计算机科学
COMPUTER SCIENCE

自适应约束上界的对抗攻击优化方法

周强, 李哲, 陶蔚, 陶卿

引用本文

周强, 李哲, 陶蔚, 陶卿. 自适应约束上界的对抗攻击优化方法[J]. 计算机科学, 2026, 53(1): 404-412.

ZHOU Qiang, LI Zhe, TAO Wei, TAO Qing. Adaptive Box-constraint Optimization Method for Adversarial Attacks [J]. Computer Science, 2026, 53(1): 404-412.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.250600144

自适应约束上界的对抗攻击优化方法

周 强[1] 李 哲[1] 陶 蔚[2] 陶 卿[3,4]
1 陆军兵种大学防空兵学院 郑州 450000
2 军事科学院战略评估中心 北京 100091
3 陆军兵种大学信息工程学院 合肥 230000
4 合肥理工学院 合肥 238076
(1071391319@qq.com)

摘 要 深度神经网络易受对抗样本攻击。现有迁移攻击优化方法普遍使用固定的约束上界表示不可察觉性强度,重点关注如何提升攻击成功率,忽略了样本间的敏感性差异,导致不可察觉性(FID)效果有待提高。受自适应梯度方法的启发,以提高不可察觉性为主要目的,提出了一种自适应约束上界的对抗攻击优化方法。首先,通过梯度幅值建立敏感性指标,量化不同样本的敏感性差异程度;在此基础上,自适应确定对抗攻击优化方法的约束上界,实现敏感样本低强度、非敏感样本高强度对抗扰动的差异化处理;最后,通过替换投影算子和步长,将自适应约束机制无缝集成至现有攻击方法。ImageNet-Compatible 数据集上的实验表明,所提方法在相同的黑盒攻击成功率下,FID 较传统固定约束方法降低 2.68%~3.49%;基于该方法的 MI-LA 对抗攻击算法较对抗攻击领域表现优异的 5 种攻击方法,FID 降低 6.32%~26.35%。

关键词:对抗攻击;自适应;约束上界;样本敏感性;黑盒迁移性;不可察觉性
中图分类号 TP391

Adaptive Box-constraint Optimization Method for Adversarial Attacks

ZHOU Qiang[1], LI Zhe[1], TAO Wei[2] and TAO Qing[3,4]
1 College of Air Defense, Army Branch University, Zhengzhou 450000, China
2 Institute of Evaluation and Assessment Research, PLA Academy of Military Science, Beijing 100091, China
3 College of Information Engineering, Army Branch University, Hefei 230000, China
4 Hefei Institute of Technology, Hefei 238076, China

Abstract Deep neural networks are vulnerable to adversarial example attacks. Existing transfer-based attack optimization methods commonly employ fixed constraint upper bounds to represent imperceptibility intensity, focusing primarily on improving attack success rates. However, such approaches overlook inter-sample sensitivity variations, resulting in suboptimal imperceptibility (measured by Fréchet Inception Distance, FID). Inspired by adaptive gradient methods, this paper proposes an adversarial attack optimization method with adaptive constraint upper bounds, aiming to enhance imperceptibility. Firstly, a sensitivity metric based on gradient magnitudes is established to quantify sensitivity differences across samples. Building on this, adaptive constraint upper bounds are determined to enable differentiated perturbation handling — applying low-intensity perturbations to sensitive samples and high-intensity perturbations to non-sensitive ones. Furthermore, by replacing the projection operator and step size, the adaptive constraintmechanism is seamlessly integrated into existing attack methods. Experiments on the ImageNet-Compatible dataset demonstrate that, under equivalent black-box attack success rates, the proposed method reduces FID by 2.68%~3.49% compared to traditional fixed-constraint methods. Additionally, the MI-LA attack algorithm based on this approach achieves 6.32%~26.35% lower FID than five state-of-the-art adversarial attack methods.

Keywords Adversarial attack, Adaptive, Upper bound constraint, Sample sensitivity, Black-box transferability, Imperceptibility

1 引言

深度神经网络(Deep Neural Networks, DNNs)在计算机视觉、自然语言处理、医疗影像分析等领域展现了革命性的性能突破,成为现代人工智能技术的核心驱动力。其脆弱性亦引发了广泛关注。研究表明,DNNs 易受对抗样本攻击,攻击

到稿日期:2025-06-20 返修日期:2025-09-08
基金项目:国家自然科学基金(62076252,62576351)
This work was supported by the National Natural Science Foundation of China(62076252,62576351).
通信作者:陶卿(taoqing@gmail.com)

者通过向输入数据注入人类难以察觉的微小扰动，即可导致模型输出高置信度的错误结果[1-2]。例如，在自动驾驶场景中，对抗样本可能使目标检测模型忽略交通标志[3]；在医疗诊断中，对抗样本会误导影像分类结果[4]。此类安全问题不仅会威胁模型可靠性，还会为实际应用带来重大风险。因此，提升对抗攻击的效能并探索其防御机制，成为当前的研究热点。

当前主流的对抗攻击方法基于梯度优化框架，通过固定振动上界表征不可察觉性，利用梯度更新生成扰动，实现攻击目标。其中，快速梯度符号方法（Fast Gradient Sign Method，FGSM）[1]和投影梯度下降方法（Projected Gradient Descent，PGD）[5]被视为基准方法。FGSM 通过单步梯度更新生成扰动；PGD 则通过多步迭代优化扰动，增强攻击效果。尽管这些方法在攻击成功率（Attack Success Rate，ASR）上表现优异，但其核心缺陷在于采用全局统一的扰动约束（如固定 $L\infty$ 或 L2 范数边界）。这种“一刀切”策略忽略了不同样本对扰动的敏感性差异，导致不可察觉性有待提高。如图 1 所示，某些样本（如语义信息模糊的领结）仅需微弱扰动（$\varepsilon=1/255$）即可诱导 ResNet-152 模型误判，而另一些样本（如语义信息明确的 T 型车）需显著扰动（$\varepsilon=6/255$）方可达到近似效果。这一现象表明，固定约束策略将导致部分样本扰动强度冗余或不足，难以平衡攻击效能与隐蔽性。

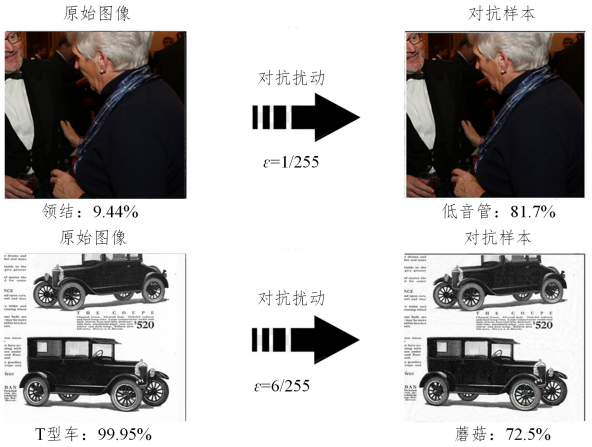


图 1　样本敏感性差异示意图

Fig. 1　Schematic diagram of sample sensitivity differences

针对上述问题，本文受自适应梯度方法[6]（AdaGrad）的启发，提出动态约束机制。AdaGrad 通过累积历史梯度信息为不同参数分配差异化学习率，对梯度幅值小的维度使用较大的步长，而对梯度幅值大的维度使用较小的步长，从而显著提升了优化过程的收敛效率与泛化能力。类比其核心思想，本文发现，对抗扰动的生成须考虑样本敏感性差异：对于高敏感性样本（如领结图片），低强度扰动即可破坏模型的特征表示；而对于低敏感性样本（如 T 型车），则需增强扰动强度以维持攻击的有效性。基于此，本文设计了一种基于梯度幅值的敏感性量化策略，动态调整扰动约束的上界，实现“敏感样本低约束、非敏感样本高约束”的差异化处理。该方法在相同攻击成功率的条件下，能够有效提升对抗样本的不可察觉性。

不可察觉性是衡量对抗攻击隐蔽性的核心指标。在相同的攻击成功率下，不可察觉性越好意味着越容易绕过防御机制。传统方法通常依赖 Lp 范数（如 L2/L_∞ 范数）度量扰动幅度，但此类方法难以反映人类视觉系统[7]（Human Visual System，HVS）对图像失真的感知特性。近年来，基于深度特征的 Fréchet Inception Distance[8]（FID）被广泛应用于对抗样本评估，通过计算对抗样本与原始样本在特征空间中的分布差异，提供更符合人类感知的不可察觉性度量。然而，现有攻击方法在优化过程中往往忽略了 FID 的指导作用，导致生成的对抗样本虽满足像素级约束，但在特征空间中仍与原始数据分布存在显著偏移。本文将 FID 指标作为度量不可觉察性的指标，更加客观地反映对抗攻击算法的视觉隐蔽性。

现有基于显著性分析的攻击方法（如 JSMA[9]）通过雅可比矩阵定位关键像素，但其计算复杂度为 $O(d)$（d 为输入维度）难以扩展至高分辨率图像（如 $299\times299\times3$）；C&W[10]等方法虽引入了动态优化，但约束强度仍与样本特性解耦。此类方法的共性局限在于：未建立样本敏感性与扰动约束的定量关联模型，导致扰动生成过程缺乏自适应性。针对上述挑战，本文提出三阶段优化框架。1）敏感性量化：通过单步反向传播提取梯度幅值，构建样本敏感性指标，揭示扰动阈值的内在规律。2）动态约束映射：设计敏感性驱动的约束系数，实现“低敏感-高约束、高敏感-低约束”的差异化扰动分配。3）自适应约束攻击：通过替换投影算子和步长实现动态约束，将 AdaConst 方法无缝集成至基于梯度的攻击算法。

本文主要贡献包括：1）首次建立样本敏感性与对抗约束强度的定量关联模型，突破固定约束假设，为自适应攻击提供可解释性理论支撑；2）提出梯度幅值显著性近似方法，其较 JSMA 降低了 $O(d)$ 的计算复杂度，突破了高分辨率场景的应用瓶颈；3）所提方法可无缝集成至 FGSM，PGD 和 MI-FGSM 等主流攻击框架，仅需替换投影算子和步长即可实现动态约束，具备工程普适性。

2　相关工作

2.1　基于优化算法的攻击

2014 年，Szegedy 等[11]首次提出了一种名为对抗样本攻击的方法（L-BFGS），其原理是在输入数据中引入微小的变化，导致模型预测时出现高置信度的错误结果。设 f 为目标模型，攻击者的目的是创造尽可能小的扰动 $\boldsymbol{\eta}$，用以生成对抗性样本 $\boldsymbol{x}^{\mathrm{adv}}=\boldsymbol{x}+\boldsymbol{\eta}$，使得模型对这些样本做出错误的分类。这种攻击的数学表述如下：

$$\min_{x^{\mathrm{adv}}} \| \boldsymbol{x}^{\mathrm{adv}}-\boldsymbol{x} \|_p \quad \text{s.t. } f(\boldsymbol{x})=l, f(\boldsymbol{x}^{\mathrm{adv}})=l', l\neq l', \boldsymbol{x}\in[0,1]^m \tag{1}$$

其中，$\| \boldsymbol{x} \|_p$ 表示 Lp 范数，用于度量对抗样本与干净样本之间的距离。

L-BFGS 的优化流程繁琐，算法的效率不高。Carlini 和 Wagner[10]在此基础上提出了式（1）的一个变体，鉴于分类函数的非线性特性，对其进行替换，确定了攻击的优化目标：

$$\min_{\eta} \| \boldsymbol{\eta} \|_p + c \cdot f(\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{\eta}) \tag{2}$$

2.2 基于梯度的攻击

基于梯度的对抗攻击方法通过约束扰动空间实现隐蔽攻击。Goodfellow 等[1]受到深度神经网络线性假设的启发，提出了一种名为 FGSM 的攻击技术，该技术通过计算损失函数的梯度方向迅速生成有效的扰动。其扰动定义为：

$$\boldsymbol{\eta}=\varepsilon \cdot sign(\nabla_x \mathcal{L}(\boldsymbol{\theta};\boldsymbol{x},y)) \tag{3}$$

其中，ε 负责调控对抗扰动的强度，而 $sign$ 函数则用于指示梯度的正负方向。这种攻击方式仅需计算梯度向量，大大缩短了生成对抗样本的时间，但会导致一定的攻击精度损失[12]。

为了提升攻击的迁移性，Dong 等[13]提出了一种新的方法——动量迭代快速梯度符号方法(MI-FGSM)，其更新方式如下：

$$\boldsymbol{x}_0^{\text{adv}}=\boldsymbol{x},\boldsymbol{g}_0=0 \tag{4}$$

$$\boldsymbol{g}_{t+1}=\mu\boldsymbol{g}_t+\frac{\nabla\mathcal{L}(\boldsymbol{x}_t^{\text{adv}},y)}{\|\nabla\mathcal{L}(\boldsymbol{x}_t^{\text{adv}},y)\|_1} \tag{5}$$

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

其中，$Clip\{\cdot\}$为裁剪函数，用于裁剪超出约束范围的数值。实验结果表明，通过利用梯度动量来增强 FGSM 的迭代攻击，显著提高了黑盒攻击的成功率。

Madry 等对 I-FGSM[14]攻击进行了深化研究，通过采用 PGD[6]技术来探索输入数据在 p 范数球范围内的扰动。该方法首先对样本进行初始化：

$$\boldsymbol{x}_0^{\text{adv}}=\boldsymbol{x}+\boldsymbol{\delta}_0 \tag{7}$$

其中，$\boldsymbol{\delta}_0\sim\mathcal{U}[-\varepsilon,\varepsilon]^d$。其迭代式为：

$$\boldsymbol{g}_{t+1}=\nabla_{x_t^{\text{adv}}}\mathcal{L}(f(\boldsymbol{x}_t^{\text{adv}}),y) \tag{8}$$

$$\boldsymbol{x}_{t+1}^{\text{adv}}=Clip_{x,\varepsilon}(\boldsymbol{x}_t^{\text{adv}}+\alpha \cdot sign(\boldsymbol{g}_{t+1})) \tag{9}$$

由于 PGD 攻击能够在整个 p 范数球内生成扰动，因此能够有效抵御 PGD 攻击，即等同于能够全面防御 p 范数球内所有潜在的对抗样本。

2.3 基于显著性分析的攻击

Papernot 等[9]开创了一种基于显著性分析的攻击方法，即 JSMA 攻击。该方法定义像素 x_i 的显著性如下：

$$s(x_i)=\left(\frac{\partial F_t(\boldsymbol{x})}{\partial x_i}\right)\cdot\sum_{j\neq t}\left(-\frac{\partial F_j(\boldsymbol{x})}{\partial x_i}\right) \tag{10}$$

其中，$F_t(\boldsymbol{x})$表示模型输出为类别 t 的 logit。通过这种方式，JSMA 攻击能够精确定位并操纵那些对模型输出影响最大的像素，从而有效地生成对抗性样本。然而，这种方法的计算成本相对较高。Oseledets 等[15]也利用雅可比矩阵进行创新，计算特征图的奇异值向量，提出了一种名为 SV-UAP 的方法，用以构建通用对抗扰动。这种方法不仅能够针对单个模型生成有效的对抗样本，而且能在多个模型之间实现迁移，增强了对抗样本的通用性。

3 自适应约束上界方法

为改善对抗攻击算法的不可察觉性，本文提出了基于自适应约束上界的对抗攻击优化算法。首先，根据样本像素点的梯度幅值，构建样本敏感性指标；然后，基于样本的敏感性指标，建立动态约束映射机制，实现“低敏感-高约束、高敏感-低约束”的差异化扰动分配；最后，通过替换投影算子和步长，将 AdaConst 方法无缝集成至基于梯度的攻击算法。

3.1 样本敏感性分析

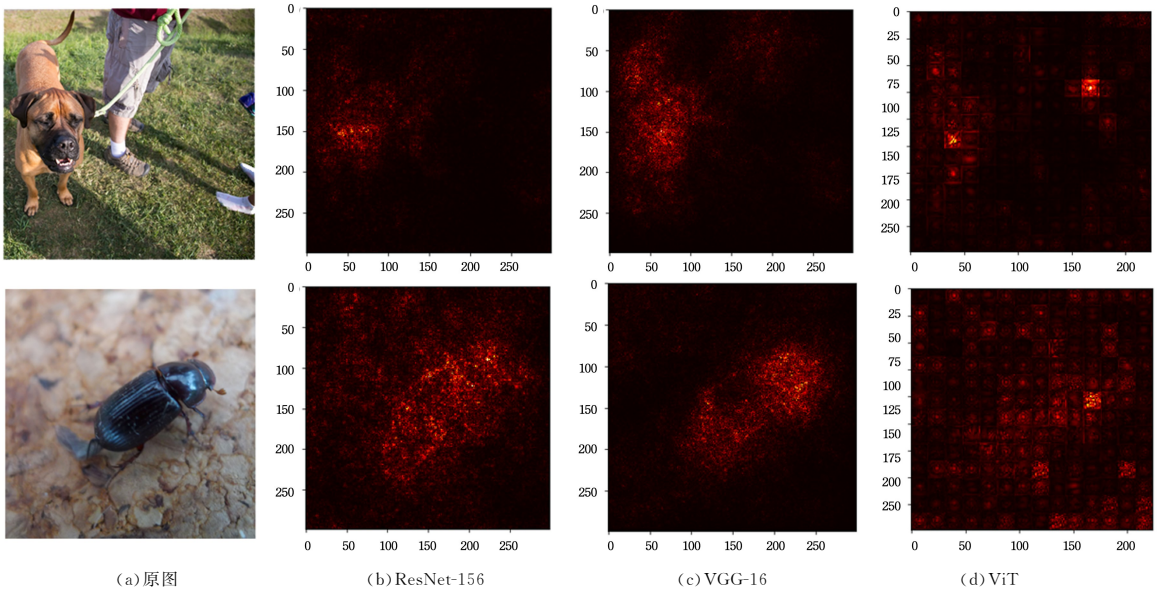
样本敏感性表示样本微小扰动对模型推理结果的影响程度。本文基于像素级显著性指标构建样本敏感性度量方法。

1)像素显著性计算。现有方法(如 JSMA 攻击)通过雅可比矩阵生成显著图，但其计算复杂度较高。为降低计算开销，本文提出简化的显著性计算方法：

$$s(x_i)=\left|\frac{\partial\mathcal{L}(f(\boldsymbol{x}),y)}{\partial x_i}\right| \tag{11}$$

其中，x_i 为样本的第 i 个像素点，f 为模型，$\mathcal{L}$为交叉熵损失函数，y 为目标标签(有目标攻击)或真实标签(无目标攻击)。

基于式(11)计算不同模型(如 ResNet-152，VGG-16，ViT 等)下样本像素点的显著性，并绘制相应的热力图(见图 2)。



(a)原图 (b)ResNet-156 (c)VGG-16 (d)ViT

图 2 跨模型显著性热力图的对比

Fig. 2 Comparison of cross model significant thermograms

热力图显示，不同模型的高显著性区域均聚焦于目标对象（如动物主体），验证了特征提取的一致性。特别地，ViT 的显著性呈块状分布，这源于其将输入图像分割为 16×16 像素块进行注意力计算的特性[16]。总体而言，ViT 模型的高亮区域仍然存在于目标附近，敏感性分析同样适用。

2）样本敏感性量化方法。基于像素显著性，定义样本敏感性为：

$$S(\boldsymbol{x})=\|s(\boldsymbol{x})\|_2 \tag{12}$$

其中，$\|\cdot\|_2$ 表示 L2 范数，$s(\boldsymbol{x})$ 表示样本各像素的显著性矩阵。通过对比 L1，L2 和 L_∞ 范数（见图 3）发现，3 种度量方法具有一致性。其中，L2 范数对异常值敏感度适中，故选定其为敏感性指标。

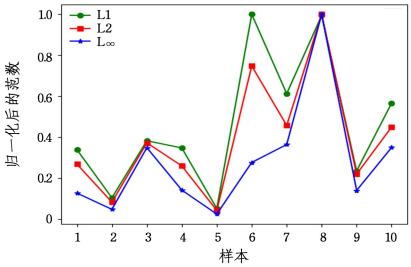


图 3　归一化范数敏感性的对比

Fig. 3　Comparison of normalized norm sensitivity

3.2　自适应约束强度设计

1）敏感性分布分析。在 ImageNet-Compatible 数据集（$I=1000$）上统计样本敏感性，其分布近似泊松分布（见图 4）。

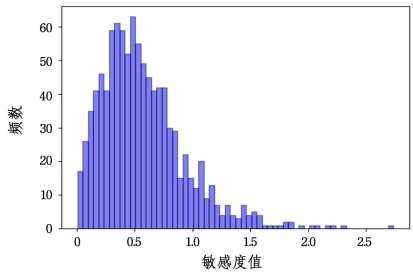


图 4　样本敏感性分布直方图

Fig. 4　Histogram of sample sensitivity distribution

为消除长尾效应，设定截断阈值：

$$S_{\text{cut}}=S_{\lceil I\cdot\tau\rceil} \tag{13}$$

其中，τ 为截断系数，S_k 为升序排列后的第 k 个敏感性值。对于大于阈值的予以截断：

$$S_i'=\min(S_{\text{cut}},S_i) \tag{14}$$

2）约束系数映射。将截断后的样本敏感性值线性映射至约束系数空间：

$$c_i=\theta+(1-\theta)\cdot\frac{S_{\text{cut}}-S_i'}{S_{\text{cut}}-S_{\min}} \tag{15}$$

其中，θ 为约束系数的下界，$S_{\min}$ 为最小敏感性值。

本节以 FGSM 算法为框架，以 ResNet-152 模型为替代模型，以 VGG-16 和 ViT 等模型为待攻击模型，系统研究了参数 θ 对黑盒攻击成功率、不可察觉性指标 FID 的影响。实验结果如图 5 所示。

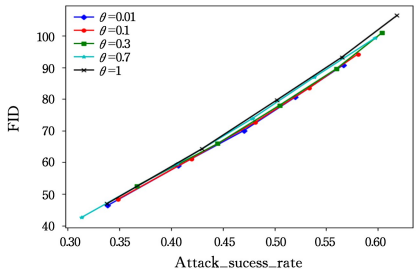


图 5　参数 θ 对算法性能的影响

Fig. 5　Influence of parameter θ on algorithm performance

由图 5 可知：当对抗扰动 ε 较小时，θ 对攻击算法性能的影响不大；当对抗扰动 ε 较大时，θ 的效果逐渐明显。该现象与理论分析是一致的，即当对抗扰动较小时，样本均未达到成功攻击的阈值，对约束上界进行调整的效果并不明显；当对抗扰动增大时，部分高敏感样本出现扰动冗余，对约束上界进行调整会得到更好的效果。总之，扰动越大，自适应约束上界方法的效果越好。另外，当 θ 大于 0.1 时，θ 越小，攻击效果越好；当 θ 小于 0.1 时，效果变化不再明显。综上所述，设计 θ 值的变化如下：

$$\theta=\begin{cases}1, & \varepsilon<0.004\\ \dfrac{0.004}{\varepsilon}, & 0.004\leqslant\varepsilon\leqslant0.04\\ 0.1, & \varepsilon>0.04\end{cases} \tag{16}$$

其中，ε 为无穷范数距离；L1 范数的大致对应关系为 $2\varepsilon_1/d$，L2 范数的大致对应关系为 $\varepsilon_2/\sqrt{d}$。

算法 1 完整描述了自适应约束上界方法的步骤。

算法 1　AdaConst

输入：模型$\{f_j\}$，样本集$\{\mathbf{x}_i\}$，标签集$\{y_i\}$；截断系数 τ

输出：约束系数$\{c_i\}$

1. for i=0 to I−1 do
2.　for j=0 to J−1 do
3.　　计算交叉熵损失$\mathcal{L}(f_j(x_i),y_i)$
4.　　计算样本敏感性$S_i^{(j)}\leftarrow\|\nabla\mathcal{L}(f_j(\mathbf{x}_i),y_i)\|_2$
5.　end for
6.　多模型集成求平均$\bar{S}_i\leftarrow(\sum_j S_i^{(j)})/J$
7. end for
8. 从小到大排序$\{\bar{S}_i\}$得$\{\bar{S}_{(k)}\}$
9. 计算截断阈值$S_{\text{cut}}\leftarrow\bar{S}_{\lceil I*\tau\rceil}$
10. 截断长尾：$S_i'\leftarrow\min(S_{\text{cut}},\bar{S}_i)$
11. 由式(16)计算映射系数 θ
12. 返回样本约束系数 $c_i=\theta+(1-\theta)\cdot\dfrac{S_{\text{sut}}S_i'}{S_{\text{cut}}-S_{\min}}$

3.3　基于自适应约束上界的对抗攻击方法

自适应约束上界方法适用于所有基于约束优化的对抗攻击算法。本节以 MI-L1mask[17] 为例介绍其拓展方法，原方法迭代方式如下：

$$\boldsymbol{x}_{t+1}^{\text{adv}}=P_{x,\varepsilon_1}\{\boldsymbol{x}_t^{\text{adv}}+\alpha\cdot sign(\mathcal{G}_{t+1})\odot\boldsymbol{m}\} \tag{17}$$

其中，P_{x,ε_1} 表示向以 $\boldsymbol{x}$ 为中心、ε_1 为半径的 L1 范数球投影，$\boldsymbol{m}$ 表示遮盖矩阵。MI-L1mask 的自适应约束版本 MI-L1mask-AdaConst（简称 MI-LA）将式(17)换成自适应约束：

$$\boldsymbol{x}_{t+1}^{\text{adv}}=P_{x,\varepsilon_1\cdot c}\{\boldsymbol{x}_t^{\text{adv}}+\alpha\cdot c\cdot sign(\boldsymbol{g}_{t+1})\odot\boldsymbol{m}\} \tag{18}$$

即在每步迭代中，迭代步长和投影算子都乘自适应约束系数 c，以实现扰动距离的差异化处理。对于高敏感样本，扰动距离减小，不可察觉性得到改善。

MI-LA 算法的具体步骤如算法 2 所示。

算法 2 MI-LA

输入：模型 f，干净样本 x，标记 y，约束系数 c，扰动 ε

默认参数：损失函数 $\mathcal{L}$ 为交叉熵，迭代步数 steps=10，动量系数 $\mu=1$

输出：对抗样本 $\mathbf{x}^{adv}$，满足 $\|\mathbf{x}^{adv}-\mathbf{x}\|_1 \leqslant \varepsilon \cdot c$。

1. 计算步长 $\alpha=\sqrt{\frac{51\varepsilon \cdot c}{steps \cdot d}}$

2. 计算 mask 系数 $\lambda=\sqrt{\frac{\varepsilon \cdot c}{12.75\times steps \cdot d}}$

3. 初始化：$\mathbf{g}_0=0$；$\mathbf{x}_0^{adv}=0$

4. for t=0 to T−1 do

5. 计算梯度 $\mathbf{g}\leftarrow\nabla_x\mathcal{L}(\mathbf{x}_t^{adv},y)$

6. if t=0 do

7. 计算各像素点显著性并排列 $s_1<s_2<\cdots<s_d$

8. 计算阈值 $\tau=s_{d\times(1-\lambda)}$

9. 计算遮盖矩阵 $\mathbf{m}\leftarrow q_i=\begin{cases}1, & s_i>\tau\\ 0, & s_i<\tau\end{cases}$

10. end if

11. 更新动量 $\mathbf{g}_{t+1}\leftarrow\mu\cdot\mathbf{g}_t+\frac{\mathbf{g}}{|\mathbf{g}|_1}$

12. 更新 $\mathbf{x}_{t+1}^{adv}\leftarrow\mathbf{x}_t^{adv}+\alpha\cdot c\cdot sign(\mathbf{g}_{t+1})\odot\mathbf{m}$

13. 向 L1 范数球投影 $\mathbf{x}_{t+1}^{adv}\leftarrow P_{x,\varepsilon\cdot c}\{\mathbf{x}_{t+1}^{adv}\}$

14. end for

4 实验及分析

为系统评估自适应约束上界方法的有效性，本文首先设计并开展消融实验。选取 FGSM，MI-FGSM 和 PGD 这 3 类经典对抗攻击方法作为基线方法，通过 AdaConst 框架对原始方法进行改进，对比分析改进前后在攻击成功率和不可察觉性等性能指标的差异性。在此基础上，进一步将 MI-LA 方法与近年提出的 APGD，Autoattack，Admix，FIA 和 DeCoWA 等对抗攻击方法进行横向对比。

4.1 实验设置

1）数据集。遵循对抗训练的通用范式，实验采用 Image Net-Compatible 数据集验证自适应约束上界方法的攻击效能。该数据集与 ImageNet 结构兼容，包含 1000 张尺寸为 299×299×3 的 RGB 图像，被广泛用于评估对抗攻击的迁移性，为比较对抗攻击和防御算法提供了公认基准。

2）模型。实验选取 5 种经典卷积神经网络（CNN）和 1 种视觉 Transformer（ViT）构成基准模型。

（1）卷积神经网络：Inception-v3（Inc-v3，改进的 Inception 模块）[18]，GoogLeNet[19]（经典 Inception 模块），ResNet-152（Res-152，深度残差学习的标杆）[20]，VGG-16[21]（经典深层网络），MobileNet-v2（Mob-v2，轻量化模型代表）[22]。所有 CNN 模型基于 Torchvision 库加载预训练参数。

（2）视觉 Transformer：ViT_base_patch16_224（ViT-B，Transformer 在 CV 领域成功应用的经典模型）[23]。基于 Timm 库加载预训练参数。实验设置中，将 ResNet-152 作为黑盒攻击替代模型，其余 5 种模型作为黑盒攻击目标模型。

3）实现细节。实验环境搭建与算法实现具体配置如下。

（1）算法框架：基于 PyTorch[24] 深度学习框架构建实验系统。

（2）攻击方法：采用 Torchattacks 库[25] 实现基准对抗攻击算法，并在此基础上改进约束方法。通过 pytorch_fid 库计算 FID。

（3）硬件配置：所有实验在 NVIDIA RTX A6000 GPU 平台完成。

4）评价标准。实验从两个维度评估对抗样本的质量。

（1）迁移性：通过黑盒攻击成功率（Attack Success Rate，ASR）度量。

$$ASR=\frac{N_{succ}}{N_{total}}\times 100\% \tag{17}$$

其中，N_{succ} 为成功攻击的样本数，N_{total} 为总样本数。ASR 值越高，表明迁移性越强。

（2）不可察觉性：采用 FID 度量，FID 值越小，表明对抗扰动的隐蔽性越好。

4.2 消融实验

为验证自适应约束方法对攻击算法的效果，本节选取 FGSM，MI-FGSM 和 PGD 这 3 种方法，使用 AdaConst 算法自适应获取样本的约束强度，对比原方法与使用 AdaConst 方法的性能差异，研究在相同 ASR 下，FID 的变化情况。

1）FGSM

按照 FGSM 算法的默认设置，扰动大小 ε 从 4 到 28 等间隔变化。两种攻击方法的性能对比如表 1 所列。根据表 1 的实验数据绘制 ASR 与 FID 的距离关系图，如图 6 所示。可以看出，在相同的 ASR 情况下，FGSM-AdaConst 的不可察觉性指标 FID 小于原算法。

表 1 FGSM 采用 AdaConst 后的性能对比

Table 1 Performance comparison of FGSM using AdaConst

FGSM			FGSM-AdaConst		
eps	*ASR*/%	*FID*	*eps*	*ASR*/%	*FID*
4	22.10	23.31	4	17.62	15.57
8	33.80	46.93	8	26.68	33.36
12	42.98	64.24	12	34.84	48.28
16	50.20	79.61	16	41.94	61.09
20	56.50	93.17	20	48.1	72.71
24	61.84	106.38	24	53.32	83.42
28	65.92	118.48	28	58.06	94.18

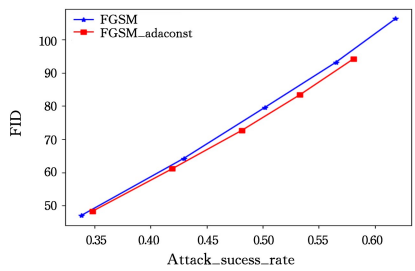


图 6 自适应约束方法对 FGSM 算法的影响

Fig. 6 Impact of adaptive constraint methods on FGSM

用线性插值法计算 ASR 为 40%，45%，50%，55%，60% 时两种方法的 FID，结果如表 2 所列。可以看出，在相同的 ASR 下，改进方法的 FID 值较原方法下降了约 2.68%。

表 2　FGSM 线性插值后相同 ASR 对应 FID 的下降情况

Table 2　Decrease of FID corresponding to the same ASR after FGSM linear interpolation

方法	*ASR*					*FID* 下降/%
	0.40	0.45	0.50	0.55	0.60	
FGSM	58.62	68.54	79.19	89.95	101.83	—
FGSM-AdaConst	57.59	66.81	76.53	87.23	99.03	↓2.68

2）MI-FGSM

按照 MI-FGSM 算法的默认设置，取攻击步数 $step=10$，步长 $\alpha=\varepsilon/4$，扰动大小 ε 从 4 到 28 等间隔变化。两种攻击方法的性能对比如表 3 所列。根据表 3 的实验数据绘制黑盒攻击成功率 ASR 与 FID 的距离关系图，如图 7 所示。

表 3　MI-FGSM 采用 AdaConst 后的性能对比

Table 3　Performance comparison of MI-FGSM using AdaConst

MI-FGSM			MI-FGSM-AdaConst		
eps	*ASR*/%	*FID*	*eps*	*ASR*/%	*FID*
4	23.16	25.74	8	31.18	39.67
8	40.26	57.18	12	42.48	60.60
12	52.04	82.27	16	51.62	78.73
16	60.78	103.79	20	59.24	95.13
20	67.34	122.1	24	65.40	112.10
24	72.22	139.12	28	68.56	125.54

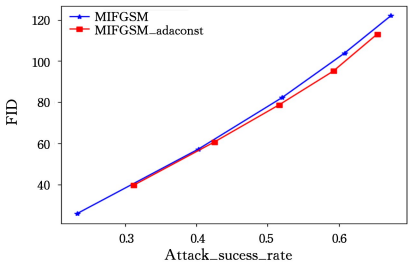


图 7　自适应约束方法对 MI-FGSM 算法的影响

Fig. 7　Impact of adaptive constraint methods on MI-FGSM

用线性插值法计算 ASR 为 40%，45%，50%，55%，60% 时两种方法的 FID，结果如表 4 所列。可以看出，在相同的 ASR 下，改进方法的 FID 值较原方法下降了约 3.04%。

表 4　MI-FGSM 线性插值后相同 ASR 对应 FID 的下降情况

Table 4　Decrease of FID corresponding to the same ASR after MI-FGSM linear interpolation

方法	*ASR*					*FID* 下降/%
	0.40	0.45	0.50	0.55	0.60	
MI-FGSM	56.70	67.28	77.93	89.56	101.87	—
MI-FGSM-AdaConst	56.01	65.60	75.52	86.01	97.35	↓3.04

3）PGD

按照 PGD 算法的默认设置，取攻击步数 step＝10，步长 $\alpha=\varepsilon/4$，扰动大小 ε 从 4 到 32 等间隔变化。两种攻击方法的性能对比如表 5 所列。根据表 5 的实验数据绘制黑盒攻击成功率与 FID 的距离关系图，如图 8 所示。

表 5　PGD 采用 AdaConst 后性能对比

Table 5　Performance comparison of PGD using AdaConst

PGD			PGD-AdaConst		
eps	*ASR*/%	*FID*	*eps*	*ASR*/%	*FID*
4	18.28	17.22	8	22.20	25.86
8	28.72	38.84	12	29.48	40.27
12	37.36	57.27	16	36.80	53.72
16	45.62	75.93	20	42.54	66.15
20	51.74	91.42	24	48.64	81.67
24	57.18	106.55	28	54.24	93.45
28	62.72	119.82	32	58.00	103.64

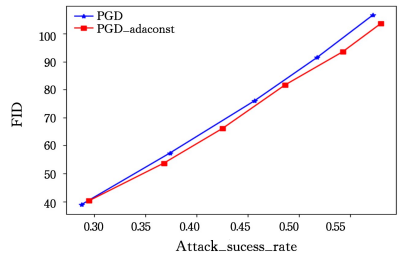


图 8　自适应约束方法对 PGD 算法的影响

Fig. 8　Impact of adaptive constraint methods on PGD

用线性插值法计算 ASR 为 40%，45%，50%，55%，60% 时两种方法的 FID，结果如表 6 所列。可以看出，在相同的 ASR 下，改进方法的 FID 值较原方法下降了约 3.49%。

表 6　PGD 线性插值后相同 ASR 对应 FID 的下降情况

Table 6　Decrease of FID corresponding to the same ASR after PGD linear interpolation

方法	*ASR*					*FID* 下降/%
	0.40	0.45	0.50	0.55	0.60	
PGD	63.23	74.53	87.02	100.49	106.55	—
PGD-AdaConst	60.65	72.41	84.53	95.51	103.64	↓3.49

4.3　对比实验

为验证 MI-LA 算法的效果，本节选取近年来表现优异的 5 种对抗攻击算法作为对比方法，在相同条件下进行测试。这 5 种对比算法分别是 APGD[26]，Autoattack[27]，Admix[28]，FIA[29] 和 DeCoWA[30]。各算法均按默认参数设置，6 种攻击方法的性能对比如表 7 所列。

表 7　不同对抗攻击算法黑盒攻击成功率与 FID 距离的对比

Table 7　Comparison of black box attack success rate and FID distance under different adversarial attack algorithms

攻击方法	*eps*	攻击目标模型						黑盒攻击成功率/%	*FID* 距离
		Inc-v3	Googlenet	VGG-16	Res-152	Mob-v2	ViT-B		
APGD[27]	4	6.20	13.50	21.70	68.30	23.30	7.80	14.50	12.90
	8	12.70	21.30	32.10	70.90	37.20	8.40	22.34	29.08
	12	18.80	31.60	43.30	71.10	47.50	8.20	29.88	43.80
	16	23.90	39.50	48.50	71.10	55.10	9.20	35.24	53.57
	20	29.80	47.60	52.80	71.10	59.50	11.30	40.20	65.36
	24	34.90	53.20	55.30	71.10	63.10	11.50	43.60	73.90
	30	41.60	58.00	60.40	71.10	67.40	12.90	48.06	88.55
	40	50.80	63.50	65.90	71.10	70.60	19.00	53.96	106.26

（续表）

攻击方法	*eps*	攻击目标模型						黑盒攻击成功率/%	*FID* 距离
		Inc-v3	Googlenet	VGG-16	Res-152	Mob-v2	ViT-B		
Autoattack-L1[28]	40 万	12.90	23.80	34.60	69.90	33.80	7.50	22.52	26.98
	80 万	18.90	33.00	43.40	70.90	44.70	8.60	29.72	40.48
	160 万	33.90	51.20	55.30	71.10	60.30	10.30	42.20	66.88
	300 万	51.00	64.40	65.90	71.10	69.60	16.40	53.46	105.02
	450 万	59.90	69.40	70.30	71.10	72.60	26.70	59.78	143.86
Admix[29]	4	15.30	24.00	30.10	89.00	35.00	10.30	22.94	22.99
	8	31.10	43.20	49.50	98.30	57.40	12.80	38.80	54.17
	12	46.10	59.60	65.10	99.50	71.20	15.00	51.40	80.40
	16	59.50	73.70	73.00	99.90	80.70	18.30	61.04	102.01
	20	67.00	82.50	79.00	99.90	85.70	22.10	67.26	122.46
	24	73.10	86.20	84.30	100.00	88.60	26.30	71.70	137.32
FIA[30]	4	14.50	22.70	29.00	92.00	34.40	9.80	22.08	23.56
	8	30.30	42.10	52.60	99.40	56.60	13.00	38.92	55.34
	12	45.00	57.40	65.10	99.80	69.70	14.10	50.26	82.22
	16	58.10	72.20	73.40	99.90	79.70	18.00	60.28	101.78
	20	66.40	78.00	79.60	100.00	84.80	22.20	66.20	120.04
	24	74.20	84.40	84.90	100.00	88.70	26.30	71.70	134.79
DeCoWA[31]	2	12.60	20.10	22.00	25.10	26.60	10.40	18.34	13.56
	4	24.70	37.60	37.80	57.50	45.10	12.30	31.50	38.71
	8	53.40	70.50	65.50	91.10	75.20	18.30	56.58	87.42
	12	73.40	84.60	80.30	97.80	87.10	26.00	70.28	119.36
	16	83.50	92.70	86.30	99.00	92.10	33.10	77.54	141.28
	20	88.50	95.40	92.10	99.70	95.00	39.70	82.14	156.96
	24	92.70	96.80	93.30	99.80	96.80	48.10	85.54	172.99
MI-LA	10 万	**12.70**	**20.40**	**25.40**	**57.10**	**25.80**	**8.30**	**18.34**	**13.33**
	30 万	**21.50**	**28.40**	**33.30**	**84.10**	**36.70**	**10.10**	**26.00**	**27.69**
	90 万	**41.80**	**46.80**	**52.00**	**97.70**	**54.10**	**14.00**	**41.74**	**54.16**
	150 万	**55.90**	**61.40**	**62.50**	**98.40**	**65.70**	**18.40**	**52.78**	**74.36**
	250 万	**67.30**	**75.60**	**75.50**	**99.20**	**77.30**	**24.00**	**63.94**	**100.48**
	350 万	**75.30**	**83.90**	**83.20**	**99.60**	**83.30**	**32.10**	**71.56**	**121.76**

根据表 7 的实验数据绘制黑盒攻击成功率 ASR 与不可察觉性指标 FID 的关系图，如图 9 所示。可以看出，在相同的黑盒攻击成功率条件下，MI-LA 算法的不可察觉性指标 FID 最低，综合性能最佳。

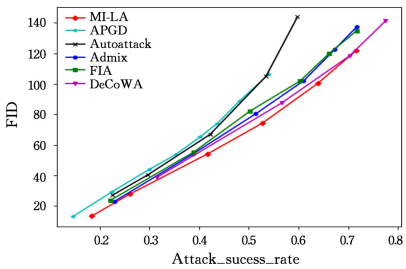


图 9 不同攻击对抗攻击算法 ASR 与 FID 的关系

Fig. 9 Relationship between ASR and FID using different adversarial attack algorithms

为方便比较，采用线性插值法计算 ASR 为 40%，45%，50%，55%，60%时 6 种方法的 FID，结果如表 8 所列。可以看出，在相同的 ASR 下，MI-LA 算法的 FID 值较原方法下降了 6.32%～26.35%，充分说明了改进方法的有效性。

表 8 线性插值后相同 ASR 对应 FID 的下降情况

Table 8 Decrease of FID corresponding to the same ASR after linear interpolation

攻击方法	*ASR*/%					MI-LA *FID* 下降/%
	40	45	50	55	60	
APGD	64.88	78.50	94.37	106.26	106.26	↓22.06
Autoattack	62.23	76.36	93.30	105.02	143.86	↓26.35
Admix	56.67	67.08	77.49	88.47	99.68	↓9.82
FIA	57.90	69.75	81.60	91.47	101.23	↓12.66
DeCoWA	55.22	64.93	74.64	84.35	95.15	↓6.32
MI-LA	51.23	60.13	69.27	79.55	91.26	—

5 总结与展望

5.1 总结

本文针对现有对抗攻击方法中固定约束策略导致的不可察觉性不足的问题，提出了一种基于自适应约束上界的对抗攻击优化框架。通过理论分析与实验验证，取得了以下成果。

1）样本敏感性与约束强度的定量建模：首次建立样本敏感性指标与对抗扰动约束强度的动态映射机制，突破了传统固定约束的假设。基于梯度幅值构建敏感性量化模型，揭示不同语义特征样本的扰动阈值差异规律，为自适应攻击提供可解释性理论支撑。

2）轻量化计算与动态约束机制：提出单步梯度幅值近似策略，将显著性计算复杂度从 O(d)降至 O(1)，突破了高分辨率图像的计算瓶颈。结合敏感性分布特征设计动态约束映射函数，实现了“高敏感-低约束、低敏感-高约束”的差异化扰动分配，在相同攻击成功率下使 FID 指标较固定约束方法降低 2.68%～3.49%。

3）算法通用性与有效性验证：所提方法可无缝集成至 FGSM，PGD 和 MI-FGSM 等主流攻击框架，仅需替换投影算子和步长即可实现动态约束。基于 MI-LA 算法的对比实验表明，其 FID 值较 APGD 和 AutoAttack 等 5 种方法降低了 6.32%～26.35%，验证了框架的优越性。

5.2 展望

本文方法尽管在不可察觉性与计算效率上取得了提升，但仍存在以下待解决问题。

1）跨域泛化性：当前实验基于 ImageNet-Compatible 数据集，未来须验证方法在低光照和遮挡等复杂场景下的鲁棒性。另外，可以探索面向视频、点云等多模态数据的扩展性，其中

视频任务须处理时序一致性与运动模糊，点云任务须处理空间稀疏性与不规则结构。在不同模态的数据上，敏感性量化模型和动态约束机制需要相应的适配与创新。

2）防御策略的对抗性：现有研究聚焦攻击效能的提升，但动态约束机制可能被防御模型逆向破解（如检测输入样本梯度幅值分布的异常模式），须进一步研究对抗训练中动态约束的隐蔽性增强策略。同时，随着防御技术的发展，本文方法在最新防御模型（如基于扩散模型的净化方法 DiffPure，更强的对抗训练变体如 TRADES＋＋，或新兴架构如 ConvNeXt 和 Vision MLP 等）下的有效性也须持续评估和适配。

3）感知-攻击联合优化：当前 FID 指标仅作为后验评估指标，未来可设计感知-攻击一致性损失函数，实现攻击过程中特征分布的直接优化。

本文为对抗攻击提供了新的研究范式，相关代码已开源[1)]。后续将围绕动态约束的可解释性理论、跨任务迁移机制（如目标检测与语义分割）及轻量化防御框架展开深入研究。

参 考 文 献

[1] GOODFELLO I J, JONATHON S, SZEGEDY C. Explaining and harnessing adversarial examples [J]. arXiv: 1412. 6572, 2014.

[2] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks [C] // ICLR. 2018.

[3] ZHENG J H, LIN C H, SUN J H, et al. Physical 3D adversarial attacks against monocular depth estimation in autonomous driving [C] // Proceedings of the 42nd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 24452-24461

[4] WANG J S, MAO X T, WANG Y, et al. Automatic Generation of Pathological Benchmark Dataset from Hyperspectral Images of Double Stained Tissues [J]. Optics and Laser Technology, 2023, 163: 109331-109331.

[5] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks [C] // Proceedings of the 6th International Conference on Learning Representations. 2018: 1-23.

[6] DUCHI J, HAZAN E, SINGER Y. Adaptive subgradient methods for online learning and stochastic optimization [J]. Journal of Machine Learning Research, 2011, 12: 2121-2159.

[7] PROVENZI E. Rudiments of Human Visual System (hvs) Features [M] // Computational Color Science Variational Retinex-Like Methods. John Wiley & Sons, Inc., 2017: 1-11.

[8] MARTIN H, HUBER R, THOMAS U, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium [C] // Advances in Neural Information Processing Systems. San Diego: NIPS, 2017.

[9] PAPERNOT N, MCDANIEL P D, JHA S, et al. The limitations of deep learning in adversarial settings [C] // Proceedings of the IEEE European Symposium on Security and Privacy. 2016: 372-387.

[10] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks [C] // Proceedings of the 38th IEEE Symposium on Security and Privacy. 2017: 39.

[11] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks [C] // Proceedings of the 2nd International Conference on Learning Representations. 2014: 1.

[12] WANG Z B, WANG X, MA J J, et al. A review of adversarial sample attacks for computer vision systems [J]. Journal of Computer Science, 2023, 46(2): 436-468.

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

[14] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial machine Learning at scale [C] // Proceedings of the 5th International Conference on Learning Representations. 2017: 1-17.

[15] OSELEDETS I, KHRULKOV V. Art of singular vectors and universal adversarial perturbations [C] // Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018: 8562-8570.

[16] DI M, PENG R, WANG Y L, et al. Boosting the Transferability of Adversarial Attack on Vision Transformer with Adaptive Token Tuning [C] // NeurIPS. 2024.

[17] ZHOU Q, CHEN J, TAO Q. Adversarial attack optimize method based on L1-mask constraint [J]. CAAI Transactions on Intelligent Systems, 2025(3): 594-604.

[18] 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.

[19] SZEGEDY C, LIU W, JIA Y Q, et al. Going deeper with convolutions [C] // 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2014: 1-9.

[20] 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 Pattern Recognition. 2016: 770-778.

[21] KAREN S, ANDREW Z. Very deep convolutional networks for large-scale image recognition [J]. arXiv: 1409. 1556, 2014.

[22] MARK S, ANDREW H, ZHU M L, et al. Mobilenetv2: inverted residuals and linear bottlenecks [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 4510-4520.

[23] ALEXEY D, LUCAS B, ALEXANDER K, et al. An image is worth 16x16 words: transformers for image recognition at scale

1) https://github.com/AtTheMoment12/AdaConst.

[C]// International Conference on Learning Representations. 2021.

[24] PASZKE A, GROSS S, MASSA F, et al. PyTorch: an imperative style, high- performance deep learning library[J]. arXiv: 1912.01703, 2019.

[25] KIM H. Torchattacks: a pytorch repository for adversarial attacks[J]. arXiv: 2010.01950, 2020.

[26] FRANCESCO C, HEIN M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks [C]// International Conference on Machine Learning. Washington DC: ICLR, 2020.

[27] FRANCESCO C, HEIN M. Mind the box: l1-APGD for sparse adversarial attacks on image classifiers[C]// International Conference on Machine Learning. Washington DC: ICLR, 2021.

[28] WANG X S, HE X R, WANG J D, et al. Admix: Enhancing the Tranferability of Adversarial Attacks[J]. arXiv: 2102.00436, 2021.

[29] WANG Z, GUO H, ZHANG Z, et al. Feature Importance-aware Transferable Adversarial Attacks[C]// 2021 IEEE/CVF International Conference on Computer Vision(ICCV). 2021.

[30] LIN Q L, LUO C, NIU Z H, et al. Boosting Adversarial Transferability across Model Genus by Deformation-Constrained Warping[C]// AAAI. 2024.

ZHOU Qiang, born in 1990, master, is a member of CCF (No. P1378G). 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. 08091S). His main research interests include machine learning, pattern recognition and applied mathematics.

(责任编辑:柯颖)