基于 NadaMax 更新与动态正则化的对抗样本迁移性增强方法

宋亚飞，仇文博，王艺菲，冯存前
（空军工程大学防空反导学院，西安，710051）

摘要 针对深度学习模型中对抗样本迁移性和黑盒攻击能力不足的问题，研究设计了一种基于 NadaMax 优化器的迭代快速梯度方法（NM-FGSM）。该方法结合了 Nesterov 加速梯度和 Adamax 优化器的优势，通过自适应学习率和前瞻动量向量提高梯度更新精确度，并引入动态正则化增强问题凸性，优化算法稳定性和针对性。实验结果表明，NM-FGSM 在不同攻击策略下优于现有方法，尤其在先进防御场景中攻击成功率提高了 4%～8%。通过动态正则化的损失函数，对抗样本的跨模型迁移能力得到提升，进一步增强了黑盒攻击效果。最后，讨论了未来优化 NM-FGSM 算法和设计防御措施的研究方向，为深度学习模型的安全性研究提供了新的思路。

关键词 迁移性；黑盒攻击；NadaMax 优化器；动量；自适应学习率；动态正则化
DOI 10.3969/j.issn.2097-1915.2025.03.015
中图分类号 TP391.4 **文献标志码** A **文章编号** 2097-1915(2025)03-0119-09

A Method of Enhancing Adversarial Example Transferability Based on NadaMax Update and Dynamic Regularization

SONG Yafei, QIU Wenbo, WANG Yifei, FENG Cunqian
(Air Defense and Antimissile School, Air Force Engineering University, Xi'an 710051, China)

Abstract To address the problem of insufficient transferability of adversarial examples and inadequate black-box attack capabilities in deep learning models, this study designs an iterative fast gradient method based on the NadaMax optimizer (NM-FGSM). This method integrates the advantages of Nesterov Accelerated Gradient and the Adamax optimizer, improving the accuracy of gradient updates through adaptive learning rates and lookahead momentum vectors. Additionally, dynamic regularization is introduced to enhance the convexity of the problem, optimizing algorithm stability and specificity. The experimental results demonstrate that the NM-FGSM is prior to the existing methods under conditions of various attack strategies, particularly in advanced defense scenarios, attack success rate increases by 4%～8%. The dynamically regularized loss function enhances the cross-model transferability of adversarial examples, thereby further improving black-box attack effectiveness. Finally, points out the way forward for the NM-FGSM algorithm and defense measures, providing a new insight into the security research of deep learning

收稿日期：2024-10-25
基金项目：国家自然科学基金(62402521)
作者简介：宋亚飞(1988－)，男，河南汝州人，副教授，博士，研究方向为智能信息处理。E-mail:yafei_song@163.com
通信作者：王艺菲(1987－)，女，陕西西安人，讲师，研究方向为网络安全与信息处理。E-mail:1549579019@qq.com

引用格式：宋亚飞，仇文博，王艺菲，等. 基于 NadaMax 更新与动态正则化的对抗样本迁移性增强方法[J]. 空军工程大学学报，2025，26(3)：119-127. SONG Yafei, QIU Wenbo, WANG Yifei, et al. A Method of Enhancing Adversarial Example Transferability Based on NadaMax Update and Dynamic Regularization[J]. Journal of Air Force Engineering University, 2025, 26(3): 119-127.

models.

Key words transferability; black-box attack; NadaMax optimizer; momentum; adaptive learning rate; dynamic regularization

深度学习作为人工智能的核心技术，已广泛应用于图像分类、自然语言处理、自动驾驶、医学诊断及军事等关键领域。然而，深度神经网络（deep neural networks，DNNs）对对抗样本高度敏感，攻击者通过添加难以察觉的扰动即可误导模型分类，严重威胁其安全性。研究高效的对抗样本生成算法对提升模型鲁棒性和安全性至关重要。

对抗样本问题的存在，主要源于深度学习模型在高维度空间中学习复杂的决策边界，这些决策边界往往对微小的输入扰动高度敏感，微小扰动能够轻易改变模型的预测结果，从而引发严重的安全问题。此外，对抗样本还具备泛化特性，即它们能够跨模型迁移，即使目标模型与训练对抗样本的模型结构不同或训练数据不同，也可能被误导。这种泛化特性极大增加了对抗样本攻击的实用性和威胁性，使得攻击者无需访问目标模型即可实施攻击。这种泛化性又被称为迁移性，近年来，对抗样本的迁移性问题受到广泛关注，尤其在黑盒攻击场景中，迁移性决定了攻击的成功率和实用性。

为了进一步提升基于迁移的攻击效果，研究者们提出了多种策略。部分研究聚焦于优化梯度计算的算法，通过动量、偏差修正等方式来稳定更新方向；也有研究致力于通过输入变换来增强样本的多样性；基于集成学习的优势提出了同时攻击多个替代模型来生成对抗样本的方法。除了这些基础做法外，Goodfellow 等[1]提出使用基于雅各比矩阵的数据增强方法来训练替代模型，使其学习目标模型的决策边界，优化替代模型的训练过程；Chen 等[2]提出注意力攻击（attack on attention，AoA）方法，通过引入注意力损失函数来抑制正确类别的注意力热力图大小，从而提升对抗样本的迁移性；Wang 等[3]设计了附属网络来捕获图像的潜在空间信息，并结合边缘检测算法找出最小有效扰动区域；Liu 等[4]还从模型几何特性、梯度正交性等角度深入分析了对抗样本迁移性的内在机制。

与此对应，对抗防御主要关注如何保护模型免受对抗性攻击的影响。近年来，主要发展出对抗性训练[5]、输入预处理[6]、特征去噪[7]、认证防御[8]、模型集成[4]、梯度掩蔽[9]及防御蒸馏[10-11]等防御策略，对抗攻击与防御相互博弈，两者之间的平衡不断被打破与重建，共同发展。

面对对抗样本的威胁，想要深入挖掘防御技术，首先要从防御者的角度生成效果良好的对抗样本，用于对抗性训练以及攻防实践。然而对抗样本攻击研究仍存在一些问题。一方面，在黑盒情境下，大多数攻击的效果较差且生成效率低，严重影响在攻击和对抗训练中的使用。另一方面，基于迁移的攻击所产生的对抗样本通常容易过拟合训练模型，并陷入较差的局部极值，从而导致迁移性较弱。

为了更好地解决上述问题，本文提出了基于 NadaMax[12]的迭代快速梯度方法（nadamax-iterative fast gradient sigh method，NM-FGSM），通过其独特的学习率调整和动量累积机制，能够更精确地指导梯度更新方向，提升生成质量和效率；受到近端点方法（proximal point method，PPM）启发，设计了一种与当前迭代次数和变量更新量相关的二次正则化项融入目标函数，控制对抗样本更新空间，进一步提升对抗样本的泛化能力和黑盒攻击成功率。实验结果表明，NM-FGSM 具有更强的迁移性和黑盒攻击能力以及更高的生成效率，为解决当前对抗样本研究中的问题提供了新的有效途径。

1 黑盒迁移对抗样本研究

1.1 黑盒迁移攻击

对抗样本攻击通常根据攻击者所掌握的信息量被分类为白盒攻击与黑盒攻击。白盒攻击指攻击者在完全知道机器学习模型结构和参数的情况下，直接利用这些信息来生成对抗样本；黑盒攻击指攻击者不知道机器学习模型的具体结构和参数，但可以通过与模型的交互（如查询模型的输出）来生成对抗样本。

黑盒攻击进一步细分为基于迁移的攻击[1-2,13-23]、基于得分的攻击[24-25]和基于决策的攻击[26]。在现实应用中，攻击者面临的往往是黑盒环境，即无法深入了解模型内部。因此，尽管黑盒攻击的实施难度更高，但其实际应用价值却更为显著。基于得分和决策的攻击都需要对神经网络进行大量访问和查询，这在实际应用中更难实现。基于迁移的攻击方法因其独特的对抗迁移性而备受青睐，即在一个模型环境中生成的对抗样本往往对其他模型也具备对抗能力。

基于迁移的攻击分为 2 个步骤：首先，在替代模型下采用白盒攻击生成对抗样本；然后，将这些样本

转移到目标模型进行攻击。白盒攻击通常为基于梯度的攻击，因为它们相对高效且易于实现。

1.2　基于梯度的黑盒迁移攻击

基于梯度的黑盒迁移攻击近年来有很多研究角度，Chen 等[2]提出的 AoA 方法从损失函数角度来研究，Goodfellow 等[1]提出的线性反向传播（linear backpropagation，LinBP）从梯度传播角度出发。最经典的类别是基于迭代算法的迁移攻击，原因是对抗样本的生成通常依赖于对迭代过程中模型梯度方向的理解，优化梯度生成过程有利于接近模型决策边界，提高迁移性。快速梯度符号法[13]（fast gradient sign method，FGSM）是所有梯度迭代攻击的基础，整个过程只更新 1 次。迭代快速梯度符号法（iterative fast gradient sign method，I-FGSM）[14]将 FGSM 中的 1 步扰动计算细分为 t 步进行迭代，并通过裁剪操作将像素限制在有效区域内。动量迭代快速梯度符号法（momentum iterative fast gradient sign method，MI-FGSM）[13]在 I-FGSM 的基础上引入了动量的思想。基于方差调整的动量迭代快速梯度符号法（variance tuning momentum iterative fast gradient sign method，VMI-FGSM）[17]通过计算梯度方差，利用第 $t-1$ 次迭代的梯度方差来调整第 t 次迭代的对抗样本梯度，以稳定更新方向。涅斯捷罗夫迭代快速梯度符号法（Nesterov iterative fast gradient sign method，NI-FGSM）[27]将 Nesterov 加速梯度应用于梯度更新，通过前瞻梯度提高梯度更新的精确度。多样化输入迭代快速梯度符号法（diverse input iterative fast gradient sign method，DI^2-FGSM）[16]对输入图像应用随机且可微的变换，并以一定的概率 p 使用变换后的图像来最大化损失函数。基于迭代算法的黑盒迁移攻击发展如图 1 所示。

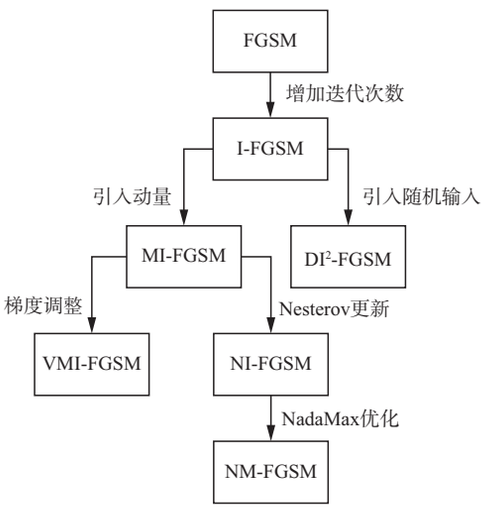

图 1　基于迭代算法的迁移攻击发展

Fig. 1　Development of transfer attacks based on iterative algorithms

由于本文提出的 NM-FGSM 追溯源头是从 MI-FGSM 引入动量这一核心做法改进的，因此在这里详细介绍 MI-FGSM 的算法流程。

1）参数初始化。在初始阶段需要设定一系列参数，包括损失函数 J、分类器 f、原始样本 $\boldsymbol{x}$、真实标签 y、扰动大小 $\boldsymbol{\delta}$、最大扰动 ε、迭代次数 T 和衰减因子 μ。

$J(f(\boldsymbol{x}),y)$：分类器 f 对样本 $\boldsymbol{x}$ 进行预测后，预测结果与真实标签 y 之间的损失；

$\boldsymbol{x}_0=\boldsymbol{x}$：初始对抗样本设置为原始样本；

$\boldsymbol{\delta}_0=0$：初始扰动设置为零向量；

其他参数如 ε、T、μ 根据具体问题设定。

2）计算梯度。在每次迭代中，需要计算当前对抗样本 $\boldsymbol{x}_t$ 在模型 f 上的梯度。梯度指示了损失函数相对于输入样本的变化率，是优化过程中调整对抗样本的关键。

$$\boldsymbol{g}_t=\nabla_x J(f(\boldsymbol{x}_t),y) \tag{1}$$

式中：$\boldsymbol{g}_t$ 为在第 t 次迭代时，损失函数 J 关于输入样本 $\boldsymbol{x}_t$ 的梯度。

3）梯度更新公式。通过引入动量项累积梯度信息，从而稳定梯度的更新方向，避免陷入局部最优解。动量项计算：

$$\boldsymbol{m}_t=\mu\boldsymbol{m}_{t-1}+\frac{\boldsymbol{g}_t}{\|\boldsymbol{g}_t\|_1} \tag{2}$$

式中：μ 为衰减因子，用于控制历史梯度的权重。

对抗扰动更新：

$$\boldsymbol{\delta}_{t+1}=\mathrm{Clip}_{\varepsilon}(\boldsymbol{\delta}_t+\alpha\,\mathrm{sign}(\boldsymbol{m}_t)) \tag{3}$$

式中：α 为步长；$\mathrm{Clip}_{\varepsilon}$ 函数确保扰动在$[-\varepsilon,\varepsilon]$范围内。

对抗样本更新：

$$\boldsymbol{x}_{t+1}=\boldsymbol{x}_t+\boldsymbol{\delta}_{t+1} \tag{4}$$

尽管 MI-FGSM 通过引入动量项在一定程度上提高了对抗样本的迁移性，但动量项的累积可能导致梯度方向过度偏移，使得对抗样本在某些情况下偏离最优解。其次，算法对初始参数的选择较为敏感，不恰当的参数设置可能导致算法性能下降。此外，MI-FGSM 在面临复杂决策边界的模型时，无法有效找到具有强迁移性的对抗样本。为了克服这些缺陷，本文提出了 NM-FGSM，通过结合 NadaMax 优化器的前瞻优势，引入融合动态正则化的目标函数提供可控空间，更加稳健地选择梯度方向，且算法中的自适应参数调整使得在不同初始参数设置下保持相对稳定的性能，使用二阶信息或更高级的梯度估计有效探索和利用模型的梯度信息。

2　NM-FGSM 算法设计

MI-FGSM 通过引入动量项，NI-FGSM 通过

NAG 算法引入前瞻性，较好地稳定了更新方向。动量和 NAG 算法都是梯度下降优化算法，能对深度学习模型的训练产生好的效果，在体现延伸到对抗样本生成中时，能够帮助对抗样本获得很好的迁移性。因此，能够假设其他梯度下降优化算法也可以应用到对抗攻击当中，提升迁移性。

本文考虑 NadaMax 应用于基于梯度的迭代攻击，优化梯度计算过程。其基础算法 Adam 具有较差的收敛性，Nadam 算法结合了 Nesterov 动量和 Adam 算法，其改进核心在于计算当前时刻的梯度时使用了“未来梯度”的预估，进一步提升了算法在复杂地形中的收敛能力，使得参数更新更加高效和准确。NadaMax 算法在 Nadam 的基础上引入了 AdaMax 的思想，即使用梯度的 L_∞ 范数替代二阶矩的平方根来进行缩放，更新参数。这种改变旨在保持自适应学习率特性的同时，通过更稳定的范数估计来避免极端值对学习过程的影响，有助于处理各种复杂的梯度分布，从而进一步提升算法的稳定性和泛化能力。因此，NadaMax 优化算法在生成对抗样本方面具有一定的潜力。具体流程如下：

1)参数初始化。开始阶段对一系列参数进行初始化，主要包括损失函数 J、分类器 f、原始样本 $\boldsymbol{x}$、真实标签 y、步长 α、一阶矩估计 $\boldsymbol{m}$、二阶矩估计 $\boldsymbol{u}$、衰减率 β_1 和 β_2、扰动大小 $\boldsymbol{\delta}$、最大扰动 ε 以及迭代次数 T。

$J(f(\boldsymbol{x}),y)$：分类器 f 对样本 $\boldsymbol{x}$ 进行预测后，预测结果与真实标签 y 之间的损失；

$\boldsymbol{x}_0=\boldsymbol{x}$：初始对抗样本设置为原始样本；

$\boldsymbol{\delta}_0=0$：初始扰动设置为零向量；

$\alpha=\varepsilon\times\sqrt{N}/T$：步长，控制每次迭代中对抗样本更新的幅度，其中，输入样本尺寸为 $N\times N$；

$\boldsymbol{m}_0=0$：一阶矩估计的初始值，通常设置为零向量；

$\boldsymbol{u}_0=0$：二阶矩估计的初始值，也设置为零向量或小的正数向量以防止除零错误；

其他参数如 ε、β_1、β_2、T 根据具体问题设定。

2)计算梯度。在每一次迭代中，需要计算当前对抗样本的梯度。这通常涉及将当前对抗样本输入到目标模型 f 中，并通过模型的前向传播和反向传播过程来计算梯度：

$$\boldsymbol{g}^{*}=\nabla_x J(f(\boldsymbol{x}_t),y) \tag{5}$$

式中：$\boldsymbol{g}^{*}$ 为梯度。

归一化梯度为：

$$\boldsymbol{g}_t=\frac{\boldsymbol{g}_t^{*}}{\|\boldsymbol{g}_t^{*}\|_1} \tag{6}$$

式中：$\boldsymbol{g}_t$ 为归一化的梯度；$\|\cdot\|_1$ 为 1－范数。

3)梯度更新公式。一阶矩估计的 Nesterov 更新：

$$\hat{\boldsymbol{g}}_t=\boldsymbol{g}_t+\beta_1\boldsymbol{m}_{t-1} \tag{7}$$

$$\boldsymbol{m}_t=\beta_1\boldsymbol{m}_{t-1}+(1-\beta_1)\hat{\boldsymbol{g}}_t \tag{8}$$

式中：$\hat{\boldsymbol{g}}_t$ 为前瞻梯度，根据当前动量方向“前瞻”一步，到达预估的下一位置，在预估位置上计算梯度；$\boldsymbol{m}_t$ 为基于前瞻梯度进行的 Nesterov 更新。

二阶矩估计为：

$$\boldsymbol{u}_t=\max(\beta_2\boldsymbol{u}_{t-1},\|\boldsymbol{g}_t\|_\infty) \tag{9}$$

使用梯度的 L_∞ 范数替代传统 Adam 算法中的二阶矩平方根估计，可以简化计算并保持数值稳定性。

动态学习率为：

$$\alpha_t=\alpha\frac{\sqrt{1-\beta_2^t}}{1-\beta_1^t} \tag{10}$$

确保学习率随着迭代次数的增加而逐渐减小，有助于模型在训练后期更加稳定地收敛。

计算扰动更新量为：

$$\boldsymbol{\Delta}_t=\alpha_t\frac{\hat{\boldsymbol{m}}^t}{\boldsymbol{u}_t+1\mathrm{e}-8} \tag{11}$$

式中：$\boldsymbol{\Delta}_t$ 为扰动更新量。

对抗扰动更新：

$$\boldsymbol{\delta}_{t+1}=\mathrm{Clip}_\varepsilon(\boldsymbol{\delta}_t+\boldsymbol{\Delta}_t) \tag{12}$$

式中：$\mathrm{Clip}_\varepsilon$ 函数确保扰动在$[-\varepsilon,\varepsilon]$范围内。

对抗样本更新：

$$\boldsymbol{x}_{t+1}=\boldsymbol{x}_t+\boldsymbol{\delta}_{t+1} \tag{13}$$

式(7)～式(13)体现了 NadaMax 的高效梯度更新机制。在该过程中，一阶矩 Nesterov 更新提高了算法逃离局部最优解的能力，二阶矩的更新使用 L_∞ 范数替代了 FGSM 方法中 sign 函数的规范作用，使得数值稳定。整体过程相较于 Nadam 的更新，还取消了数值矫正过程，因此能够得到更好的样本质量和更快的生成速度。

3 融入动态正则化的目标函数

为了进一步优化对抗样本的生成过程，提高算法的稳定性和收敛速度，本文在算法设计上做了进一步探索，引入动态正则化概念，通过对目标函数进行改进，以期进一步提升 NM-FGSM 性能。

在一系列基于梯度的黑盒迁移攻击算法中，损失函数多为交叉熵损失，但面临过拟合风险，且当训练数据分布不均匀或初始化参数不当时，会导致训练过程不稳定，收敛速度缓慢或陷入局部最优解。

在优化过程中，PPM 通过在每一步迭代中求解

1 个简化的子问题来逐步逼近原问题的解。这个子问题通常是在当前迭代点的 1 个邻域内定义的，并且包含了某种形式的正则化项，通常是与原问题相关的某种形式的距离函数，以确保解的稳定性或平滑性。PPM 的核心思想是利用近端算子来限制搜索空间，使得每一步迭代都更加可靠和可控。PPM 的数学表达为：对于 1 个给定的优化问题，例如 minimize $F(x)$，其中，$F(x)$是 1 个可能不平滑或不可微的函数。PPM 通过以下迭代步骤求解：

1)初始化，选择 1 个初始点 x_0。

2)迭代更新，对于 $k=1,2,\cdots$，直到收敛。迭代过程为：

$$x_{k+1}=\operatorname{argmin}_x\{F(x)+\frac{1}{2\lambda_k}\|x-x_k\|_2^2\} \tag{14}$$

式中：$\lambda_k>0$ 为 1 个序列，控制正则化项的强度。

受 PPM 的逐步逼近和优化思路启发，本文提出了动态正则化项，可以被视为一种特殊的近端算子，通过向损失函数中添加与当前迭代次数和变量更新量相关的二次正则化项，使问题表现出更易处理的凸性。凸性在理论上虽然不利于对抗样本理论需要的决策边界，但如果找到一个平衡点，是有助于控制对抗扰动的规模和方向的。同时，正则化项使得算法能够更好地适应不同阶段优化需求，从而在初期允许更大的探索空间，而在后期则集中搜索最优解，有助于优化算法的稳定性和加速收敛。将这个思想应用到对抗样本生成过程中，可以设计一种动态调整对抗样本搜索策略的方法，更有效地找到能够误导模型的对抗样本。

$$L_{\text{adv}}(\boldsymbol{x},y,\boldsymbol{\delta},k)=L(\boldsymbol{x}+\boldsymbol{\delta},y)+\frac{\lambda_k}{2}\|\boldsymbol{\delta}_k-\boldsymbol{\delta}_{k-1}\|^2 \tag{15}$$

式中：$\boldsymbol{x}$ 为原始样本；y 为原始样本的真实标签；$\boldsymbol{\delta}_k$ 为第 k 次迭代中的对抗扰动(在第 1 次迭代时可以是零向量或小的随机向量)；$L(\boldsymbol{x}+\boldsymbol{\delta},y)$为将对抗扰动添加到原始样本后，模型对修改后样本的预测损失；λ_k 为与迭代次数 k 相关的动态正则化强化系数，用于平衡原始损失和正则化项的重要性。

与大多优化算法中的应用类似，考虑使用逐渐减小的 λ_k 来鼓励对抗扰动的连续变化，并在搜索后期更加专注于能够最大化模型损失的对抗样本。本文采用的策略为：

$$\lambda_k=\frac{\lambda_0}{\sqrt{k+1}} \tag{16}$$

式中：λ_0 为超参数，用于控制正则化项的初始强度。

融入动态正则化项后，NM-FGSM 相对于传统的 FGSM 及其变种在复杂度上有所增加，但这一增加是合理的。NM-FGSM 的总计算复杂度可以表示为 $O(T(G+U+R))$，其中，G 为梯度计算的复杂度，U 为 NadaMax 优化器更新的复杂度，R 为动态正则化项计算的复杂度，T 为迭代次数。

动态正则化项的引入需要在每次迭代中计算正则化损失，这通常涉及向量的点积和二次方运算，其复杂度与输入样本的维度成正比，都为 $O(n)$，是比较简单的操作，且由于点积和二次方运算都是对向量中每个元素独立进行的操作，具有良好的可并行性。而梯度计算通常涉及整个神经网络的前向传播和反向传播，其复杂度远高于简单的向量运算。因此，即使动态正则化项引入了一些额外的计算量，但在整个算法的计算量中所占比例是很小的。

动态正则化项为对抗样本提供了稳定性和方向性控制机制，通过动态调整正则化项，在优化中平衡了探索与利用，使得对抗样本保持隐蔽性的同时，增强了对不同模型结构的适应能力。

4　实验与分析

4.1　数据集与实验设置

数据集：使用来自 ImageNet 数据集的 1 000 张图像，这些图像是从不同类别中随机选择的。几乎所有图像都能被此次测试的网络正确分类。在使用前，图像大小被预处理为 299×299×3。

模型：在 7 个网络上测试本文所提出的攻击方法，包括正常训练的模型 Inception-v3 (Inc-v3)、Inception-v4 (Inc-v4)、Inception-Resnet-v2 (IncRes-v2)、Resnet-v2-101 (Res-101)，以及对抗性训练的模型 Inc-v3ens3、Inc-v3ens4、IncRes-v2ens。

超参数：根据文献[13]和文献[27]中的实验经验，设置每个像素的最大扰动 $\varepsilon=16$，总迭代次数 $T=10$，步长 $\alpha=2\times\varepsilon/255$，NadaMax 衰减因子 $\beta_1=0.99$ 和 $\beta_2=0.999$ 根据初步实验调整得出，以在攻击效果和计算成本之间取得平衡。正则项中的强度因子 $\lambda_0=0.001$ 通过交叉验证选择，以在正则化效果和模型性能之间找到最佳折中。输入图像的维度 N 设置为 299×299×3。

4.2　攻击单个模型实验

本节测试并比较了 FGSM、I-FGSM、MI-FGSM、NM-FGSM 在 7 个模型上的攻击性能，实验结果如表 1 所示。其中，* 表示白盒攻击，加粗数据为 4 种算法在测试同一模型时的最高成功率。该实验对源模型(表格第 1 列)使用 4 种算法(表格第 2 列)生成对抗样本，用于攻击目标模型(表头行)。当源模型与目标模型不同时，属于黑盒环境，用于测试对抗样

本的可转移性。攻击成功率作为评估对抗样本可转移性的指标，表示能够导致测试模型误分类的对抗图像数量占总生成对抗图像数量的百分比，攻击成功率越高，算法性能越好。

表 1 各算法对单个模型攻击成功率对比

Tab. 1 Comparison of attack success rates against single models by various algorithms

模型	算法	成功率/%							平均单个样本产生时间/s
		Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3ens3	Inc-v3ens4	IncRes-v2ens	
Inc-v3	FGSM	65.6 *	27.3	24.3	24.6	10.0	9.8	4.5	0.3
	I-FGSM	96.5 *	22.1	18.9	15.4	5.6	6.7	3.5	2.2
	MI-FGSM	**99.2** *	37.8	36.8	34.0	11.5	11.2	5.7	3.5
	NM-FGSM	98.8 *	**49.2**	**45.2**	**39.1**	**15.9**	**15.4**	**8.2**	1.2
Inc-v4	FGSM	26.5	54.7 *	23.5	23.6	9.7	9.5	5.5	0.4
	I-FGSM	31.8	96.8 *	20.8	22.5	6.8	6.2	4.3	4.3
	MI-FGSM	54.4	97.2 *	45.2	42.3	16.3	15.2	7.8	6.5
	NM-FGSM	**60.4**	**99.5** *	**49.1**	**45.3**	**20.4**	**16.9**	**10.2**	3.2
IncRes-v2	FGSM	27.5	21.2	43.3 *	24.6	9.8	9.6	5.6	0.5
	I-FGSM	32.7	26.3	97.6 *	21.2	7.5	6.4	4.8	4.5
	MI-FGSM	**60.5**	52.5	**98.0** *	45.3	20.8	15.6	10.3	7.7
	NM-FGSM	60.3	**55.5**	97.4 *	**46.4**	**24.8**	**19.6**	**15.1**	3.4
Res-101	FGSM	35.7	30.8	30.2	80.3 *	15.7	14.5	7.5	0.5
	I-FGSM	32.4	25.5	24.2	**99.8** *	9.6	8.9	6.4	5.2
	MI-FGSM	58.6	51.8	49.3	99.3 *	24.6	22.3	13.5	8.2
	NM-FGSM	**60.2**	**56.7**	**51.4**	99.7 *	**28.4**	**27.2**	**17.7**	4.0

根据实验结果对比，在黑盒情况下，NM-FGSM 在这 4 种攻击方法中成功率最高。例如，当在 Inc-v3 上生成的对抗样本转移到 Inc-v4 和 IncRes-v2 时，NM-FGSM 的成功率分别为 49.2% 和 45.2%，而 MI-FGSM 的成功率分别为 37.8% 和 36.8%，这充分证明了 NM-FGSM 在提高攻击可转移性方面的优势。

此外，从表中可以观察到，除了单步攻击 FGSM 外，其他 3 种迭代攻击在白盒设置下的成功率几乎为 100%，这表明迭代攻击相对于单步攻击上具有显著优势。同时，在 6 个黑盒测试中，I-FGSM 的表现最差。因此，在后续实验中，本文方法只与基于动量的攻击进行比较，主要为 MI-FGSM、NI-FGSM 和 NM-FGSM。

4.3 攻击模型集合实验

本节分别使用 MI-FGSM、NI-FGSM 和 NM-FGSM 对多个网络模型同时实施集合攻击，攻击 Inc-v3、Inc-v4、IncRes-v2、Res-101 的集合，每个模型被赋予相同的集合权重，即 $\omega_k=1/4$。

表 2 展示了在攻击模型集合后，各种基于动量的迭代攻击方法成功率。加粗数据表示同一类型攻击的最高成功率。

表 2 各算法对模型集合攻击成功率对比

Tab. 2 Comparison of attack success rates against model ensembles by various algorithms

%

算法	Inc-v3 *	Inc-v4 *	IncRes-v2 *	Res-101 *	Inc-v3ens3	Inc-v3ens4	IncRes-v2ens
MI-FGSM	**99.9** *	98.3 *	96.2 *	**99.8** *	38.5	34.1	22.6
NI-FGSM	99.8 *	**99.8** *	98.7 *	99.7 *	38.2	23.4	**43.1**
NM-FGSM	99.7 *	99.4 *	**99.5** *	99.6 *	**57.0**	**50.7**	35.1

从表 2 可观察到，NM-FGSM 在攻击 3 个对抗训练模型时的平均成功率比 MI-FGSM 高出 15% 以上，比 NI-FGSM 高 12% 以上，进一步证明了 NM-FGSM 的优势。同时，NM-FGSM 仍然保持与其他 2 种基于动量的迭代攻击相似的白盒成功率。

4.4 攻击先进防御方法实验

除了上述使用不同攻击策略测试算法的性能

外，本节通过 MI-FGSM、NI-FGSM 和 NM-FGSM 在其他先进防御机制上的有效性，来说明 NM-FGSM 攻击和迁移性能的全面提升，这些方法包括：NIPS 竞赛中排名前 3 名的防御解决方案高级表示引导去噪器（HGD）、随机调整大小和填充（R&P）、编号为 NIPS-r3 的防御方案，以及 3 种近年提出的防御方法特征蒸馏（FD）、通过图像压缩模型净化扰动（ComDefend）和随机平滑（RS）。实验结果如表 3 所示。

表 3　各算法对先进防御方案攻击成功率对比

Tab. 3　Comparison of attack success rates against advanced defense schemes by various algorithms

%

算法	HGD	R&P	NIPS-r3	FD	ComDefend	RS	Average
MI-FGSM	36.9	29.3	40.8	51.6	47.5	27.1	38.9
NI-FGSM	38.2	30.5	42.3	48.7	47.0	**29.3**	39.3
NM-FGSM	**42.8**	**34.4**	**45.1**	**59.6**	**49.5**	26.7	**43.0**

由表 3 可以观察到，NM-FGSM 在攻击先进防御方法中，在对除了 RS 方法之外的其他防御方法均有较高的攻击成功率，尤其在面对 FD 方法时，比另外 2 种算法中较高的 MI-FGSM 还要高出 8%的攻击成功率。算法在 RS 方法上表现不佳的原因与 RS 方法性质相关，RS 在高斯噪声扰动下对分类函数进行了平滑处理，能够平滑输入空间的扰动，对小的扰动具有鲁棒性，而基于梯度的扰动通常是沿梯度方向的微小修改。

4.5　算法实用性实验

本节研究迭代次数对算法攻击性能的影响。在本实验中，迭代次数以 2 为单位，从 2～16 分布。白盒模型为 Inc-v3 模型，目标模型为 Inc-v3ens3、Inc-v3ens4 和 IncRes-v2ens 模型。图 2、图 3、图 4 显示了 3 种动量迭代攻击（MI-FGSM、NI-FGSM、NM-FGSM）不同迭代次数对攻击效果的影响。

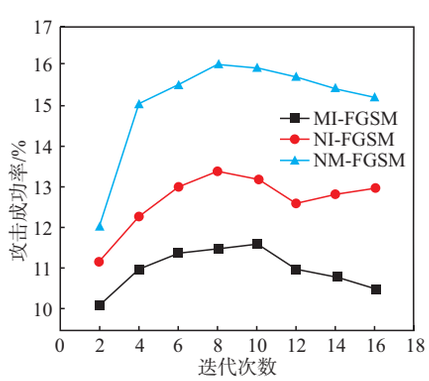

图 2　各算法攻击 Inc-v3ens3 成功率随迭代次数变化关系

Fig. 2　Relationship between attack success rate and iteration number for various algorithms targeting Inc-v3ens3

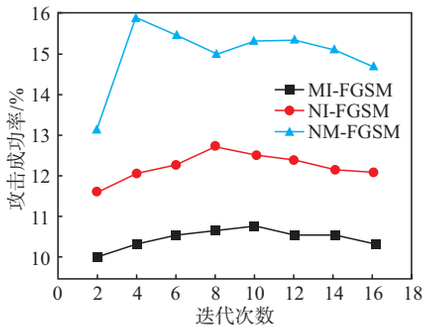

图 3　各算法攻击 Inc-v3ens4 成功率随迭代次数变化关系

Fig. 3　Relationship between attack success rate and iteration number for various algorithms targeting Inc-v3ens4

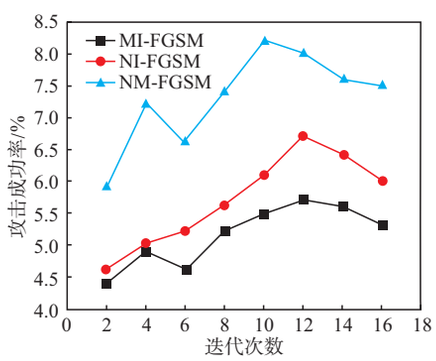

图 4　各算法攻击 IncRes-v2ens 成功率随迭代次数变化关系

Fig. 4　Relationship between attack success rate and iteration number for various algorithms targeting IncRes-v2ens

从图中可以看出，在不同迭代次数场景下，NM-FGSM 总体成功率超过 MI-FGSM 和 NI-FGSM。从另一个角度看，NM-FGSM 只需较少的迭代次数，就能获得与其他 2 种攻击类似的成功率。这表明，在成功率相同情况下，NM-FGSM 所需时间成本更少。例如，在攻击 Inc-v3ens4 时，NM-FGSM 仅用 4 次迭代就获得了约 16%的成功率，而 MI-FGSM 和 NI-FGSM 到达最大成功率分别需要 10 次和 8 次迭代，充分体现了 NM-FGSM 的优势。另外，结合表 1 的内容，从单个对抗样本生成时间来看，除了 FGSM 由于不需迭代，生成过程只有 1 步之外，NM-FGSM 具有明显的时间优势。综合来看，这种时间优势和收敛速度对对抗样本在实际中的应用有重要意义，能够为后续增强模型鲁棒性的对抗性训练提供有效数据来源。

4.6　动态正则化方法验证实验

本节分别为 MI-FGSM、NI-FGSM 和 NM-FGSM 3 种算法加入本文提出的动态正则项，仍然以 Inc-v3 为替代模型产生对抗样本，比较改进之后，各个目标模型上的攻击成功率提升效果，如表 4 所示。

表 4 动态正则化引入前后的攻击成功率对比

Tab. 4 Comparison of attack success rates before and after the introduction of dynamic regularization

%

算法	Inc-v3 *	Inc-v4	IncRes-v2	Res-101	Inc-v3ens3	Inc-v3ens4	IncRes-v2ens
MI-FGSM	99.2 *	37.8	36.8	34.0	11.5	11.2	5.7
MI-FGSM(动态正则)	**99.7** *	**44.8**	**43.1**	**36.2**	**13.8**	**13.9**	**6.7**
NI-FGSM	99.2 *	42.5	40.6	**35.3**	13.4	13.0	6.7
NI-FGSM(动态正则)	**99.4** *	**45.6**	**44.1**	34.9	**16.6**	**16.3**	**9.5**
NM-FGSM	98.8 *	**49.2**	45.2	39.1	15.9	15.4	8.2
NM-FGSM(动态正则)	**99.2** *	48.7	**48.6**	**43.0**	**18.3**	**19.6**	**14.3**

由结果分析得到,动态正则项的引入,在白盒攻击和黑盒迁移攻击中均取得较好效果,使得 3 种算法的迁移攻击平均成功率均取得了 3%以上的提升效果,能够有效提升对抗样本的生成质量和迁移性。但可以看出,NI-FGSM 在 Res-101 模型和 NM-FGSM 在 Inc-v4 模型上的优化效果不太理想,与模型的复杂度以及动态正则项的敏感性相关,对于该模型的决策边界,该优化未能很好地适应。

表 5 统计了 4 种模型中,NM-FGSM 在融入动态正则化项前后的时间成本对比,主要为单个样本生成时间的变化。可以看出,融入动态正则化项后,NM-FGSM 在生成单个对抗样本的时间上会增加 5%～8%,增加量相对较小,符合对于时间复杂度的分析,这个时间增加量在大多数应用场景下都是可接受的。特别是在批处理或离线生成对抗样本的场景中,即使时间有所增加,也不会对整体应用产生显著影响。

表 5 动态正则化项导致的 NM-FGSM 时间成本变化

Tab. 5 Change in time cost of NM-FGSM due to dynamic regularization term

模型	原时间/s	融入后时间/s
Inc-v3	1.20	1.26
Inc-v4	3.22	3.40
IncRes-v2	3.41	3.63
Res-101	4.04	4.32

5 结语

本文在深入分析当前对抗攻击与防御领域现状的基础上,提出了基于 NadaMax 优化器的迭代快速梯度方法(NM-FGSM),通过引入自适应学习率、前瞻动量向量和动态正则化,提升了对抗样本的迁移性和黑盒攻击能力。实验结果表明,NM-FGSM 在多种攻击策略下均优于现有方法。未来工作将集中于通过调整参数、结合其他优化技术等方式深度整合算法,或引入更多样化的输入变换、结合生成对抗网络等方式进行策略创新,进一步完善和优化 NM-FGSM,以提升其在不同场景下的表现。同时,针对 NM-FGSM 生成的对抗样本,探索设计更加有效的防御措施,以期为深度学习模型的安全性研究贡献更多有价值的成果。

参考文献

[1] GOODFELLOW I J, SHLENS J, SZEGEDY C, et al. Explaining and Harnessing Adversarial Examples [EB/OL]. (2015-03-20)[2024-10-20]. https://arxiv.org/abs/1412.6572.

[2] CHEN S Z, HE Z B, SUN C J, et al. Universal Adversarial Attack on Attention and the Resulting Dataset Damagenet[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(4): 2188-2197.

[3] WANG H B, ZHU C X, CAO Y J, et al. ADSAttack: An Adversarial Attack Algorithm via Searching Adversarial Distribution in Latent Space[J]. Electronics, 2023, 12(4): 816.

[4] LIU Y P, CHEN X Y, LIU C, et al. Delving into Transferable Adversarial Examples and Black-Box Attacks[EB/OL]. (2016-02-07)[2024-10-20]. https://arxiv.org/abs/1611.02770.

[5] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards Deep Learning Models Resistant to Adversarial Attacks [EB/OL]. (2019-09-04) [2024-10-20]. https://arxiv.org/abs/1706.06083.

[6] GUO C, RANA M, CISSE M, et al. Countering Adversarial Images Using Input Transformations [EB/OL]. (2018-01-25)[2024-10-20]. https://arxiv.org/abs/1711.00117.

[7] XIE C H, WU Y X, VAN DER MAATEN L, et al. Feature Denoising for Improving Adversarial Robustness[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA: IEEE, 2019: 501-509.

[8] COHEN J, ROSENFELD E, KOLTER Z. Certified Adversarial Robustness via Randomized Smoothing [C]//Proceedings of the 36th International Confer-

ence on Machine Learning (ICML). Long Beach, CA: PMLR, 2019: 1310-1320.

[9] TRAMÈR F, PAPERNOT N, GOODFELLOW I, et al. The Space of Transferable Adversarial Examples [EB/OL]. (2017-08-28)[2024-10-20]. https://arxiv.org/abs/1704.03453.

[10] PAPERNOT N, MCDANIEL P, WU X, et al. Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks[C]//2016 IEEE Symposium on Security and Privacy (SP). San Jose, CA: IEEE, 2016: 582-597.

[11] CARLINI N, WAGNER D. Towards Evaluating the Robustness of Neural Networks [C]//2017 IEEE Symposium on Security and Privacy (SP). San Jose, CA: IEEE, 2017: 39-57.

[12] KINGMA D P, BA J, HAMMAD M M. Adam: A Method for Stochastic Optimization[EB/OL]. (2017-01-30)[2024-10-20]. https://arxiv.org/abs/1412.6980.

[13] DONG Y P, LIAO F Z, PANG T Y, et al. Boosting Adversarial Attacks with Momentum [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT: IEEE, 2018: 9185-9193.

[14] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial Examples in the Physical World[M]//Artificial Intelligence Safety and Security. Boca Raton: Chapman and Hall/CRC, 2018: 99-112.

[15] WANG X S, HE K. Enhancing the Transferability of Adversarial Attacks through Variance Tuning[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN: IEEE, 2021: 1924-1933.

[16] XIE C H, ZHANG Z S, ZHOU Y Y, et al. Improving Transferability of Adversarial Examples with Input Diversity[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA: IEEE, 2019: 2730-2739.

[17] WANG X S, HE X R, WANG J D, et al. Admix: Enhancing the Transferability of Adversarial Attacks [C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC: IEEE, 2021: 16138-16147.

[18] WANG Z Y, ZHANG Z L, LIANG S Y, et al. Diversifying the High-Level Features for Better Adversarial Transferability[EB/OL]. (2023-09-15)[2024-10-20]. https://arxiv.org/abs/2304.10136.

[19] WU W B, SU Y X, LYU M R, et al. Improving the Transferability of Adversarial Samples with Adversarial Transformations[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN: IEEE, 2021: 9024-9033.

[20] HUANG H, CHEN Z Y, CHEN H R, et al. T-SEA: Transfer-Based Self-Ensemble Attack on Object Detection[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, BC: IEEE, 2023: 20514-20523.

[21] HUANG Q, KATSMAN I, GU Z Q, et al. Enhancing Adversarial Example Transferability with an Intermediate Level Attack[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul: IEEE, 2019: 4733-4742.

[22] 王茂源. 高迁移的对抗样本生成算法研究[D]. 南京：南京信息工程大学，2024.
WANG M Y. Research on High-Transferable Adversarial Example Generation Algorithms[D]. Nanjing: Nanjing University of Information Science and Technology, 2024. (in Chinese)

[23] 林志. 可迁移对抗样本生成算法的研究[D]. 绵阳：西南科技大学，2024.
LIN Z. Research on Transferable Adversarial Example Generation Algorithms[D]. Mianyang: Southwest University of Science and Technology, 2024. (in Chinese)

[24] 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. Dallas, Texas: ACM, 2017: 15-26.

[25] SU J W, VARGAS D V, SAKURAI K. One Pixel Attack for Fooling Deep Neural Networks[J]. IEEE Transactions on Evolutionary Computation, 2019, 23(5): 828-841.

[26] BRENDEL W, RAUBER J, BETHGE M. Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models[EB/OL]. (2018-02-16) [2024-10-20]. https://arxiv.org/abs/1712.04248.

[27] LIN J D, SONG C B, HE K, et al. Nesterov Accelerated Gradient and Scale Invariance for Adversarial Attacks [EB/OL]. (2020-02-03)[2024-10-20]. https://arxiv.org/abs/1908.06281.

（编辑：杜娟）