抗攻防平台。

6.2.1　威胁模型

本章构建的鲁棒性基准在多种威胁模型下进行鲁棒性测评，其包含第1.2.2节中所介绍的各种攻击形式。具体而言，对于攻击者的目标，鲁棒性基准包含有目标攻击和无目标攻击；对于攻击者的能力，鲁棒性基准包含在 ℓ_∞ 和 ℓ_2 范数限制下的攻击；对于攻击者的知识，鲁棒性基准包含白盒攻击、黑盒迁移攻击、黑盒得分攻击和黑盒决策攻击。

6.2.2　数据集

本章构建的鲁棒性基准使用 CIFAR-10[134]和 ImageNet[87]两个典型的图像分类数据集进行鲁棒性测评。鲁棒性基准选取 CIFAR-10 测试集中的 10000 张图片，并从 ImageNet 验证集中随机选择 1000 张图片。在进行有目标攻击时，每张图片的目标类别通过随机采样的方式选取。

6.2.3　攻击算法

对抗攻击通过优化攻击目标函数生成对抗样本，根据攻击目标函数的不同可以被分为两种类型。第一种攻击类型在最大扰动规模 ϵ 的限制下优化生成对抗样本，可以形式化表示为：

$$x^* = \underset{x' : \|x^* - x\|_p \leqslant \epsilon}{\arg\max} \; J(f(x'), y), \tag{6.1}$$

① 源代码参见https://github.com/thu-ml/ares。

表 6.1　鲁棒性基准中的对抗攻击算法

攻击算法	攻击者的目标	攻击者的能力	攻击者的知识	攻击目标函数
FGSM	无目标 & 有目标	ℓ_∞ & ℓ_2	白盒 & 黑盒迁移	约束攻击
BIM	无目标 & 有目标	ℓ_∞ & ℓ_2	白盒 & 黑盒迁移	约束攻击
MIM	无目标 & 有目标	ℓ_∞ & ℓ_2	白盒 & 黑盒迁移	约束攻击
DeepFool	无目标	ℓ_∞ & ℓ_2	白盒	最优化攻击
C&W	无目标 & 有目标	ℓ_2	白盒	最优化攻击
DIM	无目标 & 有目标	ℓ_∞ & ℓ_2	黑盒迁移	约束攻击
ZOO	无目标 & 有目标	ℓ_2	黑盒得分	最优化攻击
NES	无目标 & 有目标	ℓ_∞ & ℓ_2	黑盒得分	约束攻击
SPSA	无目标 & 有目标	ℓ_∞ & ℓ_2	黑盒得分	约束攻击
$\mathcal{N}$ATTACK	无目标 & 有目标	ℓ_∞ & ℓ_2	黑盒得分	约束攻击
Boundary	无目标 & 有目标	ℓ_2	黑盒决策	最优化攻击
Evolutionary	无目标 & 有目标	ℓ_2	黑盒决策	最优化攻击

其中 x 代表原始样本，y 代表 x 的真实类别，f 代表图像分类模型，$J(\cdot,\cdot)$ 代表攻击目标函数（如交叉熵损失）。本章将通过求解约束优化问题（6.1）生成对抗样本的方式称为*约束攻击方法*。可以看到，第2章介绍的动量迭代法和第3章介绍的平移不变攻击方法均为约束攻击方法。

第二种攻击类型在满足攻击者目标的情况下最小化添加扰动的范数，可以形式化表示为：

$$x^* = \underset{x':C(x')=1}{\arg\max} \|x^* - x\|_p, \tag{6.2}$$

其中 $C(\cdot)$ 代表对抗判别准则，即对抗样本满足攻击者的目标时其取值为 1，否则其取值为 0。本章将通过求解优化问题（6.2）生成对抗样本的方式称为*最优化攻击方法*。可以看到，第4章介绍的进化攻击方法为最优化攻击方法。

鲁棒性基准包含多种威胁模型下的 15 种对抗攻击算法，如下所示。

- 对于白盒攻击，鲁棒性基准包含快速梯度符号法（FGSM）[20]、基础迭代法（BIM）[48]、动量迭代法（MIM）、DeepFool[52]和 C&W[49]在内的 5 种算法。
- 对于黑盒迁移攻击，由于白盒攻击从原理上也可用于黑盒迁移攻击，鲁棒性基准采用 FGSM、BIM 和 MIM 进行黑盒迁移攻击，并加入了多样输入法（DIM）[56]，即其总共包含 4 种黑盒迁移攻击算法。
- 对于黑盒得分攻击，鲁棒性基准包含零阶优化法（ZOO）[59]、自然进化策略

表 6.2　鲁棒性基准中 CIFAR-10 数据集上的防御模型

防御模型	类型	模型结构	目标威胁模型	准确率
Res-56	正常训练	ResNet-56	-	92.6%
PGD-AT	鲁棒训练	Wide ResNet-28-10	ℓ_∞ ($\epsilon = 8/255$)	87.3%
DeepDefense	鲁棒训练	五层卷积神经网络	ℓ_2	79.7%
TRADES	鲁棒训练	Wide ResNet-34-10	ℓ_∞ ($\epsilon = 8/255$)	84.9%
Convex	可证实防御	ResNet	ℓ_∞ ($\epsilon = 2/255$)	66.3%
JPEG	图像变换	ResNet-56	通用	80.9%
RSE	随机化 & 模型集成	VGG	ℓ_2	86.1%
ADP	模型集成	3×ResNet-110	通用	94.1%

（NES）[60]、SPSA[88]和 $\mathcal{N}$ATTACK[61]在内的 4 种算法。

- 对于黑盒决策攻击，鲁棒性基准包含边界攻击（Boundary）[63]和进化攻击（Evolutionary）在内的 2 种算法。

表6.1总结了鲁棒性基准中对抗攻击算法的攻击目标、扰动范数、所获取的知识以及攻击目标函数。值得强调的是：1）鲁棒性基准没有采用投影梯度下降法（PGD）[37]，这是因为投影梯度下降法与基础迭代法十分类似，其取得的效果也十分接近；2）在黑盒迁移攻击中，攻击算法针对替代模型生成对抗样本；3）对于引起混淆梯度的防御，白盒攻击算法中均加入了适应性攻击策略，即在模型的梯度不存在时使用后向传递可微近似（BPDA）[36]或在模型的梯度随机时使用期望变换（EoT）[55]方法。基于上述对抗攻击算法的选取与实现，鲁棒性基准可以更好地测评不同防御模型的鲁棒性。

6.2.4　防御模型

鲁棒性基准总共包含 16 个模型，且这些模型覆盖第1.3.2节中所介绍的鲁棒训练、图像变换、随机化、模型集成和可证实防御五种类型。在 CIFAR-10 数据集上，鲁棒性基准包含 8 个模型，具体包括：

- 正常训练的 ResNet-56（Res-56）模型[6]；
- 基于投影梯度下降法的对抗训练模型（PGD-AT）[37]；
- 基于最大化扰动范数的 DeepDefense 防御模型[73]；
- 基于对抗训练的 TRADES 防御模型[43]；
- 基于外凸多胞体的可证实防御（Convex）[83]；
- 基于图像变换的 JPEG 压缩[53]；

表 6.3　鲁棒性基准中 ImageNet 数据集上的防御模型

防御模型	类型	模型结构	目标威胁模型	准确率
Inc-v3	正常训练	Inception v3	-	78.0%
Ens-AT	鲁棒训练	Inception v3	ℓ_∞ ($\epsilon = 16/255$)	73.5%
ALP	鲁棒训练	ResNet-50	ℓ_∞ ($\epsilon = 16/255$)	49.0%
FD	鲁棒训练	ResNet-152	ℓ_∞ ($\epsilon = 16/255$)	64.3%
JPEG	图像变换	Inception v3	通用	77.3%
Bit-Red	图像变换	Inception v3	通用	61.8%
R&P	图像变换 & 随机化	Inception v3	通用	77.0%
RandMix	可证实防御 & 随机化	Inception v3	通用	52.4%

- 随机自集成（random self-ensemble，RSE）防御模型[78]；
- 基于模型集成的 ADP 防御模型[79]。

表6.2展示了这些防御模型的具体信息，包括防御类型、模型结构、目标威胁模型（即防御模型在训练时采用的威胁模型）和在 CIFAR-10 测试集上的准确率。

在 ImageNet 数据集上，鲁棒性基准也包含 8 个防御模型，具体包括：

- 正常训练的 Inception v3（Inc-v3）模型[5]；
- 集成对抗训练（Ens-AT）防御模型[104]；
- 对抗 logit 匹配（ALP）防御模型[155]；
- 基于特征去噪（feature denoising，FD）的对抗训练模型[159]；
- 基于图像变换的 JPEG 压缩[53]；
- 基于图像变换的位深缩减防御（bit-depth reduction，Bit-Red）[160]；
- 随机放缩与填充（R&P）[77]；
- 可证实的 RandMix 防御模型[86]。

与表6.2类似，表6.3也展示了 ImageNet 数据集上防御模型的具体信息。可以看到，鲁棒性基准包含了多个基于随机化或图像变换的防御模型。虽然这些防御已经被攻破[36]，说明其鲁棒性较差，但是本章旨在全面地测评这些防御在多种威胁模型下的效果，因此鲁棒性基准也包含了这些较差的防御。

6.2.5　评估指标

本小节介绍鲁棒性基准中使用的评估指标。将攻击算法记为 $\mathcal{A}_{\epsilon,p}$，其可以在 ℓ_p 范数的限制下为原始样本 x 生成对抗样本 $x^* := \mathcal{A}_{\epsilon,p}(x)$，使得扰动的范数不超过 ϵ，即 $\|x^* - x\|_p \leqslant \epsilon$。与第2章的符号表示类似，给定分类器模型 f，其预测类

别记为 C_f。分类器在对抗攻击 $\mathcal{A}_{\epsilon,p}$ 下的分类准确率定义为：

$$\mathrm{Acc}(f, \mathcal{A}_{\epsilon,p}) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\big(C_f(\mathcal{A}_{\epsilon,p}(x_i)) = y_i\big), \tag{6.3}$$

其中 $\{x_i, y_i\}_{i=1}^{N}$ 代表数据集，$\mathbb{I}(\cdot)$ 代表指标函数。

对于无目标攻击，其针对分类器模型 f 的攻击成功率定义为：

$$\mathrm{Asr}(\mathcal{A}_{\epsilon,p}, f) = \frac{1}{M}\sum_{i=1}^{N} \mathbb{I}\big(C_f(x_i) = y_i \wedge C_f(\mathcal{A}_{\epsilon,p}(x_i)) \neq y_i\big), \tag{6.4}$$

其中 $M = \sum_{i=1}^{N} \mathbb{I}\big(C_f(x_i) = y_i\big)$ 代表模型分类正确的原始样本数量。

对于有目标攻击，其针对分类器模型 f 的攻击成功率定义为：

$$\mathrm{Asr}(\mathcal{A}_{\epsilon,p}, f) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\big(C_f(\mathcal{A}_{\epsilon,p}(x_i)) = y_i^*\big). \tag{6.5}$$

其中 y_i^* 代表 x_i 的攻击目标类别。

已有工作大多选定最大扰动规模 ϵ 并计算在此扰动规模下防御模型的分类准确率或攻击算法的攻击成功率作为评估指标，但是该指标过于简单，无法全面评估攻防算法的性能。本章构建的鲁棒性基准采用两条鲁棒性曲线作为评估指标，可以更加全面地展示防御模型在对抗攻击下的鲁棒性以及攻击算法针对防御模型的攻击效果。

第一条鲁棒性曲线为*分类准确率或攻击成功率随扰动规模变化曲线*。该曲线可以全面地评估在不同扰动下防御模型的鲁棒性及攻击算法的有效性。第二条鲁棒性曲线为*分类准确率或攻击成功率随攻击强度变化曲线*，其中攻击强度定义为攻击算法的迭代轮数或查询次数。该曲线可以更好地展示防御模型对攻击算法的抵抗能力及攻击算法的效率。

6.2.6　对抗攻防平台

为了搭建鲁棒性基准并针对上述对抗攻防算法进行全面测评，本章开发了对抗攻防平台 ARES。ARES 平台采用模块化的实现方式，包含数据集、攻击算法、防御模型和测评指标方面的通用实现。使用 ARES 平台进行鲁棒性测评较为便捷，如图6.2所示。ARES 平台的开发可以大幅降低对抗攻防模型及算法的研发和测评门槛，并可以通过通用算法模块降低新模型和新算法的开发成本。除了上述用于鲁棒性测评的对抗攻防算法，ARES 平台还包含了多种其他典型算法，并且涵盖了前面章节中所提出的攻防算法。ARES 平台目前已开源，为后续攻防算法的研发提供了便捷的测评工具。

```python
# import whatever benchmark you want
from ares.benchmark.distortion import DistortionBenchmark
from ares.model.loader import load_model_from_path
from ares.dataset import cifar10

session = ...  # load tf.Session
model = load_model_from_path('path/to/the/model.py').load(session)
dataset = cifar10.load_dataset_for_classifier(model, load_target=True)

# read documentation for the benchmark for all parameters
benchmark = DistortionBenchmark(attack_name='mim', model=model, ...)
# config the attack method
benchmark.config(decay_factor=1.0)

result = benchmark.run(dataset, some_logger)
```

图 6.2　使用对抗攻防平台进行鲁棒性测评示例

6.3　测评结果与分析

本节在多种威胁模型下针对所选取的攻防算法进行大规模实验，并采用两条鲁棒性曲线展示测评结果。本节主要展示防御模型针对 ℓ_∞ 范数限制下的无目标攻击的分类准确率随扰动规模与攻击强度的变化曲线。更为详细的实验结果请参见[99]。第6.3.1节和第6.3.2节分别展示了在 CIFAR-10 和 ImageNet 数据集上的测评结果。第6.3.3节总结了鲁棒性测评得出的重要结论。

6.3.1　CIFAR-10 数据集上的测评结果

本小节展示在 CIFAR-10 数据集上的 8 个防御模型针对白盒攻击、黑盒迁移攻击、黑盒得分攻击和黑盒决策攻击的分类准确率变化曲线。在计算分类准确率随攻击强度的变化曲线时，ℓ_∞ 范数限制下的最大扰动规模设置为 $\epsilon = \frac{8}{255}$，ℓ_2 范数限制下的最大扰动规模设置为 $\epsilon = 1.0$。

白盒攻击：图6.3和图6.4分别展示了 CIFAR-10 数据集上的防御模型针对 ℓ_∞ 范数限制下的无目标白盒攻击 FGSM、BIM、MIM 和 DeepFool 的分类准确率随扰动规模和攻击强度（即梯度迭代轮数）变化曲线。随着扰动规模增大，防御模型在基于多步梯度迭代的攻击算法下的分类准确率逐渐下降到 0。可以看到，在白盒攻击下对抗训练模型（包括 PGD-AT 和 TRADES）取得了更好的鲁棒性。此外，在不同的扰动规模或迭代轮数下，防御模型在同样对抗攻击下鲁棒性的相对好坏可能存在不同。例如，当扰动规模较小时（$\epsilon = 0.05$），TRADES 在白盒攻击下的分类准确率高于 PGD-AT，但在扰动规模较大时（$\epsilon = 0.15$），TRADES 在白盒攻击下的分类准确率低于 PGD-AT。这一现象意味着在特定的扰动规模或迭代轮数下防御模型之间的比较不能完全评估其性能，然而这种方式在已有工作中十分常见。本章构建的鲁棒性基准中采用的鲁棒性曲线可以更加全面地展示攻防算法的性能。

黑盒迁移攻击：图6.5和图6.6分别展示了 CIFAR-10 数据集上的防御模型针对

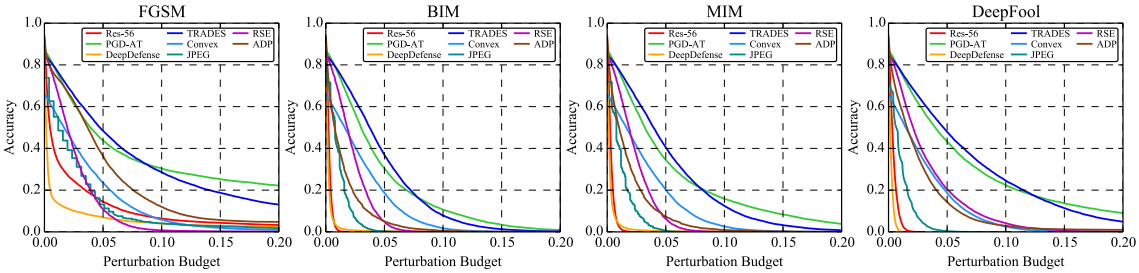


图 6.3　CIFAR-10 数据集上的防御模型针对 ℓ_∞ 范数限制下的无目标白盒攻击的分类准确率随扰动规模变化曲线

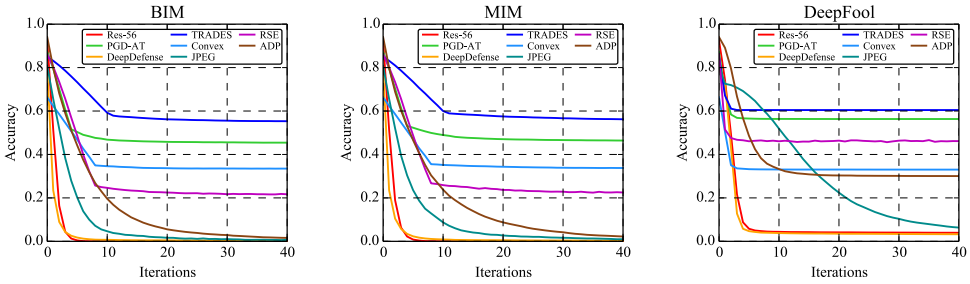


图 6.4　CIFAR-10 数据集上的防御模型针对 ℓ_∞ 范数限制下的无目标白盒攻击的分类准确率随攻击强度变化曲线

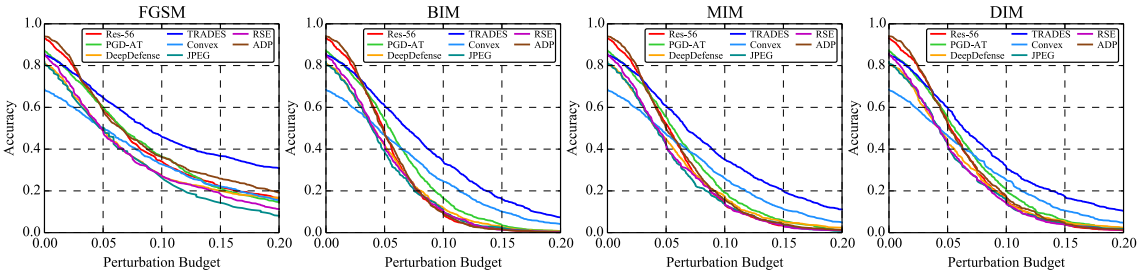


图 6.5　CIFAR-10 数据集上的防御模型针对 ℓ_∞ 范数限制下的无目标黑盒迁移攻击的分类准确率随扰动规模变化曲线

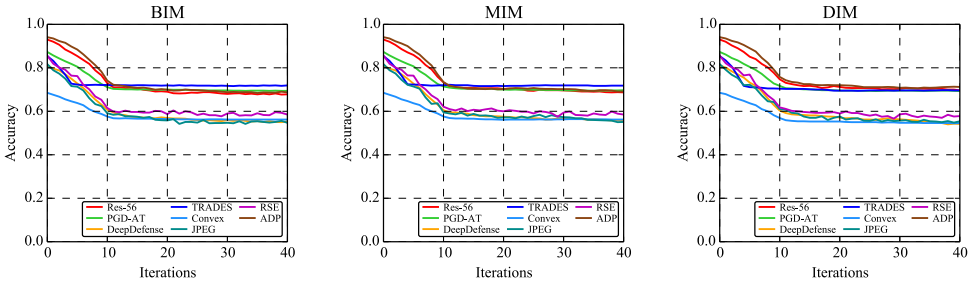


图 6.6　CIFAR-10 数据集上的防御模型针对 ℓ_∞ 范数限制下的无目标黑盒迁移攻击的分类准确率随攻击强度变化曲线

ℓ_∞ 范数限制下的无目标黑盒迁移攻击 FGSM、BIM、MIM 和 DIM 的分类准确率随扰动规模和攻击强度（即梯度迭代轮数）变化曲线。在攻击 TRADES 时，替代模型选取为 PGD-AT；在攻击其他模型时，替代模型选取为 TRADES。从实验结果中可以看到，一些提高黑盒迁移攻击成功率的方法（如 MIM 和 DIM）在 CIFAR-10

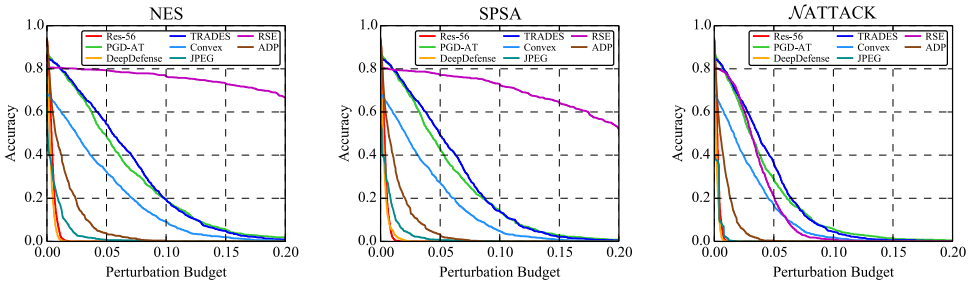


图 6.7　CIFAR-10 数据集上的防御模型针对 ℓ_∞ 范数限制下的无目标黑盒得分攻击的分类准确率随扰动规模变化曲线

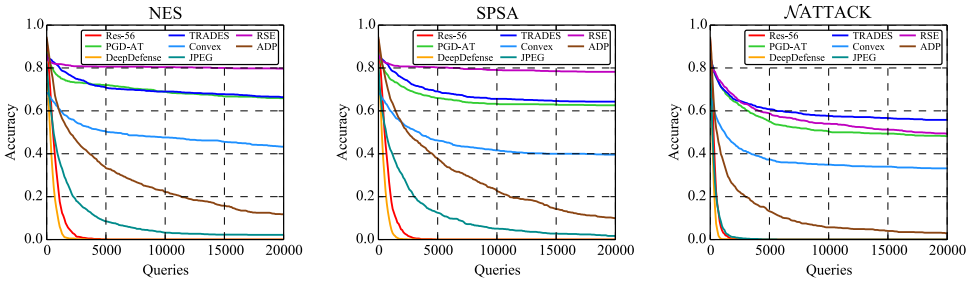


图 6.8　CIFAR-10 数据集上的防御模型针对 ℓ_∞ 范数限制下的无目标黑盒得分攻击的分类准确率随攻击强度变化曲线

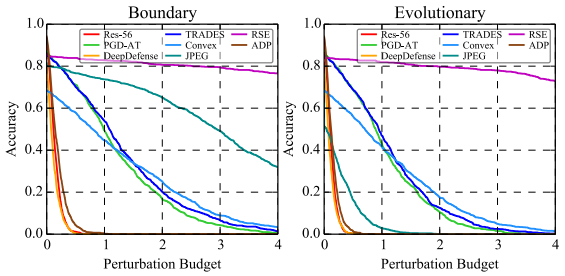


图 6.9　CIFAR-10 数据集上的防御模型针对 ℓ_2 范数限制下的无目标黑盒决策攻击的分类准确率随扰动规模变化曲线

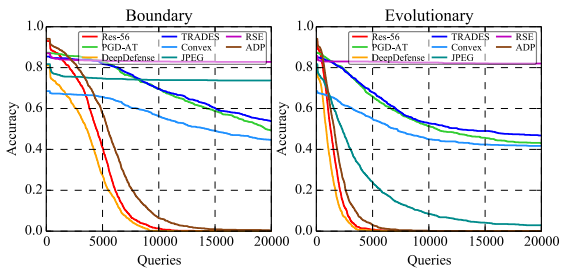


图 6.10　CIFAR-10 数据集上的防御模型针对 ℓ_2 范数限制下的无目标黑盒决策攻击的分类准确率随攻击强度变化曲线

数据集上没有取得比基准方法 BIM 更好的效果。

黑盒得分攻击：图6.7和图6.8分别展示了 CIFAR-10 数据集上的防御模型针对 ℓ_∞ 范数限制下的无目标黑盒得分攻击 NES、SPSA 和 $\mathcal{N}$ATTACK 的分类准确率随扰动规模和攻击强度（即查询次数）变化曲线。这些攻击算法的最大查询次数设置为 20000。从实验结果中可以看到，基于随机化的防御模型 RSE 对黑盒得分攻击具有很强的抵抗力，这是因为黑盒得分攻击估计梯度的随机性较大，难以有效生成对抗样本。

黑盒决策攻击：由于鲁棒性基准中的黑盒决策攻击算法仅支持 ℓ_2 范数限制下的攻击，图6.9和图6.10分别展示了 CIFAR-10 数据集上的防御模型针对 ℓ_2 范数限制下的无目标黑盒决策攻击 Boundary 和 Evolutionary 的分类准确率随扰动规模和

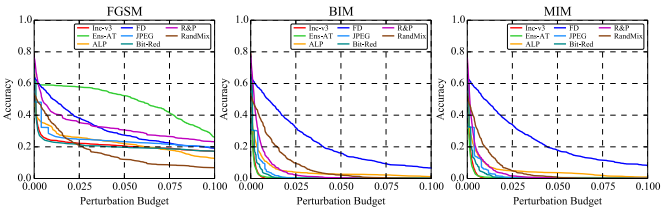


图 6.11　ImageNet 数据集上的防御模型针对 ℓ_∞ 范数限制下的无目标白盒攻击的分类准确率随扰动规模变化曲线

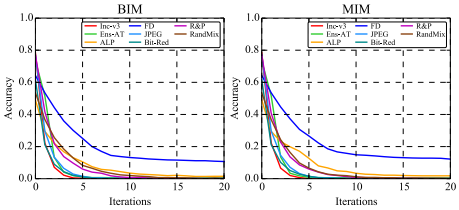


图 6.12　ImageNet 数据集上的防御模型针对 ℓ_∞ 范数限制下的无目标白盒攻击的分类准确率随攻击强度变化曲线

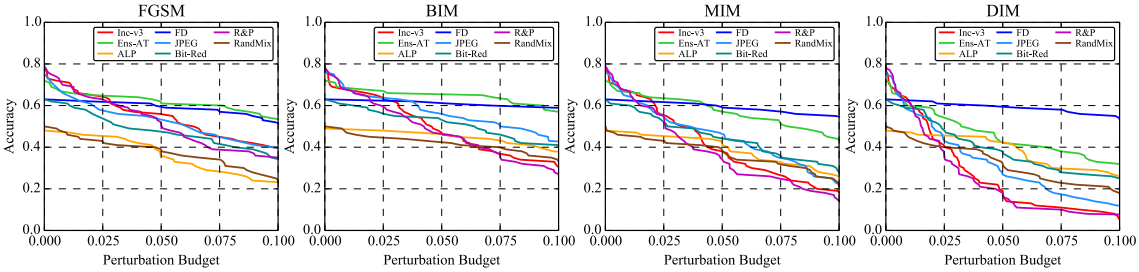


图 6.13　ImageNet 数据集上的防御模型针对 ℓ_∞ 范数限制下的无目标黑盒迁移攻击的分类准确率随扰动规模变化曲线

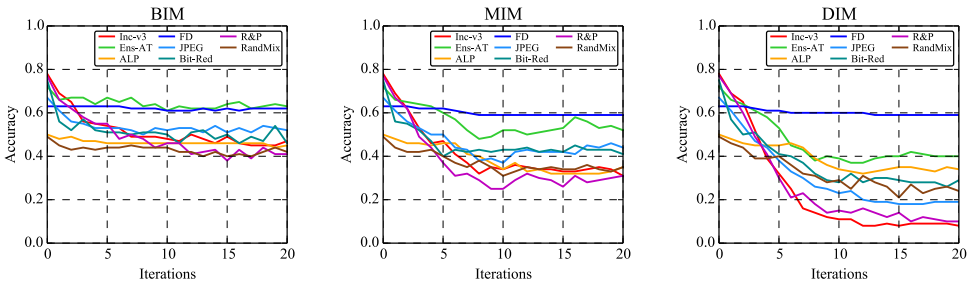


图 6.14　ImageNet 数据集上的防御模型针对 ℓ_∞ 范数限制下的无目标黑盒迁移攻击的分类准确率随攻击强度变化曲线

攻击强度（即查询次数）变化曲线。防御模型在黑盒决策攻击下的性能与其在黑盒得分攻击下的性能类似，RSE 也可以有效抵抗黑盒决策攻击。

6.3.2　ImageNet 数据集上的测评结果

本小节展示在 ImageNet 数据集上的鲁棒性测评结果。实验设置与 CIFAR-10 数据集类似。在计算分类准确率随攻击强度的变化曲线时，ℓ_∞ 范数限制下的最大扰动规模设置为 $\epsilon = \frac{16}{255}$，$\ell_2$ 范数限制下的最大扰动规模设置为 $\epsilon = \sqrt{0.001 \cdot D}$，其中 D 代表模型输入图片的维度。

白盒攻击：图6.11和图6.12分别展示了 ImageNet 数据集上的防御模型针对 ℓ_∞ 范数限制下的无目标白盒攻击 FGSM、BIM 和 MIM 的分类准确率随扰动规模和攻击强度（即梯度迭代轮数）变化曲线。可以看到，基于对抗训练的 FD 模型取得了

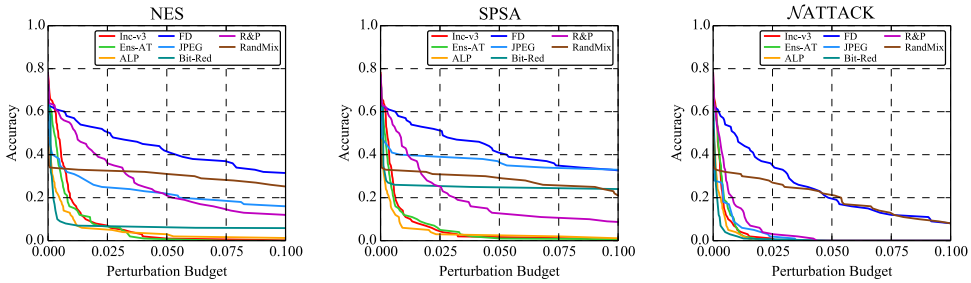


图 6.15　ImageNet 数据集上的防御模型针对 ℓ_∞ 范数限制下的无目标黑盒得分攻击的分类准确率随扰动规模变化曲线

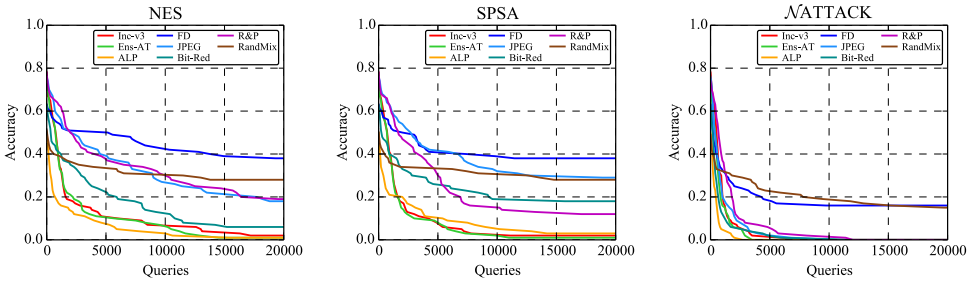


图 6.16　ImageNet 数据集上的防御模型针对 ℓ_∞ 范数限制下的无目标黑盒得分攻击的分类准确率随攻击强度变化曲线

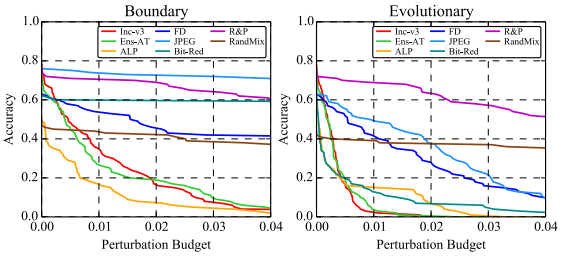


图 6.17　ImageNet 数据集上的防御模型针对 ℓ_2 范数限制下的无目标黑盒决策攻击的分类准确率随扰动规模变化曲线

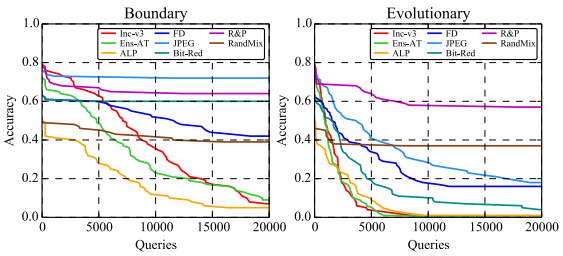


图 6.18　ImageNet 数据集上的防御模型针对 ℓ_2 范数限制下的无目标黑盒决策攻击的分类准确率随攻击强度变化曲线

最好的鲁棒性，说明了对抗训练在 ImageNet 数据集上也是最有效的防御方式。

黑盒迁移攻击：图6.13和图6.14分别展示了 ImageNet 数据集上的防御模型针对 ℓ_∞ 范数限制下的无目标黑盒迁移攻击 FGSM、BIM、MIM 和 DIM 的分类准确率随扰动规模和攻击强度（即梯度迭代轮数）变化曲线。实验使用 ResNet-152[6] 作为替代模型。与 CIFAR-10 数据集上的实验结果不同，MIM 和 DIM 显著提高了黑盒迁移攻击成功率。产生此现象的原因可能是：ImageNet 数据集中的图片维度较高，使得 BIM 生成的对抗样本会“过拟合”白盒模型[94]，难以迁移到其他黑盒模型上。

黑盒得分攻击与黑盒决策攻击：图6.15和图6.16分别展示了 ImageNet 数据集上的防御模型针对 ℓ_∞ 范数限制下的无目标黑盒得分攻击 NES、SPSA 和 $\mathcal{N}$ATTACK 的分类准确率随扰动规模和攻击强度（即查询次数）变化曲线。图6.17和图6.18分

别展示了 ImageNet 数据集上的防御模型针对 ℓ_2 范数限制下的无目标黑盒决策攻击 Boundary 和 Evolutionary 的分类准确率随扰动规模和攻击强度（即查询次数）变化曲线。与 CIFAR-10 数据集上的结果类似，基于随机化的两个防御模型 R&P 和 RandMix 比其他防御取得了更高的分类准确率。

6.3.3　实验结论

基于上述的鲁棒性测评结果，本章得出了以下结论。

第一，防御模型在同样攻击的不同参数（如扰动规模或迭代轮数）下鲁棒性的相对好坏存在差异。除了图6.3中 PGD-AT 和 TRADES 的结果可以证实此结论，其他场景下的结果也呈现出类似的现象。基于这一结论，在特定攻击参数下比较防御模型的效果并不能完全说明其优劣。采用鲁棒性曲线作为评估指标可以更加全面地比较不同模型的鲁棒性。

第二，对抗训练是目前最有效的防御方式，其鲁棒性也可以泛化到其他威胁模型下。例如，在 ℓ_∞ 范数限制下进行对抗训练得到的模型针对 ℓ_2 范数限制下的对抗攻击也具有一定防御能力。然而，对抗训练通常会导致模型在正常数据上的准确率下降且需要较高的训练代价。

第三，随机化防御在黑盒查询攻击（包括黑盒得分攻击与黑盒决策攻击）下具有良好的防御能力。由于随机化防御会对输入数据产生随机的预测结果，黑盒查询攻击在生成对抗样本的过程中梯度估计方向或搜索方向很可能被随机的预测结果误导，无法有效生成对抗样本。一个潜在的研究方向是针对随机化防御研究更加有效的黑盒查询攻击。

第四，基于图像变换的防御（如 JPEG、Bit-Red 等）可以略微提高正常训练模型的鲁棒性，并且在黑盒查询攻击下具有更好的防御能力。这些防御方法的实现十分简单，可以与其他类型的防御方式相结合，以构建更加鲁棒的防御模型。

第五，不同黑盒迁移攻击算法在 CIFAR-10 数据集上表现出相似的性能，而最近的方法（如 MIM、DIM 等）可以提高对抗样本在 ImageNet 数据集上的黑盒迁移攻击成功率。产生此现象的原因可能是：ImageNet 数据集中的图片维度远高于 CIFAR-10，因此 BIM 生成的对抗样本很容易 “过拟合” 白盒替代模型[94]，导致其迁移能力较差。

6.4　本章小结

本章面向图像分类任务构建了对抗鲁棒性测评基准，用于公平、全面地评估深度学习模型的鲁棒性及对抗攻防算法的有效性。鲁棒性基准包含 15 种对抗攻击

算法与 16 个对抗防御模型。鲁棒性基准采用两条不同的鲁棒性曲线作为公正的评估指标，并在多种威胁模型下针对对抗攻防算法进行大规模实验。通过实验分析，本章得出了一些与对抗攻防算法有效性相关的重要结论，有助于理解已有算法的性能并为后续研究工作提供新的思路。本章开发了对抗攻防平台 ARES 用于鲁棒性测评，也为今后攻防算法的研发提供了便捷的测评工具。

第 7 章　总结与展望

7.1　本文总结

本文围绕深度学习的对抗攻击与鲁棒性测评开展研究，旨在解决黑盒迁移攻击成功率较低、黑盒决策攻击效率较低、对抗训练模型鲁棒性不足以及对抗鲁棒性测评较为欠缺的问题。本文通过构建对抗攻防测评基准与平台，并面向不同场景研发高效对抗攻击算法解决上述问题，取得的主要创新性成果总结如下：

第2章提出了动量迭代法，在对抗样本生成的迭代优化过程中引入动量项，用于记录梯度更新的历史方向。动量迭代法可以使对抗样本生成过程中的更新方向更加平稳，同时避免对抗样本落入优化问题的局部极值点，有效缓解对抗样本的白盒攻击效果与迁移能力的相互制约，提高黑盒迁移攻击成功率。相比于已有攻击方法，动量迭代法将黑盒迁移攻击成功率提高了一倍左右，验证了现有深度学习模型在黑盒迁移攻击场景下鲁棒性不足的问题。

第3章提出了平移不变对抗攻击方法，其构建平移不变攻击目标函数，对一组经过平移变换的图片生成对抗样本，有效减轻对抗样本对所采用白盒模型的依赖程度。基于卷积神经网络的平移不变性，可以高效计算平移不变攻击目标函数的梯度。在不增加计算复杂度的情况下，平移不变攻击方法可以与所有基于梯度的攻击方法相结合。实验结果验证了平移不变攻击方法的有效性，其可以大幅提高针对典型防御模型的黑盒迁移攻击成功率。通过将平移不变攻击方法与多样输入法相结合得到的 TI-DIM 攻击方法，对八个典型防御模型的黑盒迁移攻击成功率达到了 82%。实验结果说明了所研究的对抗防御模型在黑盒迁移攻击场景下的脆弱性。

第4章针对人脸识别任务提出了进化攻击方法，可以在黑盒决策攻击场景下更加高效地生成对抗样本。进化攻击方法通过对搜索方向的局部几何结构进行建模，并降低搜索空间的维度，以提升黑盒决策攻击的查询效率。实验结果验证了进化攻击方法的有效性，其相比于已有方法可以通过更少的模型查询次数生成扰动更小的对抗样本。进化攻击方法成功攻破了商用人脸识别系统，表明其具有很好的实用性。实验结果表明现有人脸识别模型极易被黑盒决策攻击攻破，这很可能会对真实世界中的人脸识别系统带来严重的安全威胁。

第5章提出了对抗分布训练框架，其建模每个原始样本周围对抗扰动的分布。对抗分布训练被描述为最小最大优化问题，其中内层最大化旨在学习对抗分布，外层最小化旨在通过优化模型在对抗分布下的期望损失训练鲁棒的分类器。通过加

入熵正则项，对抗分布能够生成多样化的对抗样本，增强模型在不同攻击下的泛化能力和鲁棒性。对抗分布训练通过三种对抗攻击方式建模并学习参数化的对抗分布。实验结果验证了对抗分布训练的有效性，其相比于多种对抗训练方法可以有效增强模型的鲁棒性。第5章进一步分析对抗训练对深度学习模型可解释性的影响，并提出了加入特征表示一致性损失的对抗训练方式，以提升模型的可解释性。

第6章面向图像分类任务构建了对抗鲁棒性测评基准，用于公平、全面地评估深度学习模型的鲁棒性及对抗攻防算法的有效性。鲁棒性基准采用两条不同的鲁棒性曲线作为公正的评估指标，并针对对抗攻防算法在多种威胁模型下进行大规模实验。实验得出了与对抗攻防算法有效性相关的重要结论，有助于理解已有算法的性能并为后续研究工作提供新的思路。同时开发了对抗攻防平台 ARES 用于鲁棒性测评，也为今后对抗攻防算法的研发提供了便捷的测评工具。

7.2　未来工作展望

本文系统地研究了深度学习的对抗攻击与鲁棒性测评中存在的问题，面向不同场景提出了多种高效对抗攻击算法，构建了公平、全面的对抗鲁棒性测评基准与平台，一定程度上解决了上述问题，取得了阶段性成果。但是深度学习鲁棒性的研究还有很多极具挑战的问题需要进一步探索，具体包含以下三个方面：

- **真实世界中的对抗攻击**：本文主要研究数字世界中的对抗攻击，而真实世界中的对抗攻击会对深度学习的实际应用带来更加现实的安全威胁。在真实世界中，对抗攻击需要解决对于环境变化、空间相对位置变化、对抗扰动的物理可实现性等挑战，因此更加困难。尽管目前已有一些研究工作可以在真实世界中生成对抗样本[48,55,65-66]，但是其面临复现性较差的问题，即生成的对抗样本在不同的物理条件下难以复现，无法有效地评估真实世界中深度学习模型的鲁棒性。因此，面向真实世界中的不同场景研发对抗攻击方法具有重要意义，也是本文后续工作的重点研究方向。
- **构建安全可靠的人工智能**：以深度学习为代表的人工智能技术在多种场景下均存在鲁棒性与安全性问题。如何解决现有模型与算法的鲁棒性问题，构建更加安全可靠的人工智能成为重要的研究方向，对人工智能的进一步发展与应用具有重要的理论与现实意义。虽然目前主流的防御方式通过对抗训练等技术提升模型的鲁棒性，但其并未真正解决模型的本质问题。基于数据驱动的人工智能算法通过统计学习方式建模数据特征间的规律，拟合输入到输出的关系，而没有获取数据本质的特征及因果关系，因此无法完全解决鲁棒性的问题。为了构建安全可靠的人工智能算法，需要从理论与算法层面研究融

合先验知识、逻辑规则的新一代人工智能技术，这也是本文的长远目标。

- **多场景下的鲁棒性测评：**本文面向图像分类任务构建了对抗鲁棒性测评基准与平台，然而深度学习模型在实际中可能面临多种类型的攻击（如数据投毒攻击、模型窃取攻击等），如何面向多种攻击场景构建全面的鲁棒性测评基准是重要的研究问题。此外，随着深度学习被广泛应用于人脸识别、语音识别、自然语言处理、自动驾驶等各种任务中，针对不同任务与应用进行鲁棒性测评也具有重要意义。因此，在已有工作的基础上继续构建面向多场景的鲁棒性测评基准与平台是本文后续工作的另一重点研究方向。

参考文献

[1] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.

[2] Goodfellow I, Bengio Y, Courville A. Deep learning[M]. MIT press, 2016.

[3] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in Neural Information Processing Systems. 2012: 1097-1105.

[4] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [C]//International Conference on Learning Representations. 2015.

[5] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2818-2826.

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

[7] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 580-587.

[8] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[C]//Advances in Neural Information Processing Systems. 2015: 91-99.

[9] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 779-788.

[10] Taigman Y, Yang M, Ranzato M, et al. Deepface: Closing the gap to human-level performance in face verification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 1701-1708.

[11] Sun Y, Wang X, Tang X. Deep learning face representation from predicting 10,000 classes [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 1891-1898.

[12] Schroff F, Kalenichenko D, Philbin J. Facenet: A unified embedding for face recognition and clustering[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 815-823.

[13] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]// Advances in Neural Information Processing Systems. 2014: 3104-3112.

[14] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019: 4171-4186.

[15] Silver D, Huang A, Maddison C J, et al. Mastering the game of go with deep neural networks and tree search[J]. Nature, 2016, 529(7587): 484-489.

[16] Senior A W, Evans R, Jumper J, et al. Improved protein structure prediction using potentials from deep learning[J]. Nature, 2020, 577(7792): 706-710.

[17] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with alphafold [J]. Nature, 2021, 596(7873): 583-589.

[18] Biggio B, Corona I, Maiorca D, et al. Evasion attacks against machine learning at test time [C]//Joint European Conference on Machine Learning and Knowledge Discovery in Databases. 2013: 387-402.

[19] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[C]// International Conference on Learning Representations. 2014.

[20] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[C]// International Conference on Learning Representations. 2015.

[21] 张钹, 朱军, 苏航. 迈向第三代人工智能[J]. 中国科学: 信息科学, 2020, 50(09): 1281-1302.

[22] Wei X, Zhu J, Yuan S, et al. Sparse adversarial perturbations for videos[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019: 8973-8980.

[23] Carlini N, Wagner D. Audio adversarial examples: Targeted attacks on speech-to-text[C]//2018 IEEE Security and Privacy Workshops. 2018: 1-7.

[24] Qin Y, Carlini N, Cottrell G, et al. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition[C]//Proceedings of the 36th International Conference on Machine Learning. 2019: 5231-5240.

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

[26] Alzantot M, Sharma Y, Elgohary A, et al. Generating natural language adversarial examples[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 2890-2896.

[27] Muñoz-González L, Biggio B, Demontis A, et al. Towards poisoning of deep learning algorithms with back-gradient optimization[C]//Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. 2017: 27-38.

[28] Koh P W, Liang P. Understanding black-box predictions via influence functions[C]// Proceedings of the 34th International Conference on Machine Learning. 2017: 1885-1894.

[29] Shafahi A, Huang W R, Najibi M, et al. Poison frogs! targeted clean-label poisoning attacks on neural networks[C]//Advances in Neural Information Processing Systems. 2018: 6106-6116.

[30] Gu T, Dolan-Gavitt B, Garg S. Badnets: Identifying vulnerabilities in the machine learning model supply chain[J]. arXiv preprint arXiv:1708.06733, 2017.

[31] Chen X, Liu C, Li B, et al. Targeted backdoor attacks on deep learning systems using data poisoning[J]. arXiv preprint arXiv:1712.05526, 2017.

[32] Hendrycks D, Dietterich T. Benchmarking neural network robustness to common corruptions and perturbations[C]//International Conference on Learning Representations. 2019.

[33] Geirhos R, Rubisch P, Michaelis C, et al. Benchmarking neural network robustness to common corruptions and perturbations[C]//International Conference on Learning Representations. 2019.

[34] Akhtar N, Mian A. Threat of adversarial attacks on deep learning in computer vision: A survey [J]. IEEE Access, 2018, 6: 14410-14430.

[35] Biggio B, Roli F. Wild patterns: Ten years after the rise of adversarial machine learning[J]. Pattern Recognition, 2018, 84: 317-331.

[36] Athalye A, Carlini N, Wagner D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples[C]//Proceedings of the 35th International Conference on Machine Learning. 2018: 274-283.

[37] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks[C]//International Conference on Learning Representations. 2018.

[38] Liao F, Liang M, Dong Y, et al. 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.

[39] Carlini N, Athalye A, Papernot N, et al. On evaluating adversarial robustness[J]. arXiv preprint arXiv:1902.06705, 2019.

[40] Jing P, Tang Q, Du Y, et al. Too good to be safe: Tricking lane detection in autonomous driving with crafted perturbations[C]//30th USENIX Security Symposium. 2021: 3237-3254.

[41] Su D, Zhang H, Chen H, et al. Is robustness the cost of accuracy?–a comprehensive study on the robustness of 18 deep image classification models[C]//Proceedings of the European Conference on Computer Vision. 2018: 631-648.

[42] Tramer F, Carlini N, Brendel W, et al. On adaptive attacks to adversarial example defenses[C]// Advances in Neural Information Processing Systems. 2020: 1633-1645.

[43] 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. 2019: 7472-7482.

[44] Dong Y, Su H, Zhu J, et al. Towards interpretable deep neural networks by leveraging adversarial examples[J]. arXiv preprint arXiv:1708.05493, 2017.

[45] Tsipras D, Santurkar S, Engstrom L, et al. Robustness may be at odds with accuracy[C]// International Conference on Learning Representations. 2019.

[46] Engstrom L, Tran B, Tsipras D, et al. Exploring the landscape of spatial robustness[C]// Proceedings of the 36th International Conference on Machine Learning. 2019: 1802-1811.

[47] Song Y, Shu R, Kushman N, et al. Constructing unrestricted adversarial examples with generative models[C]//Advances in Neural Information Processing Systems. 2018: 8312-8323.

[48] Kurakin A, Goodfellow I, Bengio S. Adversarial examples in the physical world[C]// International Conference on Learning Representations Workshop. 2017.

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

[50] Papernot N, McDaniel P, Goodfellow I, et al. Practical black-box attacks against machine learning[C]//Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security. 2017: 506-519.

[51] Liu Y, Chen X, Liu C, et al. Delving into transferable adversarial examples and black-box attacks[C]//International Conference on Learning Representations. 2017.

[52] Moosavi-Dezfooli S M, Fawzi A, Frossard P. Deepfool: a simple and accurate method to fool deep neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2574-2582.

[53] 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.

[54] Guo C, Rana M, Cisse M, et al. Countering adversarial images using input transformations[C]// International Conference on Learning Representations. 2018.

[55] Athalye A, Engstrom L, Ilyas A, et al. Synthesizing robust adversarial examples[C]// Proceedings of the 35th International Conference on Machine Learning. 2018: 284-293.

[56] Xie C, Zhang Z, Zhou Y, et al. Improving transferability of adversarial examples with input diversity[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 2730-2739.

[57] Lin J, Song C, He K, et al. Nesterov accelerated gradient and scale invariance for adversarial attacks[C]//International Conference on Learning Representations. 2020.

[58] Wu D, Wang Y, Xia S T, et al. Skip connections matter: On the transferability of adversarial examples generated with resnets[C]//International Conference on Learning Representations. 2020.

[59] Chen P Y, Zhang H, Sharma Y, et al. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models[C]//Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. ACM, 2017: 15-26.

[60] Ilyas A, Engstrom L, Athalye A, et al. Black-box adversarial attacks with limited queries and information[C]//Proceedings of the 35th International Conference on Machine Learning. 2018: 2137-2146.

[61] Li Y, Li L, Wang L, et al. Nattack: Learning the distributions of adversarial examples for an improved black-box attack on deep neural networks[C]//Proceedings of the 36th International Conference on Machine Learning. 2019: 3866-3876.

[62] Wierstra D, Schaul T, Glasmachers T, et al. Natural evolution strategies[J]. Journal of Machine Learning Research, 2014, 15(27): 949-980.

[63] Brendel W, Rauber J, Bethge M. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models[C]//International Conference on Learning Representations. 2018.

[64] Cheng M, Le T, Chen P Y, et al. Query-efficient hard-label black-box attack: An optimization-based approach[C]//International Conference on Learning Representations. 2019.

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

[66] Eykholt K, Evtimov I, Fernandes E, et al. Robust physical-world attacks on deep learning visual classification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 1625-1634.

[67] Xu K, Zhang G, Liu S, et al. Adversarial t-shirt! evading person detectors in a physical world [C]//Proceedings of the European Conference on Computer Vision. 2020: 665-681.

[68] Cao Y, Wang N, Xiao C, et al. Invisible for both camera and lidar: Security of multi-sensor fusion based perception in autonomous driving under physical-world attacks[C]//IEEE Symposium on Security and Privacy. 2021: 176-194.

[69] Kurakin A, Goodfellow I, Bengio S. Adversarial machine learning at scale[C]//International Conference on Learning Representations. 2017.

[70] Pang T, Xu K, Dong Y, et al. Rethinking softmax cross-entropy loss for adversarial robustness [C]//International Conference on Learning Representations. 2020.

[71] Cisse M, Bojanowski P, Grave E, et al. Parseval networks: Improving robustness to adversarial examples[C]//Proceedings of the 34th International Conference on Machine Learning. 2017: 854-863.

[72] Hein M, Andriushchenko M. Formal guarantees on the robustness of a classifier against adversarial manipulation[C]//Advances in Neural Information Processing Systems. 2017: 2263-2273.

[73] Yan Z, Guo Y, Zhang C. Deep defense: Training dnns with improved adversarial robustness [C]//Advances in Neural Information Processing Systems. 2018: 417-426.

[74] Samangouei P, Kabkab M, Chellappa R. Defense-gan: Protecting classifiers against adversarial attacks using generative models[C]//International Conference on Learning Representations. 2018.

[75] Song Y, Kim T, Nowozin S, et al. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples[C]//International Conference on Learning Representations. 2018.

[76] Pang T, Xu K, Zhu J. Mixup inference: Better exploiting mixup to defend adversarial attacks [C]//International Conference on Learning Representations. 2020.

[77] Xie C, Wang J, Zhang Z, et al. Mitigating adversarial effects through randomization[C]// International Conference on Learning Representations. 2018.

[78] Liu X, Cheng M, Zhang H, et al. Towards robust neural networks via random self-ensemble [C]//Proceedings of the European Conference on Computer Vision. 2018: 369-385.

[79] Pang T, Xu K, Du C, et al. Improving adversarial robustness via promoting ensemble diversity [C]//Proceedings of the 36th International Conference on Machine Learning. 2019: 4970-4979.

[80] Raghunathan A, Steinhardt J, Liang P. Certified defenses against adversarial examples[C]// International Conference on Learning Representations. 2018.

[81] Raghunathan A, Steinhardt J, Liang P S. Semidefinite relaxations for certifying robustness to adversarial examples[C]//Advances in Neural Information Processing Systems. 2018: 10900-10910.

[82] Sinha A, Namkoong H, Duchi J. Certifying some distributional robustness with principled adversarial training[C]//International Conference on Learning Representations. 2018.

[83] Wong E, Kolter J Z. Provable defenses against adversarial examples via the convex outer adversarial polytope[C]//Proceedings of the 35th International Conference on Machine Learning. 2018: 5286-5295.

[84] Zhang B, Cai T, Lu Z, et al. Towards certifying l-infinity robustness using neural networks with l-inf-dist neurons[C]//Proceedings of the 38th International Conference on Machine Learning. 2021: 12368-12379.

[85] Cohen J M, Rosenfeld E, Kolter J Z. Certified adversarial robustness via randomized smoothing [C]//Proceedings of the 36th International Conference on Machine Learning. 2019: 1310-1320.

[86] Zhang Y, Liang P. Defending against whitebox adversarial attacks via randomized discretization [C]//The 22nd International Conference on Artificial Intelligence and Statistics. 2019: 684-693.

[87] Russakovsky O, Deng J, Su H, et al. Imagenet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

[88] Uesato J, O'Donoghue B, Oord A v d, et al. Adversarial risk and the dangers of evaluating against weak attacks[C]//Proceedings of the 35th International Conference on Machine Learning. 2018: 5025-5034.

[89] Papernot N, Faghri F, Carlini N, et al. Technical report on the cleverhans v2.1.0 adversarial examples library[J]. arXiv preprint arXiv:1610.00768, 2018.

[90] Rauber J, Brendel W, Bethge M. Foolbox: A python toolbox to benchmark the robustness of machine learning models[C]//ICML Workshop on Reliable Machine Learning in the Wild. 2017.

[91] Nicolae M I, Sinn M, Tran M N, et al. Adversarial robustness toolbox v1.0.0[J]. arXiv preprint arXiv:1807.01069, 2018.

[92] Goodman D, Xin H, Yang W, et al. Advbox: a toolbox to generate adversarial examples that fool neural networks[J]. arXiv preprint arXiv:2001.05574, 2020.

[93] Polyak B T. Some methods of speeding up the convergence of iteration methods[J]. USSR Computational Mathematics and Mathematical Physics, 1964, 4(5): 1-17.

[94] Dong Y, Liao F, Pang T, et al. Boosting adversarial attacks with momentum[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 9185-9193.

[95] Dong Y, Pang T, Su H, et al. Evading defenses to transferable adversarial examples by translation-invariant attacks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 4312-4321.

[96] Dong Y, Su H, Wu B, et al. Efficient decision-based black-box adversarial attacks on face recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 7714-7722.

[97] Dong Y, Deng Z, Pang T, et al. Adversarial distributional training for robust deep learning[C]// Advances in Neural Information Processing Systems. 2020: 8270-8283.

[98] 董胤蓬, 苏航, 朱军. 面向对抗样本的深度神经网络可解释性分析[J]. 自动化学报, 2020, 45: 1-12.

[99] Dong Y, Fu Q A, Yang X, et al. Benchmarking adversarial robustness on image classification [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 321-331.

[100] Cheng S, Dong Y, Pang T, et al. Improving black-box adversarial attacks with a transfer-based prior[C]//Advances in Neural Information Processing Systems. 2019: 10934-10944.

[101] Dong Y, Cheng S, Pang T, et al. Query-efficient black-box adversarial attacks guided by a transfer-based prior[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021: 1-1.

[102] Dong Y, Yang X, Deng Z, et al. Black-box detection of backdoor attacks with limited information and data[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 16482-16491.

[103] Dong Y, Ni R, Li J, et al. Stochastic quantization for learning accurate low-bit deep neural networks[J]. International Journal of Computer Vision, 2019, 127(11): 1629-1642.

[104] Tramèr F, Kurakin A, Papernot N, et al. Ensemble adversarial training: Attacks and defenses [C]//International Conference on Learning Representations. 2018.

[105] Duch W, Korczak J. Optimization and global minimization methods suitable for neural networks [J]. Neural computing surveys, 1998, 2: 163-212.

[106] Sutskever I, Martens J, Dahl G, et al. On the importance of initialization and momentum in deep learning[C]//Proceedings of the 30th International Conference on Machine Learning. 2013: 1139-1147.

[107] Szegedy C, Ioffe S, Vanhoucke V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning[C]//Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. 2017: 4278-4284.

[108] He K, Zhang X, Ren S, et al. Identity mappings in deep residual networks[C]//Proceedings of the European Conference on Computer Vision. 2016: 630-645.

[109] Abadi M, Barham P, Chen J, et al. Tensorflow: A system for large-scale machine learning[C]// 12th USENIX Symposium on Operating Systems Design and Implementation. 2016: 265-283.

[110] Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2921-2929.

[111] Moosavi-Dezfooli S M, Fawzi A, Fawzi O, et al. Universal adversarial perturbations[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1765-1773.

[112] LeCun Y, Bengio Y. Convolutional networks for images, speech, and time series[J]. Handbook of Brain Theory and Neural Networks, 1995, 3361(10).

[113] Goodfellow I, Lee H, Le Q V, et al. Measuring invariances in deep networks[C]//Advances in Neural Information Processing Systems. 2009: 646-654.

[114] Kauderer-Abrams E. Quantifying translation-invariance in convolutional neural networks[J]. arXiv preprint arXiv:1801.01450, 2017.

[115] Huang G B, Mattar M, Berg T, et al. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments[C]//Workshop on faces in'Real-Life'Images: detection, alignment, and recognition. 2008.

[116] Wen Y, Zhang K, Li Z, et al. A discriminative feature learning approach for deep face recognition [C]//Proceedings of the European Conference on Computer Vision. 2016: 499-515.

[117] Liu W, Wen Y, Yu Z, et al. Sphereface: Deep hypersphere embedding for face recognition [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 212-220.

[118] Wang H, Wang Y, Zhou Z, et al. Cosface: Large margin cosine loss for deep face recognition [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 5265-5274.

[119] Deng J, Guo J, Xue N, et al. Arcface: Additive angular margin loss for deep face recognition [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 4690-4699.

[120] Sharif M, Bhagavatula S, Bauer L, et al. Adversarial generative nets: Neural network attacks on state-of-the-art face recognition[J]. arXiv preprint arXiv:1801.00349, 2017.

[121] Hansen N, Ostermeier A. Completely derandomized self-adaptation in evolution strategies[J]. Evolutionary computation, 2001, 9(2): 159-195.

[122] Igel C, Suttorp T, Hansen N. A computational efficient covariance matrix update and a (1+1)-cma for evolution strategies[C]//Proceedings of the 8th Annual Conference on Genetic and Evolutionary Computation. 2006: 453-460.

[123] Ros R, Hansen N. A simple modification in cma-es achieving linear time and space complexity [C]//International Conference on Parallel Problem Solving from Nature. 2008: 296-305.

[124] Su J, Vargas D V, Sakurai K. One pixel attack for fooling deep neural networks[J]. IEEE Transactions on Evolutionary Computation, 2019, 23(5): 828-841.

[125] Rechenberg I. Evolutionsstrategien[M]//Simulationsmethoden in der Medizin und Biologie. Springer, 1978: 83-114.

[126] Kemelmacher-Shlizerman I, Seitz S M, Miller D, et al. The megaface benchmark: 1 million faces for recognition at scale[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 4873-4882.

[127] Song C, He K, Wang L, et al. Improving the generalization of adversarial training with domain adaptation[C]//International Conference on Learning Representations. 2019.

[128] Wong E, Rice L, Kolter J Z. Fast is better than free: Revisiting adversarial training[C]// International Conference on Learning Representations. 2020.

[129] Zhang H, Wang J. Defense against adversarial attacks using feature scattering-based adversarial training[C]//Advances in Neural Information Processing Systems. 2019: 1831-1841.

[130] Xiao C, Zhong P, Zheng C. Enhancing adversarial defenses by k-winners-take-all[C]// International Conference on Learning Representations. 2020.

[131] Croce F, Hein M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks[C]//Proceedings of the 37th International Conference on Machine Learning. 2020: 2206-2216.

[132] Tashiro Y, Song Y, Ermon S. Diversity can be transferred: Output diversification for white- and black-box attacks[C]//Advances in Neural Information Processing Systems. 2020: 4536-4548.

[133] Jang Y, Zhao T, Hong S, et al. Adversarial defense via learning to generate diverse attacks[C]//Proceedings of the IEEE International Conference on Computer Vision. 2019: 2740-2749.

[134] Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images[R]. University of Toronto, 2009.

[135] Netzer Y, Wang T, Coates A, et al. Reading digits in natural images with unsupervised feature learning[C]//NIPS Workshop on Deep Learning and Unsupervised Feature Learning. 2011.

[136] Jolliffe I T. Principal components in regression analysis[M]//Principal component analysis. Springer, 1986: 129-155.

[137] Danskin J M. The theory of max-min and its application to weapons allocation problems: volume 5[M]. Springer Science & Business Media, 2012.

[138] Kingma D P, Welling M. Auto-encoding variational bayes[C]//International Conference on Learning Representations. 2014.

[139] Ford N, Gilmer J, Carlini N, et al. Adversarial examples are a natural consequence of test error in noise[C]//Proceedings of the 36th International Conference on Machine Learning. 2019: 2280-2289.

[140] Salman H, Li J, Razenshteyn I, et al. Provably robust deep learning via adversarially trained smoothed classifiers[C]//Advances in Neural Information Processing Systems. 2019: 11292-11303.

[141] Ben-Tal A, Den Hertog D, De Waegenaere A, et al. Robust solutions of optimization problems affected by uncertain probabilities[J]. Management Science, 2013, 59(2): 341-357.

[142] Esfahani P M, Kuhn D. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations[J]. Mathematical Programming, 2018, 171(1-2): 115-166.

[143] Staib M, Jegelka S. Distributionally robust deep learning as a generalization of adversarial training[C]//NIPS workshop on Machine Learning and Computer Security. 2017.

[144] Blundell C, Cornebise J, Kavukcuoglu K, et al. Weight uncertainty in neural networks[C]//Proceedings of the 32th International Conference on Machine Learning. 2015: 1613-1622.

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

[146] Poursaeed O, Katsman I, Gao B, et al. Generative adversarial perturbations[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 4422-4431.

[147] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks[C]//International Conference on Learning Representations. 2016.

[148] Isola P, Zhu J Y, Zhou T, et al. Image-to-image translation with conditional adversarial networks [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 1125-1134.

[149] Dai Z, Almahairi A, Bachman P, et al. Calibrating energy-based generative adversarial networks [C]//International Conference on Learning Representations. 2017.

[150] Dai Z, Yang Z, Yang F, et al. Good semi-supervised learning that requires a bad gan[C]// Advances in Neural Information Processing Systems. 2017: 6513-6523.

[151] Zagoruyko S, Komodakis N. Wide residual networks[C]//Proceedings of the British Machine Vision Conference. 2016: 87.1-87.12.

[152] Johnson J, Alahi A, Fei-Fei L. Perceptual losses for real-time style transfer and super-resolution [C]//Proceedings of the European Conference on Computer Vision. 2016: 694-711.

[153] Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 2223-2232.

[154] Kingma D, Ba J. Adam: A method for stochastic optimization[C]//International Conference on Learning Representations. 2015.

[155] Kannan H, Kurakin A, Goodfellow I. Adversarial logit pairing[J]. arXiv preprint arXiv:1803.06373, 2018.

[156] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[C]//Proceedings of the European Conference on Computer Vision. 2014: 818-833.

[157] Zhou B, Khosla A, Lapedriza A, et al. Object detectors emerge in deep scene cnns[C]// International Conference on Learning Representations. 2015.

[158] Carlini N, Wagner D. Adversarial examples are not easily detected: Bypassing ten detection methods[C]//Proceedings of the 10th ACM workshop on artificial intelligence and security. 2017: 3-14.

[159] Xie C, Wu Y, Maaten L v d, et al. Feature denoising for improving adversarial robustness[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 501-509.

[160] 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. 2018.

致 谢

时光荏苒，我已在清华度过了八年多美好的时光。在求学的日子里，有获得新知识的喜悦，也有论文不中的烦恼，有老师一路的指导与教诲，也有同学相互的鼓励与陪伴。在博士毕业之际，我有一些话想说。

饮其流者怀其源，学其成时念吾师，首先衷心感谢我的导师朱军教授在博士期间对我的悉心指导与关怀。在刚读博士时，朱老师细心引导，帮助我找到了兴趣所在的研究方向；在科研工作中，朱老师指点迷津，指引我找对科研方法不断向前；在日常相处中，朱老师关心爱护，鼓励我健康成长。在这几年的科研工作中，朱老师严谨务实的科研态度，一丝不苟的治学精神，高屋建瓴的学术见地，勤奋谦虚的个人品质都深深感染着我，激励着我。朱老师的言传身教不仅培养了我的学术能力，更成为我未来人生路途中的榜样。

衷心感谢张钹教授对我的谆谆教诲，帮助我了解科研的意义与价值，培养自己的科研品味。衷心感谢苏航副研究员对我科研工作的精心指导与关怀。苏老师为我博士期间的所有工作都倾注了心血，提供了建设性意见。在不断的交流中，苏老师帮助我明确了研究方向，规划了科研道路。非常感谢实验室全体老师的辛勤付出，为同学们提供了自由的科研环境，促使我在博士期间不断进步。

感谢实验室的同学们，为我的科研生活提供了非常大的帮助。我曾与庞天宇、杨啸、邓志杰、廖方舟、程书宇、付祈安等同学深入合作，非常感谢这些同学对我科研工作的帮助。感谢全体同学的日常学术分享与讨论，营造出良好的科研氛围，使我受益颇多。

特别感谢我的爱人姚祎铭，在我低谷时给予我支持，在我迷茫时给予我鼓励，在我困难时给予我帮助，是她对家庭的付出与关心为我提供了依靠，帮助我不断向前。愿未来的道路我们携手前行。

感谢我的父母董克俭和徐莉的养育之恩，是他们的理解与支持让我义无反顾地在学术道路上进行探索。

最后，本人多次获得清华大学计算机系的奖学金，受到了微软亚洲研究院、百度、字节跳动等公司提供的资助，特此致谢。

声　明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签　名：_____ 日　期：_____

个人简历、在学期间完成的相关学术成果

个人简历

1995 年 2 月 2 日出生于河北省邯郸市。

2013 年 8 月考入清华大学计算机科学与技术系，2017 年 7 月本科毕业并获得工学学士学位。

2017 年 8 月免试进入清华大学计算机科学与技术系攻读工学博士学位至今。

在学期间完成的相关学术成果

学术论文（* 表示共同作者）：

[1] **Dong Yinpeng**, Liao Fangzhou, Pang Tianyu, et al. Boosting adversarial attacks with momentum[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**), 2018: 9185-9193. (CCF 推荐 A 类国际会议)

[2] **Dong Yinpeng**, Pang Tianyu, Su Hang, et al. Evading defenses to transferable adversarial examples by translation-invariant attacks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR**). 2019: 4312-4321. (CCF 推荐 A 类国际会议)

[3] **Dong Yinpeng**, Su Hang, Wu Baoyuan, et al. Efficient decision-based black-box adversarial attacks on face recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR**). 2019: 7714-7722. (CCF 推荐 A 类国际会议)

[4] **Dong Yinpeng***, Deng Zhijie*, Pang Tianyu, et al. Adversarial distributional training for robust deep learning[C]//Advances in Neural Information Processing Systems (**NeurIPS**). 2020: 8270-8283. (CCF 推荐 A 类国际会议)

[5] **董胤蓬**, 苏航, 朱军. 面向对抗样本的深度神经网络可解释性分析 [J]. 自动化学报, 2020, 45: 1-12. (CCF 推荐 A 类中文期刊)

[6] **Dong Yinpeng**, Fu Qi-An, Yang Xiao, et al. Benchmarking adversarial robustness on image classification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR**). 2020: 321-331. (CCF 推荐 A 类国际会议)

[7] Cheng Shuyu*, **Dong Yinpeng***, Pang Tianyu, et al. Improving black-box adversarial attacks with a transfer-based prior[C]//Advances in Neural Information Processing Systems (**NeurIPS**). 2019: 10934-10944. (CCF 推荐 A 类国际会议)

[8] **Dong Yinpeng**, Ni Renkun, Li Jianguo, et al. Stochastic quantization for learning

accurate low-bit deep neural networks[J]. International Journal of Computer Vision (**IJCV**), 2019, 127(11): 1629-1642. (CCF 推荐 A 类国际期刊)

[9] **Dong Yinpeng**, Yang Xiao, Deng Zhijie, et al. Black-box detection of backdoor attacks with limited information and data[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (**ICCV**). 2021: 16482-16491. (CCF 推荐 A 类国际会议)

[10] **Dong Yinpeng***, Cheng Shuyu*, Pang Tianyu, et al. Query-Efficient Black-box Adversarial Attacks Guided by a Transfer-based Prior[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence (**TPAMI**). 2021. (CCF 推荐 A 类国际期刊)

[11] Liao Fangzhou, Liang Ming, **Dong Yinpeng**, et al. Defense against adversarial attacks using high-level representation guided denoiser[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**). 2018: 1778-1787. (CCF 推荐 A 类国际会议)

[12] Pang Tianyu, Du Chao, **Dong Yinpeng**, et al. Towards robust detection of adversarial examples[C]//Advances in Neural Information Processing Systems (**NeurIPS**). 2018: 4584-4594. (CCF 推荐 A 类国际会议)

[13] Deng Zhijie, **Dong Yinpeng**, Zhang Shifeng, et al. Understanding and exploring the network with stochastic architectures[C]//Advances in Neural Information Processing Systems (**NeurIPS**). 2020: 14903-14914. (CCF 推荐 A 类国际会议)

[14] Pang Tianyu, Yang Xiao, **Dong Yinpeng**, et al. Boosting adversarial training with hypersphere embedding[C]//Advances in Neural Information Processing Systems (**NeurIPS**). 2020: 7779-7792. (CCF 推荐 A 类国际会议)

[15] Yang Xiao, **Dong Yinpeng**, Pang Tianyu, et al. Towards face encryption by generating adversarial identity masks[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (**ICCV**). 2021: 3897-3907. (CCF 推荐 A 类国际会议)

专利：

[16] 朱军, **董胤蓬**, 苏航, 等. 信息处理方法及装置、介质及计算设备: 中国, CN110020593B[P]. 2021-04-13.

学术竞赛获奖：

[17] **董胤蓬**, 廖方舟, 庞天宇, 等. NeurIPS 2017 对抗攻防竞赛无目标攻击、有目标攻击、对抗防御三个赛道冠军. 2017-12-09.

指导教师学术评语

深度学习在图像识别、自然语言处理等众多任务上取得了突出进展，但是，深度学习模型很容易受到对抗噪声的干扰，产生错误的预测。因此，如何理解和提升深度学习模型的对抗鲁棒性是当前的一个前沿课题。该论文针对深度学习模型的对抗鲁棒性，重点从对抗攻击、对抗防御以及鲁棒性评测等方面进行系统研究，深入理解和提升模型的鲁棒性。该论文选题具有重要的理论意义和广泛的应用价值。论文的主要创新成果如下：

1） 提出了一种动量迭代的攻击方法，通过在对抗样本迭代优化过程中引入动量项，大幅提高了黑盒迁移攻击成功率；并且进一步提出了一种平移不变对抗攻击方法，有效提高针对对抗防御模型的黑盒迁移攻击成功率；

2） 提出了一种进化攻击方法，通过在黑盒决策攻击中建模搜索方向的局部几何结构，降低搜索空间的维度，有效提升黑盒决策攻击的效率；

3） 提出了一种对抗分布训练方法，通过利用对抗分布刻画原始样本周围多样化的对抗样本，有效增强了深度学习模型在不同对抗攻击下的防御能力；

4） 构建了一个面向图像分类任务的对抗鲁棒性测评基准，并提出了鲁棒性曲线的评测指标，对多个典型对抗攻防算法进行了公平、全面的鲁棒性测评。

该论文由作者独立完成。论文工作表明作者具有本领域坚实宽广的基础理论和系统深入的专业知识，独立从事科研能力强。论文写作规范，结构合理，论述清晰，实验充分。

答辩委员会决议书

论文研究深度学习的对抗攻击与鲁棒性测评，选题具有重要的理论意义和应用价值。

论文的主要创新性成果如下：

1. 提出了动量迭代方法和平移不变对抗攻击方法，两种方法分别通过引入动量项以及对一组经过平移变换的图片生成对抗样本，大幅提高了黑盒迁移攻击成功率；
2. 提出了一种进化攻击方法，该方法通过建模搜索方向的几何结构，有效提升了黑盒决策攻击的效率，验证了典型人脸识别模型在此场景下鲁棒性的不足；
3. 提出了一种对抗分布训练方法，该方法利用对抗分布刻画原始样本周围多样化的对抗样本，有效增强了模型的鲁棒性，有助于提升深度学习模型的可解释性；
4. 面向图像分类任务构建了一套对抗鲁棒性测评基准，采用鲁棒性曲线针对多个典型的对抗攻防算法进行了鲁棒性测评，并且通过分析实验结果得出与对抗攻防算法有效性相关的结论。

论文工作表明，作者已掌握本学科领域坚实宽广的基础理论和系统深入的专门知识，独立从事科研工作的能力很强。论文写作规范，结构合理，叙述清楚，是一篇优秀的博士论文。答辩过程中表述流畅，回答问题正确。

答辩委员会经无记名投票表决，一致同意通过论文答辩，并建议授予董胤蓬工学博士学位。