计算机科学
COMPUTER SCIENCE

通过拉普拉斯平滑梯度提高对抗样本的可迁移性

李文婷, 肖蓉, 杨肖

引用本文

李文婷, 肖蓉, 杨肖. 通过拉普拉斯平滑梯度提高对抗样本的可迁移性[J]. 计算机科学, 2024, 51(6A): 230800025-6.

LI Wenting, XIAO Rong, YANG Xiao. Improving Transferability of Adversarial Samples Through Laplacian Smoothing Gradient [J]. Computer Science, 2024, 51(6A): 230800025-6.

相似文章推荐（请使用火狐或 IE 浏览器查看文章）

Similar articles recommended (Please use Firefox or IE to view the article)

具有对抗鲁棒性的人脸活体检测方法

Face Anti-spoofing Method with Adversarial Robustness

计算机科学, 2024, 51(6A): 230400022-7. https://doi.org/10.11896/jsjkx.230400022

基于DNN模型输出差异的测试输入优先级方法

Test Input Prioritization Approach Based on DNN Model Output Differences

计算机科学, 2024, 51(6A): 230600121-8. https://doi.org/10.11896/jsjkx.230600121

融合证据句子提取的文档级关系抽取

Document-level Relation Extraction Integrating Evidence Sentence Extraction

计算机科学, 2024, 51(6A): 230800081-6. https://doi.org/10.11896/jsjkx.230800081

基于对抗样本和自编码器的鲁棒异常检测

Robust Anomaly Detection Based on Adversarial Samples and AutoEncoder

计算机科学, 2024, 51(5): 363-373. https://doi.org/10.11896/jsjkx.230300153

基于特征拓扑融合的黑盒图对抗攻击

Black-box Graph Adversarial Attacks Based on Topology and Feature Fusion

计算机科学, 2024, 51(1): 355-362. https://doi.org/10.11896/jsjkx.230600127

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230800025

通过拉普拉斯平滑梯度提高对抗样本的可迁移性

李文婷　肖　蓉　杨　肖
湖北大学计算机与信息工程学院　武汉 430000
(1749115318@qq.com)

摘　要　深度神经网络因模型自身结构的脆弱性,容易受对抗样本的攻击。现有的对抗样本生成方法具有较高的白盒攻击率,但在攻击其他DNN模型时可转移性有限。为了提升黑盒迁移攻击成功率,提出了一种利用拉普拉斯平滑梯度的可迁移对抗攻击方法。该方法在基于梯度的黑盒迁移攻击方法上做了改进,先利用拉普拉斯平滑对输入图片的梯度进行平滑,将平滑后的梯度输入利用梯度攻击的攻击方法中继续用于计算,旨在提高对抗样本在不同模型之间的迁移能力。拉普拉斯平滑的优点在于它可以有效地降低噪声和异常值对数据的影响,从而提高数据的可靠性和稳定性。通过在多个模型上进行评估,该方法进一步提高了对抗样本的迁移成功率,最佳的可迁移成功率比基线攻击方法高出2%。结果表明,该方法对于增强对抗攻击算法的迁移性能具有重要意义,为进一步研究和应用提供了新的思路。

关键词:深度神经网络;对抗攻击;对抗样本;黑盒攻击;可迁移性
中图分类号　TP393.08

Improving Transferability of Adversarial Samples Through Laplacian Smoothing Gradient

LI Wenting,XIAO Rong and YANG Xiao
School of Computer and Information Engineering,Hubei University,Wuhan 430000,China

Abstract　Deep neural networks are vulnerable to adversarial sample attacks due to the fragility of the model structure. Existing adversarial sample generation methods have a high white box attack rate,but their transferability is limited when attacking other DNN models. In order to improve the success rate of black box migration attack,this paper proposes a migration counterattack method using Laplacian smooth gradient. This method is improved on the gradient-based black box migration attack method. Firstly,Laplacian smoothing is used to smooth the gradient of the input image,and the smoothed gradient is input into the attack method using gradient attack for further calculation,aiming to improve the migration ability of the adversary-sample between different models. The advantage of Laplacian smoothing is that it can effectively reduce the impact of noise and outliers on the data,thus improving the reliability and stability of the data. The approach does further improve the migration success of adversarial samples by evaluating them on multiple models,with the best migrable success rate 2%,higher than the baseline attack method. The results show that this method is of great significance to enhance the migration performance of adversarial attack algorithms,and provides a new idea for further research and application.

Keywords　Deep neural networks,Adversarial attack,Adversarial samples,Black-box attack,Transferability

1　引言

近年来,深度学习技术快速发展,在各个领域取得了显著的成果。深度神经网络在图像分类、汽车自动驾驶[1]和视频监控[2]等任务中取得了令人瞩目的成果。然而,由于深度神经网络结构具有脆弱性,因此其容易受到对抗样本的影响[3]。对抗样本是在干净的输入数据上添加微小的扰动,这些示例经过微小扰动后,与原始样本无法区分,但会导致模型预测出现错误。此外,对抗样本具有一个“有趣”的属性——可迁移性,即利用当前模型生成的对抗样本可以成功地欺骗其他未知模型[4]。因此,学习如何生成具有高度可转移性的对抗样本变得至关重要。

在过去的几年里,生成对抗样本的方法得到了很多研究。例如局部时间序列黑盒对抗攻击[5]、多目标类别对抗样本生成算法[6]、快速梯度符号法[7]及其变体(包括迭代快速梯度符号法[8]、动量迭代快速梯度符号法[9]、Nesterov迭代快速梯度符号法[10]、基于方差调整动量的迭代方法[11]和空间动量迭代方法[12])。在这些攻击方法中,除了局部时间序列黑盒对抗攻击和多目标类别对抗样本生成算法外,其余均是基于梯度的攻击方法,此类攻击方法中梯度的重要性不言而喻。本

基金项目:基于人工智能的红外目标探测识别技术研究与应用(2022KZ00125);成果转化视角下的光电类国际专利数据聚类分析(E1KF291005)
This work was supported by the Research and Application of Infrared Target Detection and Recognition Technology based on Artificial Intelligence(2022KZ00125) and Cluster Analysis of Optoelectronic International Patent Data from The Perspective of Achievement Transformation (E1KF291005).
通信作者:肖蓉(x_rong@whu.edu.cn)

文提出的拉普拉斯平滑方法利用拉普拉斯平滑基于梯度攻击中的梯度信息，使梯度数据更加稳定，从而增强对抗样本的可迁移性。

本研究的目标是提升对抗样本在不同模型之间的迁移能力。该算法利用拉普拉斯平滑梯度信息优化生成对抗样本的过程，增强对抗样本的攻击性，最终通过全面的实验评估来验证算法的性能和有效性。本研究的意义在于增强对抗攻击算法的迁移性能。通过提升对抗样本在不同模型之间的迁移能力，可以更好地理解和应对对抗攻击的挑战，为深度学习模型的安全性提供更可靠的保护。此外，本研究也为进一步探索对抗攻击算法的改进和扩展方向提供了新的思路。主要贡献总结如下：

1)引入了一类基于拉普拉斯的可迁移对抗攻击算法来提高对抗样本的可迁移性，该算法将平滑后的梯度输入利用梯度攻击的攻击方法中，以稳定梯度信息。

2)实验证明了使用拉普拉斯平滑后的梯度有助于增强对抗样本的可迁移性，为所有基于梯度的对抗攻击方法提供新思路。

在接下来的章节中，将详细介绍相关工作、算法设计、实验评估以及讨论与分析，以全面探究基于拉普拉斯的可迁移对抗攻击算法的潜力和应用前景。

2 相关工作

若输入 x 为良性图像，输出 y 为 x 对应的真标签，则 $f(x;\theta)$ 输出预测结果的带有参数 θ 的分类器。$J(x,y;\theta)$ 是分类器 f 的损失函数，而攻击者非目标攻击的目标是找到一个对抗样本 x^{adv}，使其满足 $\|x-x^{adv}\|_p<\epsilon$，但会误导模型使其预测 $f(x;\theta)\neq f(x^{adv};\theta)$。$p$ 表示 p-范数距离，ϵ 是对抗性扰动的大小。

通过攻击先验条件可知，黑盒攻击比白盒攻击在实际场景中更为实用。黑盒攻击包括基于查询的攻击[10]和基于迁移的攻击[13]。基于查询的攻击方法通常需要大量的查询次数，这在实际应用中并不可行。相比之下，基于迁移的攻击更实用且得到了广泛研究，它利用源模型的白盒攻击来生成对抗样本，从而欺骗目标模型。本文的重点在于改进基于梯度的黑盒迁移攻击方法，以研究更具攻击性的对抗样本，从而提升深度模型的安全性和鲁棒性[14]。Zhao 等[15]深度神经网络安全性进行了综述，包括对抗攻击与防御方法的研究现状。

基于梯度的黑盒迁移攻击方法有 FGSM[7]，I-FGSM[8]，MI-FGSM[9]，NI-FGSM[16]，VMI-FGSM[11] 和 SMI-FGSM[12] 等。FGSM[7] 通过最大化损失函数 J(x,y) 来找出相应的对抗样本，该方法在特定模型上取得了一定的成功，但在迁移到其他模型上时效果有限。I-FGSM 方法将噪声 ε 的上限分成多个小的步长 α，并逐步增加噪声[17]。MI-FGSM[9] 是一种基于动量的多次迭代快速梯度符号方法，其引入了动量项，使得噪声添加方向的调整更加平滑，但边际效应递减对迭代次数的影响仍然存在[17]。Lin 等[16]提出了 NI-FGSM 方法，将 Nesterov 加速梯度法应用于迭代攻击中，使其更容易摆脱局部极值差，从而达到攻击的效果[18]。基于方差调整动量的迭代方法(VMI-FGSM)在更新过程中进一步考虑前一次迭代的梯度方差，以调整当前的梯度方向，从而稳定优化过程。空间动量迭代方法通过考虑来自不同区域的上下文梯度信息，引入从时域到空间域的动量累积机制，实现了 FGSM 类中最高的迁移成功率。

基于输入变换的迁移攻击算法有 DIM[19]，TIM[13]，SIM[16] 等，DIM 多元输入法以一定的概率对输入图像进行随机的大小调整和填充[19]，然后将处理后的图像送入分类器中计算梯度，提高可转移性。TIM 通过一组图像计算梯度，尤其对于带有防御机制的黑箱模型来说，这种方法的效果非常好。SIM 引入了尺度不变性，并在输入图像上按 $1/2^i$ 缩放的一组图像上计算梯度，以增强生成的对抗示例的可转移性，其中 i 是超参数。在 FGSM 类方法的基础上，结合不同输入(DIM)、平移不变(TIM)和尺度不变(SIM)可以进一步提高攻击成功率。

3 本文方法与算法设计

本章将详细描述本文的攻击算法 L-NI-FGSM，表 1 中列出了所有字符的含义。3.1 节介绍了目标优化问题，3.2 节介绍了基于拉普拉斯的攻击算法 L-NI-FGSM，3.3 节给出了这两者的合成攻击算法 L-NI-FGSM 的具体步骤，3.4 节说明了本文方法与现有攻击的区别。

表 1 符号总结

Table 1 Symbol summary

符号	描述	符号	描述
L	拉普拉斯函数	∇	求导
f	目标分类器	J	模型的损失函数
θ	一个特定模型的一组参数	x^{adv}，x'	添加扰动后的图像
x	原始图像	y，y^{true}	真实标签
ϵ	扰动大小	p	表示 p-范数距离
α	每次迭代图像像素更新的幅值	x_t^*，x_t^{adv}	x 经过 t 次攻击算法处理后的对抗样本
g_t	累积梯度	μ	衰减因子

3.1 目标优化问题

可以把生成对抗样本的过程视为一个优化问题，通过攻击白盒模型来生成并优化对抗样本，类似于在训练过程中使用训练数据。对抗性样本可以被视为模型的训练参数，在测试阶段，可以将评估对抗样本的黑盒模型视为模型的测试数据。从优化的角度来看，对抗样本的可转移性与训练模型的泛化能力相似[20]。因此，可以将用于提高模型泛化的方法迁移到对抗示例的生成中，从而提高对抗示例的可转移性。

给定一个参数为 θ 的目标分类器 f 和一个良性图像 $x\in X$，其中，x 在 d 维中，X 表示所有原始图像，对抗性攻击的目的是寻找一个满足以下条件的对抗样本 $x^{adv}\in x$。

$$f(x;\theta)\neq f(x^{adv};\theta) \tag{1}$$

$$\text{s.t. } \|x-x^{adv}\|<\epsilon \tag{2}$$

对于白盒攻击，可以把攻击看作一个优化问题，在 x 的邻域内搜索一个例子 x'，x' 为自变量，使目标分类器 f 的损失函数 J 最大化。

$$x^{adv}=\underset{x'}{\operatorname{argmax}}\ J(x',y;\theta)$$
$$\text{s.t. } \|x'-x\|_p\leq\epsilon \tag{3}$$

Lin 等[18]将对抗样本生成过程类比为标准的神经模型训练过程，将输入 x 视为待训练参数，将目标模型视为训练集。从这个角度来看，对抗样本的可转移性相当于正常训练模型的泛化。因此，现有的工作主要集中在更好地优化算法(如 MI-FGSM 和 NI-FGSM)[16]上。

3.2 基于拉普拉斯平滑的攻击算法

拉普拉斯平滑是一种常用的平滑技术,被广泛应用于统计学和机器学习中,用于对数据进行平滑处理,以降低噪声和增强信号。拉普拉斯平滑的核心思想是将数据中的每个值替换为其相对于整个数据集的加权平均值。具体来说,拉普拉斯平滑将每个数据点x_i替换为以下加权平均值:

$$\hat{g}_i=\frac{(n-1)g_i+\alpha\mu}{n+\alpha-1} \tag{4}$$

其中,n表示迭代的次数,迭代n次就会有n个梯度数据;μ表示所有梯度数据的平均值;g_i表示需要利用拉普拉斯进行平滑的参数,在本文中为输入图片的梯度数据;α是一个平滑参数,用于控制平滑的强度。当α越大时,平滑效果越强,数据越平滑;当α越小时,平滑效果越弱,数据越接近原始数据。

FGSM是最早的对抗攻击算法之一,FGSM类攻击算法(如FGSM[7],I-FGSM[8],MI-FGSM[9],NI-FGSM[16],VMI-FGSM[11]和SMI-FGSM[12]等)均是通过攻击方法在输入数据上利用梯度信息的符号添加扰动,从而改变模型的输出并生成对抗样本。梯度对于FGSM类攻击方法是非常重要的。而拉普拉斯平滑的优点在于,它可以有效地降低噪声和异常值对数据的影响,从而提高数据的可靠性和稳定性。另外,由于它计算简单,可以很容易地在大规模数据集上进行计算,因此被广泛应用于实际应用中。实验表明,拉普拉斯平滑梯度在提高对抗样本迁移能力方面具有潜力。通过拉普拉斯平滑对输入图片的梯度进行平滑,用平滑后的梯度参与扰动计算,这种改进使对抗样本在不同模型之间具有更好的可迁移性,从而增强了对抗样本在不同模型之间的迁移能力。

3.3 拉普拉斯平滑的梯度位置

根据拉普拉斯平滑的梯度位置不同,将梯度分为大梯度和小梯度。I-FGSM是FGSM的迭代版且它们的梯度位置固定,平滑位置也固定。将MI-FGSM和NI-FGSM中的动量项称为小梯度,将最终用于更新对抗样本的梯度,也就是$sign()$中的梯度称为大梯度。例如MI-FGSM中式(5)和NI-FGSM中式(10)中L()中的梯度即为小梯度,MI-FGSM中式(7)和NI-FGSM中式(12)中g_{t+1}即为大梯度。

L-MI-FGSM:

$$L\left(\frac{\nabla x J(x_t^*,y)}{\|\nabla x J(x_t^*,y)\|_1}\right) \tag{5}$$

$$g_{t+1}=\mu g_t+\frac{\nabla x J(x_t^*,y)}{\|\nabla x J(x_t^*,y)\|_1} \tag{6}$$

$$L(g_{t+1}) \tag{7}$$

$$x_t^*=x_{t-1}+\alpha * sign(g_{t+1}) \tag{8}$$

L-NI-FGSM:

$$x_t^{nes}=x_t^{adv}+\alpha\cdot\mu\cdot g_t \tag{9}$$

$$L\left(\frac{\nabla x J(x_t^{nes},y^{true})}{\|\nabla x J(x_t^{nes},y^{true})\|_1}\right) \tag{10}$$

$$g_{t+1}=\mu\cdot g_t+\frac{\nabla x J(x_t^{nes},y^{true})}{\|\nabla x J(x_t^{nes},y^{true})\|_1} \tag{11}$$

$$L(g_{t+1}) \tag{12}$$

$$x_{t+1}^{adv}=Clip_x^{\varepsilon}\{x_t^{adv}+\alpha\cdot sign(g_{t+1})\} \tag{13}$$

3.4 攻击算法L-NI-FGSM的具体步骤

本文将拉普拉斯应用到基于梯度攻击的算法中,基于梯度攻击的算法中,梯度是至关重要的。考虑到这一点,本文使用拉普拉斯平滑梯度数据。简单来说,拉普拉斯有对数据进行平滑处理,以降低噪声和增强信号的作用,它可以应用在有梯度信息的任何位置,对梯度数据进行平滑,然后用拉普拉斯平滑后的梯度代替原始梯度用于攻击算法中生成对抗样本。L-NI-SI-FGSM算法的具体步骤如算法1所示,第10步加上了拉普拉斯平滑公式,用于平滑小梯度信息。它同样也可以加在第12步的位置,第4章的对比实验中L-MI-FGSM和L-MI-FGSM攻击成功率的数据就是基于将拉普拉斯加在第12步的位置。根据梯度加的不同位置,在第4章中也给出了相应的攻击效果实验。输入x表示图像,y表示图像x的标签,损失函数为J,分类器F,微扰大小ε,最大迭代次数T,尺度复制数m和衰减因子μ。

算法1 L-NI-SI-FGSM

输入:$(x,y,J,F,\varepsilon,T,m,\mu)$

输出:对抗样本x^{adv}

1. $\alpha=\varepsilon/T$
2. $g_0=0;x_0^{adv}=x$
3. for t=0 to T−1 do
4. g=0
5. Get x_t^{nes} by Eq.(9) ▷3.3节中公式(9)在先前累积梯度的方向上跳跃
6. for i=0 to m−1 do ▷对输入图像的缩放副本求和梯度
7. 通过$\nabla_x J(S_i(x_t^{nes}),y^{true})$得到梯度 ▷$S_i$引入了尺度不变性,在输入图像上按$1/2^i$缩放图像,优化扰动
8. 将梯度求和为$g=g+\nabla_x J(S_i(x_t^{nes}),y^{true})$
9. 得到平均梯度为$g=\frac{1}{m}\cdot g$
10. L(g) ▷使用拉普拉斯平滑平均梯度
11. 将g_{t+1}更新为$g_{t+1}=\mu\cdot g_t+\frac{g}{\|g\|_1}$
12. $L(g_{t+1})$ ▷使用拉普拉斯平滑最终梯度
13. 用Eq.(13)更新x_{t+1}^{adv} ▷3.3节中公式(13)
14. return $x^{adv}=x_T^{adv}$

3.5 与现有攻击的区别

L-NI-FGSM攻击方法是基于FGSM的梯度攻击方法,NI-FGSM利用历史梯度稳定方向,用拉普拉斯平滑梯度。MI-FGSM使用先前的梯度更新当前的梯度。NI-FGSM通过Nesterov加速梯度对其进行改进,VMI-FGSM通过加入各种噪声来考虑梯度方差对其进行增强。这两种方法两次使用了之前的梯度,其作用不同。SMI-FGSM考虑了空间域信息,TI-FGSM是利用简单的预定义卷积核在空间域平滑未平移图像的梯度,其性能有限。通过结合时空动量,SM^{2}I-FGSM[11]不仅考虑了梯度的时间相关性,还考虑了空间相关性,可以全面稳定更新方向,获得更好的性能。以上方法均用到了输入样本的梯度信息,都是基于梯度的黑盒迁移攻击方法。本文方法与以上对抗攻击方法并不对立,而是在以上方法的基础上对梯度稍作调整以达到更好的攻击效果。

4 实验与评估

4.1 实验环境

在武汉大学的超算资源的分区GPU上完成实验,此GPU分区有125台V100服务器,其中每台节点的硬件配置及软件环境如表2所列。

表 2 实验环境

Table 2 Experimental environment

硬件配置	√	双路 Intel Xeon E5-2640v4 2.4GHz 处理器共 20 核
	√	128GB ECC 2400MHz DDR4 内存卡
	√	Nvidia Tesla V100 16GB NVLink
	√	Intel OPA 100Gbps 互联
软件环境	√	Linux Ubuntu 4.18.0
	√	Python 3.6
	√	TensorFlow 1.12.0

4.2 实验设计和数据集选择

4.2.1 数据集

从 ILSVRC 2012 验证集[21]中随机选取 1000 张属于 1000个类别的干净图像，这些图像几乎被所有测试模型正确分类，如文献[20]所示，并在表 3 中列出了完整的超参数取值。

表 3 超参数设置

Table 3 Hyperparameters setting

超参数	含义	值
n	最大扰动	16
T	迭代次数	10
α	步长	1.6
μ	衰减因子	1.0
p	对于 DIM 转换概率	0.5
eps	转化值	0.125(2 * n/255.0)
W	对于 TIM 高斯核大小	7×7
m	对于 SIM 缩放值$1/2^i$	i=0,1,2,3,4
alpha	Laplace 平滑参数 eps/i	i=1,10,30,50,130,150,200

4.2.2 分类模型

本文选用了 4 个正常训练的网络(包括 Inception-v3 (Inc-v3)[22], Inception-v4 (Inc-v4)[23], Inception-Resnet-v2 (InRes-v2)[23]和 Resnet-v2-101(Res-101)[24])和两个对抗训练的模型，即 IncRes-v2$_{ens}$[23] 和对抗训练的 Inception-v3 (Adv-Inc-v3)[18]。

4.2.3 基准方法

为更好地评估本文方法的有效性，实验以两种流行的基于动量的迭代对抗性攻击为基准，即 MI-FGSM[20] 和 NI-FGSM[16]。此外，将所提出的方法与 DIM[19]，TIM[13] 和 SIM[16]等多种输入变换相结合，记为 L-M(N)I-SI-FGSM 和 L-M(N)I-DTS-FGSM，进一步验证了方法的有效性。

4.3 对比实验

在单模型设置下使用 MI-FGSM 和 L-MI-FGSM 执行对抗性攻击，结果如表 4 所列。被攻击的 DNN 模型列在行上，第一列为生成对抗样本所使用的模型。显然，在攻击白盒模型时，L-MI-FGSM 和 MI-FGSM 一样强大，都有 100%的成功率。由表 4 可以看出，基于拉普拉斯的攻击显著提高了对抗性攻击的可转移性，并且模型是不可知的。例如，当使用 Inception-v3 作为白盒模型生成对抗样本时，MI-FGSM 在 Inception-resnet-v2，Resnet-v2-101 和 Adv-Inception-v3 上最高攻击成功率分别为 41.6%，35.6%和 17.9%，而 L-MI-FGSM 的攻击成功率分别为 43.2%，36.8%和 19.1%，可以看出本文方法领先基线方法，分别高出 1.6%，1.2%和 1.2%。当使用 Inception-Resnet-v2 作为白盒模型生成对抗样本时，L-MI-FGSM 在 Inception-v4 和 Adv-Inception-v3 上的最高攻击成功率分别为 52.7%和 24.3%，分别高出基线模型 1.8%和 1.3%，这表明，拉普拉斯对提高可迁移性的重要性。

表 4 MI-FGSM 和 L-MI-FGSM 的攻击成功率

Table 4 Attack success rate of MI-FGSM and L-MI-FGSM

(%)

Model	Attack		Inc-v3	IncRes-v2	Res-101	Adv-Inc-v3
Inc-v3	MI-FGSM		100*	41.6	35.6	17.9
	L-MI-FGSM alpha=	eps	100*	42.0	36.5	**19.1** ↑
		eps/10	100*	41.8	35.4	18.7
		eps/30	100*	40.2	**36.8** ↑	17.9
		eps/50	100*	42.9	35.5	18.3
		eps/130	100*	**43.2** ↑	35.2	18.1
		eps/150	100*	42.0	36.0	18.1
Model	Attack		Inc-v3	Inc-v4	IncRes-v2	Adv-Inc-v3
IncRes-v2	MI-FGSM		59.7	50.9	11.1	23
	L-MI-FGSM alpha=	eps	60.0	51.9	11.0	22.9
		eps/10	60.0	**52.7** ↑	**11.6** ↑	22.8
		eps/30	60.2	51.0	11.3	**24.3** ↑
		eps/50	59.7	51.4	11.3	23.3
		eps/130	59.2	52.3	11.5	23.4
		eps/150	**60.6** ↑	51.9	11.1	24.0

注：对抗性样本分别是在 Inc-v3 和 IncRes-v2 上制作的，* 表示攻击的是白盒攻击。最好的结果用粗体标出。

同样，我们在 NI-FGSM 上做了实验以验证本文方法的有效性，结果如表 5 所列。当使用 Inception-v3 作为白盒模型生成对抗样本时，L-NI-FGSM 和 NI-FGSM 一样强大，都有 100%的成功率。L-NI-FGSM 在 Inception-Resnet-v2 和 ens-Inception-Resnet-v2 模型上攻击成功率较 NI-FGSM 分别提升 1.6%和 1%。当使用 Incpeption-Resnet-v2 作为白盒模型生成对抗样本时，L-NI-FGSM 在 Inception-v3 和 Resnet-v2-101 上的最高攻击成功率分别为 64.1%和 47.2%，分别高出基线模型 2%和 1.5%。

表 5 NI-FGSM 和 L-NI-FGSM 的攻击成功率

Table 5 Attack success rate of NI-FGSM and L-NI-FGSM

(%)

Model	Attack		Inc-v3	IncRes-v2	IncRes-v2$_{ens}$	Adv-Inc-v3
Inc-v3	NI-FGSM		100*	48.5	5.8	20.3
	L-NI-FGSM alpha=	eps/10	100*	49.3	5.9	19.9
		eps/30	100*	50.0	6.1	19.2
		eps/50	100*	48.9	5.8	20.4
		eps/130	100*	49.3	6.8	20.1
		eps/150	100*	**50.1** ↑	**6.8** ↑	**20.4** ↑
Model	Attack		Inc-v3	Inc-v4	IncRes-v2	Res-101
IncRes-v2	NI-FGSM		62.1	54.5	98.9*	45.7
	L-NI-FGSM alpha=	eps	62.9	54.5	98.8*	46.4
		eps/10	62.3	**55.2** ↑	98.8*	45.8
		eps/50	6.03	55.2	99.0*	45.4
		eps/130	63.2	54.6	**99.1*** ↑	**47.2** ↑
		eps/150	63.4	54.8	98.8*	46.0
		eps/200	**64.1** ↑	54.9	98.8*	45.1

注：对抗性样本分别是在 Inc-v3 和 IncRes-v2 上制作的，* 表示攻击的是白盒攻击。最好的结果用粗体标出。

(1)拉普拉斯平滑位置的影响

在实验中，根据拉普拉斯平滑位置的不同会产生不同的效果，样本的迁移性效果也不同。取平滑参数 α=eps/130，如表 6 所列，无论是平滑哪个梯度，本文方法均比较稳定且有提升，其中 L-NI-DTS 平滑大梯度在 Ens-Inception-Resnet-v2 模型上攻击成功率较 NI-DTS 提升了 1.2%，平滑小梯度在 Resnet-v2-101 和 Ens-Inception-Resnet-v2 模型上较 NI-DTS 攻击成功率分别提升了 1%和1.9%。

表 6　NI-FGSM-DTS 和 L-NI-FGSM-DTS 的攻击成功率

Table 6　Attack success rate of NI-FGSM-DTS and L-NI-FGSM-DTS

(%)

Model	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	IncRes-v2$_{ens}$
Inc-v4	NI-DTS	88.1	99.3*	83.7	77	49.4
	+L(小)	88.1	**99.6*** ↑	**84** ↑	**78** ↑	**51.3** ↑
	+L(大)	**88.6** ↑	99.5*	83.6	77	50.6

注:对抗性样本是在 Inc-v4 上制作的,* 表示攻击的是白盒攻击。最好的结果用粗体标出。

(2)与其他方法结合

在 NI-FGSM 的基础上,加上尺度不变(SI)可以提高攻击成功率。我们将拉普拉斯与 SI-NI-FGSM 结合为 L-SI-NI-FGSM,如表 7 所列。L-SI-NI-FGSM 在 inception-v4 和 resnet-v2-101 模型上最高攻击成功率为 78.1%和 68.2%,与基线方法 SI-NI-FGSM 相比分别提升了 1.2%和 1.6%,这是一个显著的改进,表明本文方法具有良好的可扩展性,可以与现有方法相结合,进一步提高黑盒迁移攻击的成功率。同时,在 NI-FGSM 的基础上,不同输入(DI)、平移不变(TI)和尺度不变(SI)可以进一步提高攻击成功率。研究表明,两者的结合(本文称之为 DTS)可以帮助基于梯度的攻击实现很大的可转移性[16]。本文将 DTS 与 NI-FGSM 和 L-NI-FGSM 结合为 NI-FGSM-DTS 和 L-NI-FGSM-DTS,结果如表 6 所列。从表中可以看出,L-NI-FGSM 在 Resnet-v2-101 和 Ens-Inception-Resnet-v2 模型上最高攻击成功率为 78%和 51.3%。与基线方法 NI-FGSM-DTS 的可迁移成功率相比提升了 1%和 1.9%,这表明了本文方法提升的稳定性。

表 7　SI-NI-FGSM 和 L-SI-NI-FGSM 的攻击成功率

Table 7　Attack success rate of SI-NI-FGSM and L-SI-NI-FGSM

(%)

Attack		Inc-v4	Res-101	IncRes-v2$_{ens}$
NI-FGSM		52.1	41.8	5.8
SI-NI-FGSM		76.9	66.6	15.9
L-SI-NI-FGSM alpha=	eps/130	77.6	67.3	15.5
	eps/150	**78.1** ↑	66.7	**16.2** ↑
	eps/200	77.6	**68.2** ↑	15.6

注:对抗性样本是在 Inc-v3 上制作的,最好的结果用粗体标出。

(3)拉普拉斯平滑参数的影响

从表 4－表 7 的实验结果上看,拉普拉斯平滑参数 eps/i 对结果的好坏有很大的影响,如表 3 所列,采用 7 种不同的 i 值(i=1,10,30,50,130,150,200)对攻击方法中的梯度信息按照从小到大的程度进行平滑时,本文方法与基线模型相比表现出了显著的效果提升。这种多样性的取值允许本文从拉普拉斯平滑程度角度分析平滑梯度对攻击效果的影响,进一步挖掘平滑程度对不同模型的影响。本文选择这 7 种 i 值是因为它们在不同模型的攻击效果方面表现出色。平滑程度可与平滑位置结合起来研究,进一步改进基于拉普拉斯的可迁移对抗攻击算法,实现更优的攻击效果。

4.4　实验总结

本文以攻击模型的成功率为评估标准,对采用不同攻击方法生成的对抗样本进行了对比研究[14]。实验结果显示,拉普拉斯方法能够增强对抗样本的可迁移性,提高了黑盒攻击的成功率[14]。除此之外,表 5－表 7 实验结果表明:1)拉普拉斯平滑梯度的位置不同,会导致不同的攻击成功率,在实际应用中可根据攻击的目标模型选择最佳平滑位置,实现最佳攻击成功率;2)拉普拉斯方法对 FGSM 类基于梯度迭代的对抗攻击方法普遍适用且提升攻击效果稳定。

结束语　本研究提出了基于拉普拉斯的可迁移对抗攻击算法,并在深度学习模型上进行了实验评估。实验结果表明,基于拉普拉斯的可迁移对抗攻击算法在提高攻击成功率方面取得了显著效果。然而,我们也意识到该算法在某些情况下可能存在局限性,例如,拉普拉斯平滑梯度的位置及数量会直接影响攻击效果。因此,建议进一步研究和改进基于拉普拉斯的可迁移对抗攻击算法,探索该算法的适用范围和潜在局限性,以扩大其适用范围,提高鲁棒性。总之,本研究为深度学习模型中的对抗攻击和防御提供了一种新的方法。基于拉普拉斯的可迁移对抗攻击算法在提高攻击成功率和生成高质量对抗样本方面具有潜力,为进一步研究对抗攻击的抵抗能力和深度学习模型的鲁棒性提供了有价值的思路和启示。

参 考 文 献

[1] DUAN R,MAO X,QIN A K,et al. Adversarial laser beam:Effective physical-world attack to DNNs in a blink[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021:16062-16071.

[2] YAN H,WEI X:Efficient Sparse Attacks on Videos using Reinforcement Learning[C]// Proceedings of the 29th ACM International Conference on Multimedia. 2021:2326-2334.

[3] YE Q S,DAI X C. Current Status Analysis of Adversarial Sample Generation Techniques for Attack Classifiers [J]. Computer Engineering and Applications,2020,56(5):34-42.

[4] JIN S,LI M H,DU Y. Loss smoothing based countersample attack algorithm[J/OL]. Journal of Beijing University of Aeronautics and Astronautics:1-10[2023-08-02].

[5] YANG W B,YUAN J D. Local time series black box counterattack attacks[J]. Computer Science,2022,49(10):285-290.

[6] LI J,GUOY M. Multi-Objective Class Adversarial Sample Generation Algorithm based on Generative Adversarial Networks [J]. Computer Science,2022,49(2):83-91.

[7] GOODFELLOW I J,SHLENS J,SZEGEDY C. Explaining and Harnessin adversarial examples[EB/OL]. [2020-06- 20]. https://arxiv. org/abs/1412. 6572.

[8] KURAKIN A,GOODFELLOW I,BENGIO S. Adversarial examples in the physical world[EB/OL]. [2020-06-20]. https://arxiv. org//abs/1607. 02533.

[9] DONG Y,LIAO F,PANG T,et al. Boosting adversarial attacks with momentum[C]// Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. LongBeach. USA: IEEE Press,2018:9185-9193.

[10] TU C C,TING P,CHEN P Y,et al. Autozoom:Autoencoder-based zeroth order optimization method for attacking black-box neural networks[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2019:742-749.

[11] WANG X,HE K. Enhancing the transferability of adversarial

attacks through variance tuning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021:1924-1933.

[12] WANG G, YAN H, WEI X. Enhancing transferability of adversarial examples with spatial momentum[C]// Chinese Conference on Pattern Recognition and Computer Vision (PRCV). Cham: Springer International Publishing, 2022:593-604.

[13] DONG Y P, PANG T Y, SU H, et al. Evading defenses to transferable adversarial examples by translation-invariant attacks [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019:4312-4321.

[14] BAI Z X, WANG H J, GUO K X. Adversarial sample generation method based on random transformation of image color[J]. Computer Science, 2023, 50(4):88-95.

[15] ZHAO H, CHANG Y K, WANG W J. A review of adversarial attack and defense methods of deep neural networks[J]. Computer Science, 2022, 49(S2):662-672.

[16] LIN J D, SONG C B, HE K, et al. Nesterov accelerated gradient and scale invariance for adversarial attacks[C]// International Conferenceon Learning Representations. 2020.

[17] CHEN X N, HU J M, ZHANG B J. Based on the model of black box method against the attacks start increasing mobility between[J]. Computer Engineering, 2021, 47(8):162-169.

[18] DING J, XU Z W. Anti-migration attacks based on Rectified Adam and color invariance[J]. Journal of software, 2022, 33(7): 2525-2537.

[19] XIE C H, ZHANG Z S, ZHOU Y Y, et al. Improving transferability of adversarial examples with input diversity[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019:2730-2739.

[20] DONG Y P, LIAO F Z, PANG T Y, et al. Boosting adversarial attacks with momentum[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:9185-9193.

[21] RUSSAKOVSKY O, DENG J, SU H, et al. Imagenet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3):211-25.

[22] 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.

[23] SZEGEDY C, IOFFE S, VANHOUCKE V, et al. Inception-v4, inception-resnet and the impact of residual connections on learning[C]// AAAI Conference on Artificial Intelligence. 2017: 4278-4284.

[24] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Patternrecognition. 2016:770-778.

LI Wenting, born in 1999, postgraduate. Her main research interests include deep learning and adversarial attacks.

XIAO Rong, born in 1980, Ph.D, lecturer. Her main research interests include industrial Internet and intelligent analysis, natural language processing.