计算机科学
COMPUTER SCIENCE

一种基于线性插值的对抗攻击方法

陈军, 周强, 鲍蕾, 陶卿

引用本文

陈军, 周强, 鲍蕾, 陶卿. 一种基于线性插值的对抗攻击方法[J]. 计算机科学, 2025, 52(8): 403-410.

CHEN Jun, ZHOU Qiang, BAO Lei, TAO Qing. Linear Interpolation Method for Adversarial Attack[J]. Computer Science, 2025, 52(8): 403-410.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240700058

一种基于线性插值的对抗攻击方法

陈 军 周 强 鲍 蕾 陶 卿
陆军炮兵防空兵学院信息工程系 合肥 230031
(chenjun342423@sina.com)

摘 要 深度神经网络在对抗性样本面前表现出显著的脆弱性,易遭受攻击。对抗性样本的构造可被抽象为一个最大化目标函数的优化问题。然而,基于梯度迭代的方法在处理此类优化问题时往往面临收敛性挑战。这类方法主要依赖梯度符号进行迭代更新,却忽略了梯度的大小和方向信息,导致算法性能不稳定。研究表明,I-FGSM 对抗攻击算法源自优化领域中的随机投影次梯度方法。已有文献指出,在优化问题中,采用线性插值方法替代随机投影次梯度方法能够获得优异的性能。鉴于此,提出一种新型的基于线性插值的对抗攻击方法。该方法将插值策略应用于对抗攻击中,并以实际梯度替代传统的符号梯度。理论上,所提出的线性插值对抗攻击算法已被证明在一般凸优化问题中能够实现最优的个体收敛速率,从而克服符号梯度类算法的收敛难题。实验结果证实,线性插值方法作为一种通用且高效的策略,与基于梯度的对抗攻击方法相结合,能够形成新的攻击算法。相较于已有算法,这些新的攻击算法在保持对抗性样本的不可察觉性的同时,显著提升了攻击成功率,并在迭代过程中保持了较高的稳定性。

关键词:线性插值;对抗攻击;梯度符号;收敛性;稳定性
中图分类号 TP391

Linear Interpolation Method for Adversarial Attack

CHEN Jun,ZHOU Qiang,BAO Lei and TAO Qing
Department of Information Engineering,PLA Army Academy of Artillery and Air Defense,Hefei 230031,China

Abstract Deep neural networks exhibit significant vulnerability in the face of adversarial examples and are prone to attacks. The construction of adversarial examples can be abstracted as an optimization problem that maximizes the objective function. However,gradient-based iterative methods often face convergence challenges when dealing with such optimization problems. These methods primarily rely on the gradient sign for iterative updates,neglecting the magnitude and direction information of the gradient,which can lead to algorithm instability. Studies have shown that the I-FGSM adversarial attack algorithm originates from the stochastic projection subgradient method in the field of optimization. Literature has indicated that in optimization problems,using linear interpolation methods to replace stochastic projection subgradient methods can achieve superior performance. Based on this,this paper proposes a novel linear interpolation-based adversarial attack method,which applies the interpolation strategy to adversarial attacks and replaces the traditional sign gradient with the actual gradient. Theoretically,the proposed linear interpolation adversarial attack algorithm is proved can achieve the optimal individual convergence rate in general convex optimization pro-blems,thereby overcoming the convergence difficulties of sign gradient-based algorithms. Experimental results confirm that the linear interpolation method,as a universal and efficient strategy,when combined with gradient-based adversarial attack methods,can form new attack algorithms. Compared to the original algorithms,these new algorithms significantly increase the success rate of attacks while maintaining the imperceptibility of adversarial examples and exhibit high stability during the iterative process.

Keywords Linear interpolation,Adversarial attack,Gradient sign,Convergence,stability

1 引言

深度神经网络模型(Deep Neural Networks,DNNs)在人脸识别[1]、自动驾驶[2]、图像处理[3]、自然语言处理[4]等多个领域都有广泛应用。然而,研究发现,这些深度模型存在一定的脆弱性[5-6]。在图像领域,通过对原始样本添加肉眼难以察觉的噪声生成对抗样本,可以轻易地使 DNNs 模型产生误分类,从而达到攻击的目的。鉴于对抗样本对图像识别分类应用的极大影响,以及它对人身和财产安全、公共交通安全等方面构成的严重威胁,研究对抗样本的生成方式和原理,并寻找相应的防御方法以增强模型的鲁棒性变得尤为重要。根据对模型结构和参数的了解程度,对抗攻击可以分为白盒攻击

到稿日期:2024-07-09 返修日期:2024-09-24
通信作者:陶卿(taoqing@gmail.com)

和黑盒攻击。白盒攻击是在已知模型和参数的情况下，通过最大化损失来生成对抗样本的优化过程。研究表明，对抗样本具有迁移性[7]：在一个已知模型上生成的对抗样本能够欺骗结构和参数未知的模型，这意味着它具有黑盒攻击的能力。

在白盒攻击场景中，若干攻击策略已被证明具有显著的攻击效果[8-10]。然而，当这些策略应用于黑盒模型时，其迁移能力普遍较弱，这一点在对抗防御模型时尤为明显[11-12]。先前的研究成果[13-15]指出，对抗性样本的低迁移性可归因于攻击过程中陷入次优的局部极值点，或是对代理模型的过度拟合，这些情况的出现使得对抗样本在不同模型间难以实现有效的迁移。针对这一挑战，研究者从多个角度出发，提出了多种解决方案。梯度优化攻击方法[6,13-14]致力于通过精确的梯度计算来增强黑盒攻击的性能。输入变换攻击策略[14-16]专注于通过在输入数据上施加多样化的变换，以生成更具迁移性的对抗性样本。基于模型集成的攻击方法使用多个模型作为白盒，使用加权平均的方式计算对抗样本迭代梯度，提升对抗样本的迁移性[6,13]。这些方法主要从优化和泛化的理论框架出发，将白盒模型上生成对抗性样本的过程类比于标准的神经网络训练流程，并将对抗性样本的迁移能力与模型的泛化能力等同[6]。尽管这些方法在一定程度上提升了对抗性样本的迁移性，但白盒攻击与基于迁移的黑盒攻击之间仍存在一定差距。

研究表明，输入变换策略、模型集成策略可以与基于梯度优化的方法有效结合，从而增强对抗样本的迁移性[6]。本文重点研究基于梯度优化的对抗攻击方法。从快速梯度符号方法(Fast Gradient Sign Method，FGSM)[8]开始，研究者提出了许多有效提升对抗样本迁移性的对抗攻击方法，如 I-FGSM[9]，MI-FGSM[13]，NI-FGSM[14]，GI-FGSM[17]，VMI-FGSM[6]，EMI-FGSM[18]，IE-FGSM[19]，PGN[20]等，以及最近提出的 ANDA[21]算法。本质上，这些方法都是通过有效的数学优化方法，来有效地提高梯度的精确性，进而提升对抗样本的迁移性。

然而，在生成对抗性样本的过程中，上述算法均采用了梯度的符号信息来替代实际的梯度值。Karimireddy 等[22]在 2019 年就已经指出了梯度符号算法在收敛性实现方面的难题。他们通过构造具体的反例指出，即便在简单的凸性条件下，此类算法也可能无法实现收敛。对于依托梯度优化的对抗性攻击算法而言，探究其收敛性质是极其关键的，这直接关联到算法稳定性的理论保障。该问题的核心在于，梯度符号算法仅利用梯度的符号信息来决定更新方向，而未考虑梯度的实际大小和方向，构成了一种偏颇的梯度近似。由于这种做法无法确保学习率在整个迭代过程中保持单调性，因此算法的收敛性无法得到有效保证。

Tao 等[23]在随机投影次梯度优化方法的基础上提出一种嵌入线性插值操作的投影次梯度方法，证明了在一般凸情形下该方法具有 $O(1/\sqrt{t})$ 的个体最优收敛速率，并进一步得到了对应的随机方法也具有最优的个体收敛速率的结论。该方法的个体收敛结果在系统实施稳定化中也具有广泛的应用前景。而研究表明，对抗攻击中的 I-FGSM 算法是由随机投影次梯度算法发展而来的。

受上述研究启发，本文将线性插值方法运用于图像对抗攻击中，使用真正的梯度替代梯度符号，有效解决了梯度符号算法不易收敛的问题。此外，在自然语言处理、语音识别等领域，从理论上来说，只要在神经网络训练过程中使用梯度，就能够采用线性插值方法实现有效的攻击。

本文的主要贡献如下：

1)提出了一种基于线性插值的迭代对抗攻击方法(Linear Interpolation Method for Iterative Adversarial Attack，LIM)，首次将优化领域的线性插值策略应用于对抗性攻击之中。相较于 I-FGSM 算法，该线性插值方法在未增加额外计算负担的前提下，实现了更为精确的迭代方向，从而显著提升了对抗性样本的攻击成功率。

2)证明了 LIM 算法在一般凸问题上能够实现个体收敛，克服了 I-FGSM 算法不收敛的问题，从理论上确保了生成对抗样本过程的稳定性。

3)大量的实验结果表明，线性插值方法作为一种普适且高效的策略，能够与现有提出的各类梯度算法相融合，形成新型的对抗性攻击算法。与原始算法相比，这些经过融合的算法在维持生成对抗性样本不可察觉性的同时，显著提高了攻击的成功率，并且在迭代过程中保持着极高的稳定性。

2 相关工作

假设初始样本为 $\boldsymbol{x}$，y 为样本对应的真实标签。$c(\boldsymbol{x};\theta)$ 是参数为 θ、输出为预测标签值的分类器。假设该分类器已经通过训练得到了优化的参数，从而能够对输入样本进行准确分类，即当输入 $\boldsymbol{x}$ 为初始样本时，输出的 y 为正确的标签，$c(\boldsymbol{x};\theta)=y$。在此设 $f(\boldsymbol{x},y;\theta)$ 表示分类器 $c(\boldsymbol{x};\theta)$ 的交叉熵损失函数。定义对抗攻击是在初始样本 $\boldsymbol{x}$ 邻域内找到一个对抗样本 $\boldsymbol{x}^{adv}$，该样本满足以下条件：与初始样本的 p 范数距离在一定范围内，即 $\|\boldsymbol{x}^{adv}-\boldsymbol{x}\|_p\leqslant\varepsilon$ 并且使得分类器 c 分类错误，也即 $c(\boldsymbol{x}^{adv};\theta)\neq y$。这里，$\|\cdot\|_p$ 表示 p 范数距离，p 的取值根据具体情况可以设定为 0，1，2，∞。本文主要关注 L_∞ 范数下的攻击方法。

本章对 FGSM，I-FGSM，MI-FGSM，NI-FGSM 等几种常见的符号梯度算法与线性插值优化算法进行分析介绍。

对于一个训练好参数的分类模型来说，产生对抗样本的过程实际上是一个优化问题，目标是寻求最大化模型损失的对抗样本[13]。

$$\max f(\boldsymbol{x};y;\theta)$$

$$\text{s.t. } \|\boldsymbol{x}^{adv}-\boldsymbol{x}\|_\infty\leqslant\varepsilon$$

这里假设损失函数 $f(\boldsymbol{x};y;\theta)$ 一阶可微，后续将损失函数简写为 $f(\boldsymbol{x})$。

2.1 几种梯度符号方法

2.1.1 快速梯度符号算法

Goodfellow 等[8]在 2015 年提出了符号梯度攻击算法 FGSM，更新规则为：

$$\boldsymbol{x}^{adv}=\boldsymbol{x}+\varepsilon\cdot sign(f(\boldsymbol{x})) \tag{1}$$

其中，$sign(\cdot)$ 为符号函数，$\nabla f(\boldsymbol{x})$ 表示损失函数对输入样本 $\boldsymbol{x}$ 的梯度值。可以看出，FGSM 方法实际上是一种单步迭代且步长为 ε 的算法，满足 $\|\boldsymbol{x}-\boldsymbol{x}^{adv}\|_\infty\leqslant\varepsilon$ 的约束条件。

2.1.2　迭代快速梯度符号算法

Kurakin 等[9]在 FGSM 的基础上，于 2017 年提出了迭代快速梯度符号算法 I-FGSM，通过迭代若干次来生成对抗样本，主要更新规则为：

$$\boldsymbol{x}_0^{\mathrm{adv}}=\boldsymbol{x} \tag{2}$$

$$\boldsymbol{x}_{t+1}^{\mathrm{adv}}=\boldsymbol{x}_t^{\mathrm{adv}}+a\cdot sign(\nabla f(\boldsymbol{x}_t^{\mathrm{adv}})) \tag{3}$$

为了满足 $\|\boldsymbol{x}-\boldsymbol{x}^{\mathrm{adv}}\|_\infty\leqslant\varepsilon$ 的约束条件，这里设定每次更新迭代的梯度符号的步长 α 为 ε/T，T 为迭代次数。实验表明，与单步攻击方法相比，I-FGSM 方法具有更高的白盒攻击成功率，但是迁移性稍弱。

2.1.3　基于 Momentum 的迭代快速梯度符号方法

Dong 等[13]将动量方法融入 I-FGSM 算法中，形成了 MI-FGSM 方法。其更新规则描述如下：

$$\boldsymbol{x}_0^{\mathrm{adv}}=\boldsymbol{x},g_0=0,\alpha=\frac{\varepsilon}{T} \tag{4}$$

$$\boldsymbol{g}_{t+1}=\mu\boldsymbol{g}_t+\frac{\nabla f(\boldsymbol{x}_t^{\mathrm{adv}})}{\|\nabla f(\boldsymbol{x}_t^{\mathrm{adv}})\|_1} \tag{5}$$

$$\boldsymbol{x}_{t+1}^{\mathrm{adv}}=\boldsymbol{x}_t^{\mathrm{adv}}+\alpha\cdot sign(\boldsymbol{g}_{t+1}) \tag{6}$$

其中，$\boldsymbol{g}_t$ 为前 t 次迭代中累加的梯度，μ 为动量系数。MI-FGSM 方法由于添加了动量项，累积了历史梯度信息，因此能够稳定更新方向并容易跳过局部极值点。实验结果表明，该算法与一般符号梯度方法相比，显著提高了白盒与黑盒攻击的成功率，且能保持较好的稳定性。

2.1.4　基于 Nesterov 型动量的迭代快速梯度符号方法

Lin 等[14]将 NAG 算法引入对抗样本的生成过程中，提出了基于 Nesterov 型动量的迭代快速梯度符号方法 NI-FGSM。其更新规则描述如下：

$$\boldsymbol{x}_0^{\mathrm{adv}}=\boldsymbol{x},\boldsymbol{g}_0=0 \tag{7}$$

$$\boldsymbol{x}_t^{\mathrm{nes}}=\boldsymbol{x}_t^{\mathrm{adv}}+a\cdot\mu\cdot\boldsymbol{g}_t \tag{8}$$

$$\boldsymbol{g}_{t+1}=\mu\boldsymbol{g}_t+\frac{\nabla f(\boldsymbol{x}_t^{\mathrm{nes}})}{\|\nabla f(\boldsymbol{x}_t^{\mathrm{nes}})\|_1} \tag{9}$$

$$\boldsymbol{x}_{t+1}^{\mathrm{adv}}=Clip_{x,\varepsilon}\{\boldsymbol{x}_t^{\mathrm{adv}}+\alpha\cdot sign(\boldsymbol{g}_{t+1})\} \tag{10}$$

其中，$\boldsymbol{x}_t^{\mathrm{nes}}$ 为 Nesterov 项。NAG 算法相比 Heavy-ball 型动量方法多了一个本次梯度相对上次梯度的变化量，即 Nesterov 动量项，使其具有向前一步的特性。这种向前一步的特性使得 NI-FGSM 与 MI-FGSM 相比，不仅能够保持迭代方向的稳定性，还能够提高收敛速率，提升对抗样本攻击成功率。

为了进一步提升对抗性样本的迁移性，后续的研究工作相继提出了多种基于梯度符号的对抗性攻击算法。这些算法主要采用多样化的优化技术来计算更为精确的梯度更新方向，从而有效提高对抗性样本在不同模型间的迁移能力。尽管如此，这些方法仍旧采用梯度的符号信息来确定最终的更新方向，因此仍然面临着算法收敛的挑战。对于依托梯度优化的对抗性攻击算法而言，探究其收敛性质是至关重要的，这是因为收敛性直接关系到算法稳定性的理论保障。因此，第 3 章将着重分析所提线性插值方法(LIM)的收敛性特性。

2.2　线性插值优化算法

求解优化问题：

$$\begin{aligned}&\min\ f(\boldsymbol{x})\\&\text{s.t.}\ \boldsymbol{x}\subseteq Q\end{aligned} \tag{11}$$

目前常用的是随机投影次梯度算法[23]，主要更新规则为：

$$\boldsymbol{x}_{t+1}=P_Q(\boldsymbol{x}_t-\alpha_t\nabla f(\boldsymbol{x}_t)) \tag{12}$$

其中，$Q\subseteq R^N$ 为定义域内的闭凸集，$f(\boldsymbol{x})$ 为凸函数，$\nabla f(\boldsymbol{x}_t)$ 是函数 f 在 $\boldsymbol{x}_t$ 处的次梯度，P_Q 表示迭代后的值向定义域 Q 的投影操作。通过设置适当的迭代步长，可以推出一般凸问题的平均收敛速率为 $O(1/\sqrt{t})$，即 $f\left(\frac{1}{t}\sum_{k=1}^{t}\boldsymbol{x}_k\right)-f(\boldsymbol{x}^*)\leqslant O\left(\frac{1}{\sqrt{t}}\right)$。这里假设 $\boldsymbol{x}^*$ 为式(11)所求优化问题的最优解。容易得出，式(3)与式(12)算法思想基本一致，I-FGSM 对抗攻击算法由随机投影次梯度算法发展而来。

Tao 等[23]提出了一种线性插值投影次梯度优化算法求解式(11)的优化问题，并证明了在一般凸情况下具有 $O(1/\sqrt{t})$ 的个体最优收敛速率。具体算法如下：

$$\begin{cases}\boldsymbol{x}_t^+=P_Q(\boldsymbol{x}_{t-1}^+-a_t\eta_t\nabla f(\boldsymbol{x}_t))\\\boldsymbol{x}_{t+1}=\dfrac{A_t}{A_{t+1}}\boldsymbol{x}_t+\dfrac{t+1}{A_{t+1}}\boldsymbol{x}_t^+\end{cases} \tag{13}$$

该算法沿 $\boldsymbol{x}_{t-1}^+$ 方向做投影次梯度计算，在梯度运算 $\nabla f(\boldsymbol{x}_t)$ 中，$\boldsymbol{x}_t$ 是由线性插值形式获得的，因此称该算法为嵌入线性插值操作的投影次梯度方法。研究表明，这种嵌入线性插值操作的投影次梯度方法能够有效提升优化收敛效果，并被证明具有个体收敛性。

3　算法及收敛性分析

本文提出的基于线性插值的对抗攻击方法，有效解决了 I-FGSM 方法不易收敛的问题，提高了对抗样本的攻击成功率。具体伪代码如算法 1 所示。

算法 1　基于线性插值的对抗攻击方法

输入：干净样本 $\mathbf{x}$，标签 y，损失函数 f，扰动 ε，初始步长 α，步长系数 $\eta>0$，插值系数 A_t，迭代次数 T

输出：对抗样本 $\mathbf{x}_T$

1. $\mathbf{x}_0^+=\mathbf{x}_1=\mathbf{x};\mathbf{g}_0=0;A_1=1;\mu=1$

2. for t=1 to T−1 do

3. $\quad\alpha_t=\dfrac{\alpha}{\sqrt{t}};A_{t+1}=A_t+(t+1)$

4. $\quad\mathbf{x}_t^+=P_Q\left(\mathbf{x}_{t-1}^+-\alpha_t\eta\dfrac{\nabla f(\mathbf{x}_t)}{\|\nabla f(\mathbf{x}_t)\|_1}\right)$

5. $\quad\mathbf{x}_{t+1}=\dfrac{A_t}{A_{t+1}}\mathbf{x}_t+\dfrac{t+1}{A_{t+1}}\mathbf{x}_t^+$

6. end for

算法 1 中，步骤 4 与优化算法中的线性插值法在求解过程上存在两个主要差异。首先，它对梯度进行了 L_1 范数归一化处理；其次，优化算法旨在训练模型参数，以最小化模型输出的误差，而本文提出的对抗攻击算法的目的与此相反，其核心在于通过添加噪声生成对抗性样本，使模型损失最大化，因此其运算符号与优化算法是相反的。两者的相似之处在于，求解梯度过程中使用的数据都是通过步骤 5 插值方法获得的。针对 I-FGSM 算法的收敛性问题，式(3)清晰地显示在生成对抗性样本时，运用符号函数会导致无法维持迭代步长的单调性，这是算法难以收敛的根本原因。本文通过插值的方法，在迭代过程中使用归一化的梯度值 $\frac{\nabla f(\boldsymbol{x}_t)}{\|\nabla f(\boldsymbol{x}_t)\|_1}$ 代替梯

度符号，并使用单调递减的步长α_t，确保算法收敛。为了证明所提算法的收敛性，做如下假设：

假设 1 假设存在常数 $G\geqslant 0$，$\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in Q$

$$\| \boldsymbol{x}_1 - \boldsymbol{x}_2 \|_2 \leqslant G \tag{14}$$

假设 2 假设存在常数 $M\geqslant 0$，$\forall \boldsymbol{x} \in Q$

$$\| \nabla f(\boldsymbol{x}) \|_1 \leqslant M \tag{15}$$

定理 1 在式(14)和式(15)两个假设条件成立的情况下，假设$\boldsymbol{x}^*$为给定约束条件下对抗样本的最优解，对抗样本$\boldsymbol{x}_t$由算法 1 生成，则不等式(16)成立：

$$f(\boldsymbol{x}^*) - f(\boldsymbol{x}_t) < \frac{2}{3}\frac{a\eta M t^{\frac{3}{2}}}{t(t+1)} + \frac{MG^2 t^{\frac{3}{2}}}{\alpha\eta t(t+1)} \tag{16}$$

其中，$\boldsymbol{x}_t$为迭代 t 次后的对抗样本。为便于理解定理 1，做如下说明：

1)本文算法在一般凸情况下取得了个体最优收敛速率$O(1/\sqrt{t})$，从理论上克服了 I-FGSM 算法使用梯度符号导致的不收敛问题。

2)在假设样本 $\boldsymbol{x}$ 能够被分类器正确分类的前提下，经过对抗攻击算法成功生成对抗性样本$\boldsymbol{x}_t^{\text{adv}}$后，其损失函数值将增加，即 $f(\boldsymbol{x}_t^{\text{adv}}) > f(\boldsymbol{x})$。因此可以推断，损失函数在对抗样本生成的邻域内是局部凹的。在运算过程中，由于约束区域范围较小，认为上述假设成立。据此，在证明过程中，利用了凹函数的性质。综上，线性插值算法在对抗性攻击中的应用不仅对一般凸问题有效，也适用于非凸情况下的神经网络模型。

3)RMSProp[24]和 Adam[25]强调对算法收敛性的分析，对确定算法中超参数的选取具有指导性作用。相应地，本研究对算法收敛性进行探讨，不仅为算法的稳定性提供了理论上的保障，而且对于后续过程中超参数(如迭代步长)的优化选择提供了强大的理论支持。

4)在迭代过程中，线性插值方法(LIM)未使用梯度符号，相较于 I-FGSM 算法，有效地提高了对抗性样本的攻击成功率。实验结果表明，线性插值作为一种通用策略，能够与 MI-FGSM，NI-FGSM，VMI-FGSM 等梯度算法相结合，形成相应的去符号化线性插值梯度方法，即 LIM-MI，LIM-NI，LIM-VMI 等。结果还表明，线性插值算法亦能与 GI-FGSM，EMI-FGSM，IE-FGSM，PGN 等梯度方法相结合，形成基于线性插值的新型对抗性攻击算法，此处不再逐一详述。LIM-MI (Linear Interpolation Method for Momentum Iterative Adversarial Attack)算法的伪代码见算法 2，其他与线性插值相结合的对抗性攻击算法表达式可依据插值定义推导得出。

算法 2 一种基于线性插值的动量对抗攻击方法

输入：干净样本 $\mathbf{x}$，标签 y，损失函数 f，扰动 ε，初始步长 α，步长系数 $\eta>0$，插值系数A_t，迭代次数 T，衰减系数 μ

输出：对抗样本$\mathbf{x}_T$

1. $\mathbf{x}_0^+ = \mathbf{x}_1 = \mathbf{x}; \mathbf{g}_0 = 0; A_1 = 1$

2. for t=1 to T−1 do

3. $\alpha_t = \frac{\alpha}{\sqrt{t}}; A_{t+1} = A_t + (t+1)$

4. $\mathbf{g}_{t+1} = \mu\,\mathbf{g}_t + \frac{\nabla f(\mathbf{x}_t)}{\| \nabla f(\mathbf{x}_t) \|_1}$

5. $\mathbf{x}_t^+ = P_Q(\mathbf{x}_{t-1}^+ + \alpha_t \eta\, \mathbf{g}_{t+1})$

6. $\mathbf{x}_{t+1} = \frac{A_t}{A_{t+1}}\mathbf{x}_t + \frac{t+1}{A_{t+1}}\mathbf{x}_t^+$

7. end for

4 实验及结果分析

本章通过大量实验，验证线性插值算法的有效性。使用 2 个指标来衡量对抗样本性能，分别为攻击成功率和不可察觉性。

攻击成功率是指生成的对抗样本被模型误分类的概率。

$$\frac{\sum_{i=1}^{N}[f(\boldsymbol{x}_i) = y_i \wedge f(\boldsymbol{x}_i^*) \neq y_i]}{\sum_{i=1}^{N}[f(\boldsymbol{x}_i) = y_i]} \tag{17}$$

其中，$\boldsymbol{x}_i$表示干净样本，$\boldsymbol{x}_i^*$表示对抗样本，y_i表示样本的真实标签，N 是数据集样本的总数。

一种常用的不可察觉性指标是使用扰动的L_P范数，它具有简单、数据易处理及实用性强等优点[26]。本文主要使用平均L_P范数失真(ALD_P)作为对抗样本不可察觉性指标，这里 p 取 1，ALD_P的数学表达式如下：

$$ALD_P = \frac{1}{N}\sum_{(i,j)} |\boldsymbol{x} - \boldsymbol{x}^*| \tag{18}$$

其中，$\boldsymbol{x}$ 表示干净样本，$\boldsymbol{x}^*$ 表示对抗样本，(i,j)表示图像的像素位置，N 是图像中像素点的总数。ALD_P值越小，说明不可察觉性越好，反之则越差。

本文实验主要有以下目的：

1)将所提出的线性插值方法 LIM，LIM-MI，LIM-NI，LIM-VMI 分别与对应的基于梯度符号的方法 IFGSM，MI-FGSM，NI-FGSM，VMI-FGSM 在一般模型上进行比较，以验证所提算法的有效性。为进一步验证算法的性能，并便于比较分析，本研究主要以 Inception-v3 模型作为白盒环境，对 I-FGSM 和 VMI-FGSM 及其对应的 LIM 和 LIM-VMI 这 4 种典型算法进行比较实验。此外，将其他模型作为白盒环境时，上述各类算法的比较亦已实施并经实验验证。为保持文本的简洁性，此处不再逐一阐述。

2)以 Inception-v3 为白盒，使用 LIM 和 LIM-VMI 算法与对应的 IFGSM 和 VMI-FGSM 算法在防御模型上比较对抗样本攻击成功率与 ALD_P值，验证本文算法的有效性。

3)以 Inception-v3 为白盒，比较 LIM 和 LIM-VMI 算法与对应的 I-FGSM 和 VMI-FGSM 算法在集成模式下对抗样本的攻击成功率与 ALD_P值，验证本文算法的有效性。

4)以 Inception-v3 为白盒，比较 LIM-MI 与 MI-FGSM 算法在使用数据增强方法后生成对抗样本的攻击成功率与 ALD_P值，验证本文算法的有效性。

5)通过对比 LIM-MI 算法与 MI-FGSM 算法的迭代次数与攻击成功率之间的对应关系，验证本文算法在迭代过程中能够保持良好的稳定性。

4.1 实验设置

4.1.1 数据集

本文实验使用的数据集与 MI-FGSM，NI-FGSM 等算法使用的数据集相同，均为从 ILSVRC2012 验证集[27]中随机抽取的 1 000 张属于不同类别的图片。

4.1.2　模型

本文的对比模型包括8个一般模型与4个对抗训练模型。常规训练模型分别从卷积神经网络模型(Convolutional Neural Networks，CNNs)与ViTs(Vision Transformers)模型中选取。其中CNNs模型共有5种，分别为ResNet-34(Res34)[28]，Inception-v3(Inc-v3)[29]，Vgg16[30]，Densenet121(Dens-121)[31]，Mobilenet-V2(Mob-V2)[32]，采用Torchvision库提供的模型预训练参数；ViTs模型有3种，分别为Vit-Base-patch16(Vit-B)[33]，Visformer-Smal(Vis-S)和Swin-Tiny-patch4(Swin-T)[34]。对抗训练模型有4种，分别为Inc-v3adv，IncRes-v2ens，Efficient-B0adv，Efficient-B1adv[35]。ViTs模型和对抗训练模型均采用timm库提供的模型预训练参数。

4.1.3　实验平台

所有实验均在PyTorch框架下编程完成，对抗攻击算法均取自TransferAttack库[1)]，它是一个基于PyTorch框架的专门用于提升对抗样本迁移性的算法库。

4.1.4　超参设置

实验中所有参与比较的对抗攻击算法，其超参均与TransferAttack库中设置的参数保持一致。特别说明的是，本文算法实验中均采用TransferAttack平台默认设置，最大扰动量ε为16/255，步长α为16/255/10，迭代次数$T=10$，学习率系数η在LIM，LIM-MI，LIM-NI算法中取值为3.5，在VMI-FGSM算法中取值为8，其余参数设置与对应算法相同。

4.2　基于动量的对抗攻击

本节在8种一般模型上分别对比不同攻击算法的黑盒攻击成功率。实验分别以Res34，Inc-v3，Vit-B，Vis-S，Swin-T为白盒，攻击成功率如表1所列。

表1　不同算法在一般模型上的攻击成功率

Table 1　Attack success rate of different algorithms on general models

(%)

Model	Methods	Res34	Inc-v3	Vgg16	Dens-121	Mob-V2	Vit-B	Vis-S	Swin-T	Average
Res34	I-FGSM	100.0*	30.3	39.4	40.6	37.8	5.5	15.3	20.8	36.2
	LIM	100.0*	50.1	62.1	67.7	59.3	12.9	26.7	34.3	51.6
	MI-FGSM	100.0*	56.3	70.3	74.7	67.0	16.8	35.4	40.4	57.6
	LIM-MI	100.0*	58.0	72.2	75.1	68.8	17.6	35.4	40.7	58.5
	NI-FGSM	100.0*	59.7	73.4	76.7	71.6	17.8	35.8	41.3	59.5
	LIM-NI	100.0*	59.5	75.0	76.7	71.9	18.2	36.0	41.8	59.9
	VMI-FGSM	100.0*	73.9	83.5	88.4	80.2	32.2	**55.5**	58.1	71.5
	LIM-VMI	100.0*	**74.5**	**84.8**	**89.3**	**82.8**	**33.3**	54.7	**59.2**	**72.3**
Inc-v3	I-FGSM	21.5	97.7*	25.5	20.4	25.8	4.8	11.0	14.6	27.7
	LIM	39.4	98.9*	41.4	38.1	42.6	10.0	21.1	26.0	39.7
	MI-FGSM	43.2	97.8*	48.2	44.7	47.7	13.0	24.9	29.7	43.7
	LIM-MI	50.1	98.7*	54.5	50.1	54.5	13.9	26.9	31.2	47.5
	NI-FGSM	50.5	98.4*	54.2	51.7	55.9	14.2	27.2	31.8	48.0
	LIM-NI	51.5	98.9*	56.6	52.2	57.8	15.0	27.8	30.7	48.8
	VMI-FGSM	58.4	98.4*	59.5	59.1	59.5	21.3	35.6	40.3	54.0
	LIM-VMI	**59.5**	98.5*	**60.2**	**59.5**	**59.8**	**22.2**	**36.5**	**42.2**	**54.8**
Vit-B	I-FGSM	26.2	28.8	34.3	27.6	34.7	92.9*	22.6	36.1	37.9
	LIM	37.6	41.1	45.8	40.3	46.8	92.1*	33.6	47.9	48.2
	MI-FGSM	47.5	46.4	58.4	51.5	56.4	97.4*	42.8	55.4	57.0
	LIM-MI	48.2	47.9	59.2	52.8	56.4	94.7*	43.8	55.7	57.3
	NI-FGSM	49.4	49.2	59.3	52.2	57.7	96.5*	44.2	57.8	58.3
	LIM-NI	49.5	50.2	59.3	53.8	59.4	96.5*	45.6	57.4	59.0
	VMI-FGSM	56.2	54.9	64.1	60.8	60.6	98.2*	**57.5**	**67.2**	64.9
	LIM-VMI	**56.9**	**57.0**	**65.2**	**61.8**	**63.6**	98.8*	55.4	66.1	**65.6**
Vis-S	I-FGSM	25.2	26.2	34.6	27.5	36.3	11.6	91.4*	34.1	35.9
	LIM	44.1	42.0	56.9	49.1	57.1	24.4	97.0*	50.3	52.6
	MI-FGSM	52.0	50.7	65.1	59.6	63.8	29.7	97.2*	58.2	59.5
	LIM-MI	53.9	50.2	67.0	61.0	65.9	30.5	97.7*	58.8	60.6
	NI-FGSM	53.8	50.5	67.3	61.3	66.0	31.5	98.8*	60.1	61.2
	LIM-NI	55.6	52.7	70.0	63.3	69.2	30.4	98.9*	59.1	62.4
	VMI-FGSM	68.6	66.1	75.8	74.7	75.9	**54.6**	96.9*	76.9	73.7
	LIM-VMI	**69.9**	**68.1**	**80.7**	**79.1**	**78.4**	50.7	98.9*	**79.1**	**75.6**
Swin-T	I-FGSM	16.2	19.2	23.3	15.5	28.4	5.4	13.6	70.1*	24.0
	LIM	30.5	30.0	39.1	30.4	43.1	14.3	28.0	90.3*	38.2
	MI-FGSM	36.2	34.9	48.9	38.8	51.9	21.0	34.1	95.9*	45.2
	LIM-MI	37.5	36.9	50.7	41.3	53.5	19.4	34.4	96.3*	46.3
	NI-FGSM	37.2	37.1	48.8	37.6	53.5	20.3	34.7	96.5*	45.7
	LIM-NI	38.0	36.8	51.5	41.5	54.4	20.3	35.4	96.2*	46.8
	VMI-FGSM	54.8	53.3	62.6	58.5	67.4	**46.1**	60.3	97.8*	62.6
	LIM-VMI	**58.9**	**56.2**	**66.2**	**61.2**	**72.3**	44.7	**62.0**	99.2*	**65.1**

1) https://GitHub-Trustworthy-AI-Group/TransferAttack

其中“*”表示白盒攻击，线性插值算法均以加粗表示。以 Inception-v3 为白盒，不同对抗攻击算法生成的可视化对抗样本如图 1 所示。由表 1 可知，线性插值算法攻击成功率在大部分模型中均优于原梯度符号攻击算法，尤其是 LIM 算法，其在各个模型上的攻击率相较于对比算法 I-FGSM 均有显著提高。图 1 表明在扰动量相同的情况下，从视觉上来说，I-FGSM，LIM，VMI-FGSM，LIM-VMI 算法所生成的对抗样本基本在同一水平。

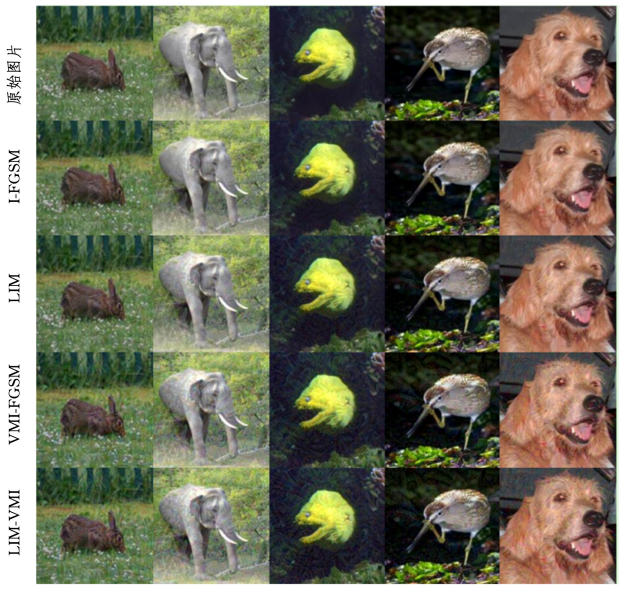


图 1 I-FGSM，LIM，VMI-FGSM，LIM-VMI 在扰动值为 16 时的对抗样本

FIG. 1 Adversarial sample with perturbation value is 16 for I-FGSM, LIM, VMI-FGSM and LIM-VMI

为了进一步验证算法的有效性，以 Inception-v3 为白盒，以 4 个防御模型为黑盒，对比不同算法在防御模型上的对抗样本攻击成功率及其对应的 ALD_P 值大小，其对抗攻击成功率如表 2 所列。由表 2 可知，在防御模型上，LIM 算法的攻击成功率与 I-FGSM 相比依然具有较大的优势，但是其不可察觉性指标 ALD_P 值比 I-FGSM 大。这表明，在一定程度上提高攻击成功率是以损害不可察觉性为代价获取的。VMI-FGSM 与 LIM-VMI 算法生成图像的 ALD_P 值基本相当，LIM-VMI 算法提升了对抗样本的攻击成本率，验证了线性插值算法的有效性。

表 2 不同算法在防御模型上的黑盒攻击成功率

Table 2 Black-box attack success rate of different algorithms on defensive models

(%)

Model	Methods	Inc-v3adv	IncRes-v2ens	Efficient-B0adv	Efficient-B1adv	ALD_P
Inc-v3	I-FGSM	25.2	12.3	13.5	11.2	0.045
	LIM	41.6	24.2	27.8	23.3	0.067
	VMI-FGSM	64.9	**47.5**	45.9	44.0	0.090
	LIM-VMI	**66.1**	46.0	**48.0**	**45.2**	0.092

4.3 集成攻击方法

研究表明，同时使用多个模型集成作为白盒，能够有效提升对抗样本的迁移性[35-36]。为了进一步验证所提算法的有效性，分别将 LIM，LIM-VMI，I-FGSM，VMI-FGSM 算法与集成攻击方法相结合，生成对抗样本并进行比较。需要说明的是，采用 Dong 等[13]提出的 logits 集成方法对模型进行集成，采用加权平均的方式获取最终输出的 logits。使用 Res34，Inc-v3，Vgg16 作为白盒生成对抗样本。攻击成功率如表 3 所列。从表 3 可以看出，在集成攻击模式下，通过对比不同算法对抗样本的攻击成率与 ALD_P 值之间的关系，进一步验证了线性插值算法的有效性。

表 3 不同算法在集成模型下的攻击成功率

Table 3 Attack success rate of different algorithms under ensemble models

(%)

Methods	Res34*	Inc-v3*	Vgg16*	Dens-121	Mob-V2	Vit-B	Vis-S	Swin-T	Inc-v3adv	IncRes-v2ens	Efficient-B0adv	Efficient-B1adv	ALD_P
I-FGSM	99.6*	99.0*	98.5*	56.1	52.4	13.3	25.8	33.5	34.3	18.1	29.2	24.6	0.049
LIM	99.9*	99.7*	99.6*	78.2	71.6	23.1	44.7	48.5	54.3	34.2	52.3	46.3	0.067
VMI-FGSM	99.9*	99.7*	99.8*	90.3	85.0	**44.2**	65.7	68.3	74.9	**61.1**	74.0	72.6	0.093
LIM-VMI	99.9*	100*	99.8*	**94.3**	**89.8**	42.1	**67.4**	**70.1**	**75.5**	60.4	**78.0**	**75.5**	0.099

4.4 数据增强方法

与其他优化算法一样，本文算法可以与输入数据增强方法相结合，从而进一步提升对抗样本的迁移性。为了进一步验证算法的有效性，分别将 MI-FGSM 和 LIM-MI 与 3 种典型的数据增强方法 TI[16]，DI[15]，SI[14] 相结合。为了便于比较，仍设置 Inception-v3 为白盒模型。攻击成功率如表 4 所列。

由表 4 可以看出，与数据增强方法结合后，线性插值算法与原算法相比，在保持对抗样本不可察觉性的基础上，仍然能有效提升攻击成功率。

表 4 算法在不同数据增强方法下的攻击成功率

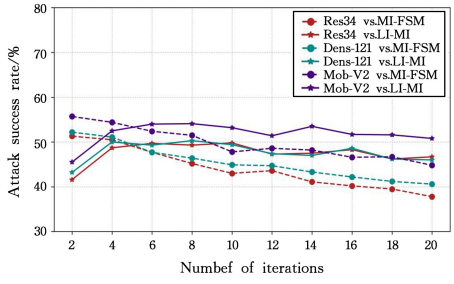
Table 4 Attack success rate of different algorithms under different data augmentation methods

(%)

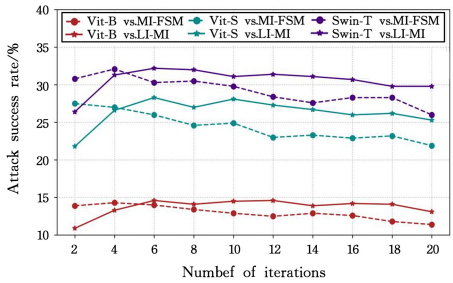
Methods	Res34	Inc-v3*	Vgg16	Dens-121	Mob-V2	Vit-B	Vis-S	Swin-T	Inc-v3adv	IncRes-v2ens	Efficient-B0adv	Efficient-B1adv	ALD_P
TI+MI-FGSM	40.7	97.2	49.2	43.2	40.8	10.5	18.7	23.4	41.7	30.9	26.5	25.9	0.092
TI+LIM-MI	45.4	98.0	52.8	46.5	43.8	10.3	19.9	25.9	47.4	34.2	30.0	29.9	0.096
SI+MI-FGSM	57.0	99.6	59.4	58.9	58.1	16.7	30.4	36.7	62.3	43.9	41.9	39.0	0.092
SI+LIM-MI	59.6	99.8	61.4	**61.4**	61.3	18.3	33.3	36.2	63.8	45.3	46.4	42.4	0.092
DI+MI-FGSM	59.6	98.0	60.7	59.0	59.4	20.7	35.7	**39.6**	**64.2**	**47.0**	46.8	**45.6**	0.091
DI+LIM-MI	**61.5**	98.6	**62.2**	59.6	**61.4**	**21.3**	**36.5**	37.2	63.8	47.3	**48.6**	45.4	0.095

4.5　算法稳定性实验

MI-FGSM类算法由于使用了动量项，因此能够稳定地更新样本数据[13]。为了验证插值方法的稳定性，以Inception-v3为白盒，研究迭代次数 T 值从2到20变化时，MI-FGSM与LIM-MI算法攻击成功率的变化情况，结果如图2所示。图2(a)表明，随着 T 值不断增大，两种算法在卷积模型上均能保持良好的稳定性，本文算法的稳定性更加突出。图2(b)表明，在ViTs模型上，随着 T 值不断增大，两种算法均出现一定程度的波动，但整体上本文算法保持了更好的稳定性。



(a) Source model: Inception-v3, Target model: CNNS modes



(b) Source model: Inception-v3, Target model: ViTs modes

图2　不同迭代次数下算法的攻击成功率

Fig. 2　Attack success rate of algorithms at different iteration counts

结束语　本文揭示了当前梯度符号方法在收敛性方面的固有难题，并提出了一种基于线性插值的新型对抗性攻击方法。理论上，该方法已被证明在一般凸问题中能够实现个体收敛，从而克服I-FGSM算法引入梯度符号导致的收敛性问题，从理论层面确保了迭代过程的稳定性。实验结果显示，在一定参数范围内，提高对抗性攻击的成功率是以牺牲对抗样本的不可察觉性为代价的。同时，大量的实验验证了线性插值方法作为一种高效策略，可广泛适用于基于梯度优化的各类对抗性攻击算法中，其能够替代传统的符号方法，充分利用梯度信息，在维持对抗样本不可察觉性的同时，显著提升其在不同模型间的迁移能力，并在迭代过程中保持良好的稳定性。此外，探究本文提出的线性插值方法在自然语言处理、语音识别等非视觉领域神经网络应用中的有效性，将是未来研究值得深入的方向。

参 考 文 献

[1] KIM M, JAIN A K, LIU X. AdaFace: Quality Adaptive Margin for Face Recognition[C]// 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 18729-18738.

[2] FENG S, SUN H W, YAN X T, et al. Dense reinforcement learning for safety validation of autonomous vehicles[J]. Nature, 2023, 615(7953): 620-627.

[3] WANG Y, YU J, ZHANG J. Zero-Shot Image Restoration Using Denoising Diffusion Null-Space Model[J]. arXiv: 2212.00490, 2022.

[4] HESSEL J, MARASOVIĆ A, HWANG J D, et al. Do Androids Laugh at Electric Sheep? Humor "Understanding" Benchmarks from The New Yorker Caption Contest[J]. arXiv: 2209.06293, 2022.

[5] GU J, JIA X, JORGE P D, et al. A Survey on Transferability of Adversarial Examples across Deep Neural Networks[J]. arXiv: 2310.17626, 2023.

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

[7] JI S L, DU T Y, DENG S G, et al. Robustness certification research on deep learning models: a survey[J]. Chinese Journal of Computers, 2022, 45(1): 190-206.

[8] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and Harnessing Adversarial Examples[J]. arXiv: 1412.6572, 2014.

[9] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial examples in the physical world[J]. arXiv: 1607.02533, 2016.

[10] CARLIN IN, WAGNER D A. Towards Evaluating the Robustness of Neural Networks[C]// 2017 IEEE Symposium on Security and Privacy. 2017: 39-57.

[11] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards Deep Learning Models Resistant to Adversarial Attacks[J]. arXiv: 1706.06083, 2019.

[12] TRAMÉR F, KURAKIN A, PAPERNOT N, et al. Ensemble Adversarial Training: Attacks and Defenses[J]. arXiv: 1705.07204, 2017.

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

[14] LIN J, SONG C, HE K, et al. Nesterov Accelerated Gradient and Scale Invariance for Adversarial Attacks[J]. arXiv: 1908.06281, 2019.

[15] XIE C, ZHANG Z, WANG J, et al. Improving Transferability of Adversarial Examples With Input Diversity[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 2725-2734.

[16] DONG Y, PANG T, SU H, et al. Evading Defenses to Transferable Adversarial Examples by Translation-Invariant Attacks[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 4307-4316.

[17] WANG J, CHEN Z, JIANG K, et al. Boosting the Transferability of Adversarial Attacks with Global Momentum Initialization[J]. arXiv: 2211.11236, 2022.

[18] WANG X, LIN J, HU H, et al. Boosting Adversarial Transferability through Enhanced Momentum[J]. arXiv: 2103.10609, 2021.

[19] PENG A, LIN Z, ZENG H, et al. Boosting Transferability of Adversarial Example via an Enhanced Euler's Method[C]// ICASSP 2023-2023 IEEE International Conference on A-

coustics, Speech and Signal Processing. 2023.
[20] GE Z, SHANG F, LIU H, et al. Boosting Adversarial Transferability by Achieving Flat Local Maxima[J]. arXiv: 2306.05225, 2023.
[21] FANG Z, WANG R, HUANG T, et al. Strong Transferable Adversarial Attacks via Ensembled Asymptotically Normal Distribution Learning[C]// CVPR2024. 2024.
[22] KARIMIREDDY S P, REBJOCK Q, STICH S U, et al. Error Feedback Fixes SignSGD and other Gradient Compression Schemes[J]. arXiv: 1901.09847, 2019.
[23] TAO W, PAN Z S, ZHU X H, et al. The Optimal individual convergence rate for the projected subgradient method with linear interpolation operation [J]. Journal of Computer Research and Development, 2017, 54(3): 529-536.
[24] MUKKAMALA M C, HEIN M. Variants of RMSProp and Adagrad with Logarithmic Regret Bounds [J]. arXiv: 1706.05507, 2017.
[25] KINGMA D P, BA J. Adam: A Method for Stochastic Optimization[J]. arXiv: 1412.6980, 2017.
[26] SITAWARIN C. New perspectives on adversarially robust machine learning systems: UCB-EECS-2024-10[R]. 2024.
[27] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet Large Scale Visual Recognition Challenge[J]. International Journal of Computer Vision, 2015, 115: 211-252.
[28] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.
[29] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the Inception Architecture for Computer Vision[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2818-2826.
[30] SIMONYAN K, ZISSERMAN A. Very Deep Convolutional Networks for Large-Scale Image Recognition[J]. arXiv: 1409.1556, 2014.
[31] HUANG G, LIU Z, WEINBERGER K Q. Densely Connected Convolutional Networks[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2261-2269.
[32] SANDLER M, HOWARD A G, ZHU M, et al. MobileNetV2: Inverted Residuals and Linear Bottlenecks[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 4510-4520.
[33] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale[J]. arXiv: 2010.11929, 2020.
[34] LIU Z, LIN Y, CAO Y, et al. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows[C]// 2021 IEEE/CVF International Conference on Computer Vision. 2021: 9992-10002.
[35] LIU Y, CHEN X, LIU C, et al. Delving into Transferable Adversarial Examples and Black-box Attacks[J]. arXiv: 1611.02770, 2016.
[36] BAO L, TAO W, TAO Q. Enhancing Adversarial Attack Transferability with an Adaptive Step-Size Strategy and Data Augmentation Mechanism[J]. Electronics Letters, 2024, 52(1): 157-169.

CHEN Jun, born in 1989, master. His main research interests include machine learning and mathematical optimization.

TAO Qing, born in 1965, Ph.D, professor, doctoral supervisor, is a senior member of CCF(No. 09081S). His main research interests include machine learning, pattern recognition and applied mathematics.

(责任编辑:何杨)