智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于L_1-mask约束的对抗攻击优化方法

周强, 陈军, 陶卿

引用本文:
周强, 陈军, 陶卿. 基于L_1-mask约束的对抗攻击优化方法[J]. 智能系统学报, 2025, 20(3): 594-604.
ZHOU Qiang, CHEN Jun, TAO Qing. Adversarial attack optimization method based on L_1-mask constraint[J]. *CAAI Transactions on Intelligent Systems*, 2025, 20(3): 594-604.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202405037

DOI: 10.11992/tis.202405037
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20250416.1145.004

基于 L_1-mask 约束的对抗攻击优化方法

周强, 陈军, 陶卿
(陆军炮兵防空兵学院 信息工程系, 安徽 合肥 230031)

摘　要: 当前的对抗攻击方法通常采用无穷范数或 L_2 范数来度量距离, 但在不可察觉性方面仍有提升空间。L_1 范数作为稀疏学习的常用度量方式, 其在提高对抗样本的不可察觉性方面尚未被深入研究。为了解决这一问题, 提出基于 L_1 范数约束的对抗攻击方法, 通过对特征进行差异化处理, 将有限的扰动集中在更重要的特征上。此外, 还提出了基于显著性分析的 L_1-mask 约束方法, 通过遮盖显著性较低的特征来提高攻击的针对性。这些改进不仅提高了对抗样本的不可察觉性, 还减少了对抗样本对替代模型的过拟合风险, 增强了对抗攻击的迁移性。在 ImageNet-Compatible 数据集上的实验结果表明: 在保持相同黑盒攻击成功率的条件下, 基于 L_1 约束的对抗攻击方法不可察觉性指标 FID(frechet inception distance) 指标较无穷范数低约 5.7%, 而基于 L_1-mask 约束的 FID 指标则低约 9.5%。

关键词: 对抗攻击; L_1 范数; 遮盖; 显著性; 不可察觉性; 迁移性; 稀疏; 约束
中图分类号: TP181　　**文献标志码:** A　　**文章编号:** 1673−4785(2025)03−0594−11

中文引用格式: 周强, 陈军, 陶卿. 基于 L_1-mask 约束的对抗攻击优化方法 [J]. 智能系统学报, 2025, 20(3): 594–604.
英文引用格式: ZHOU Qiang, CHEN Jun, TAO Qing. Adversarial attack optimization method based on L_1-mask constraint[J]. CAAI transactions on intelligent systems, 2025, 20(3): 594–604.

Adversarial attack optimization method based on L_1-mask constraint

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

Abstract: The existing adversarial attack methods generally utilize infinite or L_2 norms to measure distance. However, these methods can be improved in terms of imperceptibility. Moreover, the L_1 norm, as a conventionally employed metric method in sparse learning, has not been extensively studied in terms of improving the imperceptibility of adversarial samples. To address this research gap, an adversarial attack method based on the L_1 norm constraint is proposed, and it focuses limited perturbations on more crucial features by performing feature differentiation processing. Additionally, an L_1-mask constraint method based on saliency analysis is proposed to improve attack targeting by masking low-saliency features. The results reveal that these improvements enhance the imperceptibility of adversarial samples and reduce the risk of overfitting alternative models with adversarial samples, thereby enhancing the transferability of adversarial attacks. Experiments using the ImageNet compatible dataset reveal that the imperceptibility FID index of the L_1-constrained adversarial attack methods is approximately 5.7% lower than that of the infinite norm while maintaining the same success rate for black box attacks. Conversely, the FID index of L_1-mask-constrained adversarial attack methods is approximately 9.5% lower.

Keywords: adversarial attack; L_1 norm; mask; saliency; imperceptibility; transferability; sparse; constraint

近年来, 深度神经网络 (deep neural network, DNN) 在自动驾驶[1]、医学图像分析[2] 等诸多领域表现出了卓越性能。但文献 [3-4] 阐明了 DNN 容易对精心设计的对抗样本产生误判, 并且通常不会被人类发现, 这可能会在现实应用中引发严重错误。此外, 这些对抗样本在不同模型体系结构之间的迁移性对实际应用提出了更大的挑战。因此, 对抗攻击的研究受到越来越多的关注[5]。

收稿日期: 2024−05−27.　**网络出版日期:** 2025−04−16.
基金项目: 国家自然科学基金项目 (62076252).
通信作者: 陶卿. E-mail: taoqing@gmail.com.

对抗攻击通常分为白盒攻击和黑盒攻击。白盒攻击[4-5] 可以访问目标模型的架构和参数; 黑盒攻击[6-19] 无法获取目标模型的信息, 因此更接近真实世界的场景。基于迁移性的黑盒攻击在替代模型上构建对抗样本来欺骗目标模型。通过采用不同的优化策略, 设计各种损失函数, 利用多种数据增广等方法, 迁移性得到了较大提高。

但是, 在迁移性得到有效提高的同时, 对抗样本的不可察觉性并不令人满意。原因在于, 以上大多数方法都是在 RGB 空间中采用无穷范数或 L_2 范数作为度量人类感知的标准, 并约束原图像

和对抗样本之间的扰动幅度。采用无穷范数约束倾向于让所有像素点都产生最大限度的变化, 而采用 L_2 范数的对抗样本倾向于让所有像素点产生变化。这种对所有像素点进行修改的方法忽视了像素点的重要性差异, 从而增加了人眼的不可察觉性。

众所周知, L_1 范数具有稀疏性[20], 能够筛选重要特征, 被广泛用于压缩感知[6]、稀疏恢复[7] 等问题。然而, 目前基于 L_1 范数约束的对抗攻击研究还不够深入, 特别是利用 L_1 范数提升不可察觉性和增强迁移性的方法还没有相关文献记载。本研究认为 L_1 范数由于其内在稀疏性, 可以选择重要像素进行攻击而保持次要像素不变, 从而把有限的扰动量用在高效的目标区域, 以减少不必要的像素改变, 从而降低不可察觉性; 同时, 通过特征选择还可以降低对抗样本过拟合于替代模型的风险。该方法适用于所有基于 L_p 范数约束的对抗攻击方法, 通过将 L_p 范数替换成 L_1 范数约束可以获得不可察觉性的改善和迁移性的提高。

此外, 从可解释性[8-9] 角度分析, 深度神经网络能够识别目标主要是因为能够提取目标区域信息并进行加工整理。文献 [10-11] 给出了显著性 (saliency) 计算方法, 即计算输入图片像素点对损失函数的偏导数的绝对值, 偏导数绝对值大的像素更为显著, 对模型推理作用更大。同理, 对抗攻击的重点应该是对模型显著性较大的像素。本研究以显著性大小为依据对像素进行分类, 只攻击显著性超过阈值的像素, 并与 L_1 约束相结合, 提出 L_1-mask 约束方法, 该方法能够进一步提升对抗攻击的性能。

本研究的主要贡献包括 3 个方面:

1) 提出了一种基于 L_1 范数并充分利用像素显著性信息的约束方法 L_1-mask。该方法采用投影算法求解 L_1 范数约束优化问题, 实现对特征的差异化处理, 并且对显著性较低的特征进行遮盖, 使有限的扰动更多作用于重要特征, 改善了对抗攻击的不可察觉性, 同时降低了对抗样本过拟合于替代模型的风险, 提升了对抗攻击的迁移性。

2) 提出了部分参数设定的解析方法。提出不同约束方式下扰动值 ε 之间的对应比例关系, 并建立模型, 根据扰动值 ε 确定步长 α 和 mask 系数 λ, 使 ε 成为唯一需要调整的超参数, 有效降低了对抗攻击超参搜索维度, 节约了实验能耗及时间开销。

3) 实验验证了本研究方法的有效性。在 ImageNet-Compatible 数据集上对不同约束方法的黑盒攻击能力和不可察觉性进行检验。实验结果表明, 在相同的黑盒攻击成功率下, 基于 L_1 约束的对抗攻击方法不可察觉性指标 (frechet inception distance, FID) 比无穷范数低约 5.7%, 基于 L_1-mask 约束的 FID 比无穷范数低约 9.5%。

1　相关工作

对抗攻击度量不可察觉性的理想指标是 FID, 但该指标数学表达复杂, 难以在优化过程中直接运用。为便于优化求解, 现有研究大多通过 P 范数约束图片变化距离, 常用的范数约束包括: 无穷范数、L_2 范数、L_1 范数和 L_0 范数。

无穷范数约束要求对抗噪声在 $[-\varepsilon,\varepsilon]^d$ 内变化 (其中 d 表示输入图片的维度)。文献 [4] 提出了快速梯度符号方法 (fast gradient sign method, FGSM), 让图片的每个像素在梯度符号的方向上产生 ε 位移, 可快速生成对抗样本。文献 [5] 提出迭代版本的快速梯度符号方法 (iterative fast gradient sign method, I-FGSM), 提升了白盒攻击的成功率。文献 [12-16] 提出了动量迭代快速梯度符号方法 (momentum iterative fast gradient sign method, MI-FGSM)、Nesterov 动量迭代快速梯度符号方法 (Nesterov iterative fast gradient sign method, NI-FGSM)、平移不变迭代快速梯度符号方法 (translation-invariant iterative fast gradient sign method, TI-FGSM)、多样化输入迭代快速梯度符号方法 (diverse inputs iterative fast gradient sign method, DI-FGSM)、基于补丁的迭代快速梯度符号方法 (patch-wise iterative fast gradient sign method, PI-FGSM) 等方法, 通过引入动量等手段提升迭代攻击方向的稳定性, 从而提升了黑盒攻击的迁移性。基于无穷范数约束的对抗攻击因计算便捷而被广泛应用, 但该约束方式倾向于让所有像素点都产生最大限度的变化, 降低了对抗攻击的不可察觉性。如图 1 所示, 无穷范数约束产生的对抗样本不仅在目标区域产生噪声, 也在背景中产生了明显的条纹, 这些条纹会使人眼很容易察觉; 同时, 这些条纹会导致对抗样本过拟合于替代模型, 降低对抗迁移性。

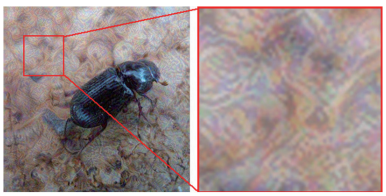


图 1　无穷范数约束生成对抗样本

Fig. 1　Adversarial sample generated by L_∞

L_2 范数约束要求对抗噪声的 L_2 范数小于扰动值 ε, 对于超出约束范围的解向 L_2 范数球投影,

即对抗噪声除以自身的 L_2 范数再乘以 ε。文献 [17-19,21] 提出的基于 L_2 范数的投影梯度下降方法 (projected gradient descent L_2, $PGDL_2$)、自动投影梯度下降 (autoprojected gradient descent, APGD)、特征对抗攻击 (feature adversarial attack, FAB)、方形攻击 (SquareAttack) 等算法都有 L_2 版本。基于 L_2 范数约束的对抗攻击因便于计算而应用也较为广泛，但该约束方式没有对特征进行差异化处理，对抗攻击的不可察觉性仍有提升空间。

L_1 范数约束要求对抗噪声的 L_1 范数小于扰动值 ε，对于超出约束范围的解向 L_1 范数球投影。但是向 L_1 范数球投影计算较为复杂，因此，L_1 范数约束在对抗攻击中的研究还比较少。文献 [22] 在 APGD 算法基础上结合 L_1 范数球投影方法提出 L_1-APGD 算法，但未进行不可察觉性分析。文献 [23] 提出的基于 L_1 范数的弹性网络攻击 (elastic-net attacks to deep neural networks based-L_1, $EADL_1$) 算法利用了 L_1 正则化项，相比于投影计算，降低了计算难度，但无法获得准确的约束范围。总之，将 L_1 范数用于度量对抗攻击距离的研究还不够深入，特别是对不可察觉性的作用还需要进一步研究。

L_0 范数约束要求对抗样本改变的像素点数量小于 ε，对于每一点的变化量不作要求，因此，其生成的对抗样本通常可察觉。目前，L_0 范数约束在稀疏对抗攻击领域应用广泛，其生成的对抗补丁可直接应用于真实世界。文献 [24] 提出了 One-Pixel 算法，文献 [25] 提出的 SparseFool 算法能够在只改变少数像素点的情况下取得较好的攻击效果，文献 [26-27] 提出的 Pixel 和基于雅可比矩阵的显著图攻击 (Jacobian-based saliency map attack, JSMA) 算法也是稀疏对抗攻击的常用方法。

2　L_1-mask 约束方法

为改善对抗样本的不可察觉性，本研究提出了 L_1-mask 约束方法。该约束方法可以无障碍推广至所有基于 L_p 范数约束的对抗攻击算法。本节以经典的 MI-FGSM 为例，设计其基于 L_1 约束的版本 MI-L_1，并在此基础上充分利用像素的显著性差异设计 MI-L_1-mask 算法，进一步提升算法效能。

2.1　L_1 范数约束

L_1 范数具有稀疏性，向 L_1 范数球投影的解是一个稀疏解，因此 L_1 范数也被叫作稀疏规则算子。通过向 L_1 范数球投影可以实现特征的稀疏选择，过滤掉无关特征而保留重要特征。基于 L_1 范数约束的无目标攻击方法可用数学公式表达为

$$\underset{\boldsymbol{x}^*}{\operatorname{argmax}} J(\boldsymbol{x}^*, y),\ \text{s.t.}\|\boldsymbol{x}^*-\boldsymbol{x}\|_1 \leqslant \varepsilon_1 \tag{1}$$

式中：$\boldsymbol{x}$ 表示真实样本，$\boldsymbol{x}^*$ 表示对抗样本，J 表示损失函数，ε_1 表示对抗扰动。

求解式 (1) 常用的方法是投影次梯度方法，但向 L_1 范数球投影的计算难度明显高于向无穷范数球和 L_2 范数球投影。值得庆幸的是 (Duchi 等[28]) 给出了向 L_1 范数球投影的方法：

令 $\boldsymbol{v}=\boldsymbol{x}^*-\boldsymbol{x}$，$\boldsymbol{w}=P_{0,\varepsilon_1}(\boldsymbol{v})$，其中 $P_{0,\varepsilon_1}(\cdot)$ 表示向中心为 0、半径为 ε_1 的 L_1 范数球投影。那么，如果 $\|\boldsymbol{v}\|_1 \leqslant \varepsilon_1$，则 $\boldsymbol{w}=\boldsymbol{v}$；如果 $\|\boldsymbol{v}\|_1 > \varepsilon_1$，则

1) 令 $\boldsymbol{u}$ 表示 $\boldsymbol{v}$ 的绝对值矢量，即 $u_i=|v_i|$，对矢量 $\boldsymbol{u}$ 进行降序排序得到，$\boldsymbol{u}: u_1 \geqslant u_2 \geqslant \cdots \geqslant u_d$；

2) $\rho(\varepsilon_1,\boldsymbol{u})=\max\left\{j\in[n]: u_j-\dfrac{1}{j}\left(\sum_{r=1}^{j}u_r-\varepsilon_1\right)>0\right\}$；

3) 定义 $\theta=\dfrac{1}{\rho}\left(\sum_{i=1}^{\rho}u_i-\varepsilon_1\right)$；

4) $\boldsymbol{w}$ 的分量 $w_i=\operatorname{sign}(v_i)\max\{u_i-\theta,0\}$。

利用上述投影方法，基于无穷范数约束和 L_2 范数约束的对抗攻击方法可以无障碍地拓展到其 L_1 范数约束形式。本文以经典的 MI-FGSM 为例介绍其拓展方法。MI-FGSM 原始更新规则描述为

$$\boldsymbol{x}_0^*=\boldsymbol{x}, \boldsymbol{g}_0=0 \tag{2}$$

$$\boldsymbol{g}_{t+1}=\mu\boldsymbol{g}_t+\frac{\nabla J(\boldsymbol{x}_t^*,y)}{\left\|\nabla J(\boldsymbol{x}_t^*,y)\right\|_1} \tag{3}$$

$$\boldsymbol{x}_{t+1}^*=\operatorname{Clip}_{\boldsymbol{x},\varepsilon_\infty}\{\boldsymbol{x}_t+\alpha\cdot\operatorname{sign}(\boldsymbol{g}_{t+1})\} \tag{4}$$

式中：$\boldsymbol{g}_t$ 为前 t 次迭代中累加的梯度，μ 为动量系数，α 为每次迭代时使用梯度符号对样本更新的步长。$\operatorname{Clip}_{\boldsymbol{x},\varepsilon_\infty}\{\cdot\}$ 为裁剪函数，能够保证生成的对抗样本符合 L_∞ 范数约束。MIFGSM 的 L_1 范数约束形式只需将式 (4) 换成向 L_1 投影，即

$$\boldsymbol{x}_{t+1}^*=P_{\boldsymbol{x},\varepsilon_1}\{\boldsymbol{x}_t+\alpha\cdot\operatorname{sign}(\boldsymbol{g}_{t+1})\} \tag{5}$$

完整的 MI-L_1 算法如算法 1。

算法 1　MI-L_1

输入　模型 f，图片样本 $\boldsymbol{x}$，标记 y，扰动 ε_1，步长 α；默认参数：损失函数 J 为交叉熵，迭代步数 T=10，动量系数 $\mu=1$。

输出　对抗样本 $\boldsymbol{x}^*$，满足 $\|\boldsymbol{x}^*-\boldsymbol{x}\|_1 \leqslant \varepsilon_1$。

1) 初始化：$\boldsymbol{g}_0=0; \boldsymbol{x}_0^*=0$;

2) 循环 t=0 至 T−1：

3) 计算梯度 $\boldsymbol{g}=\nabla_x J\left(\boldsymbol{x}_t^*,y\right)$；

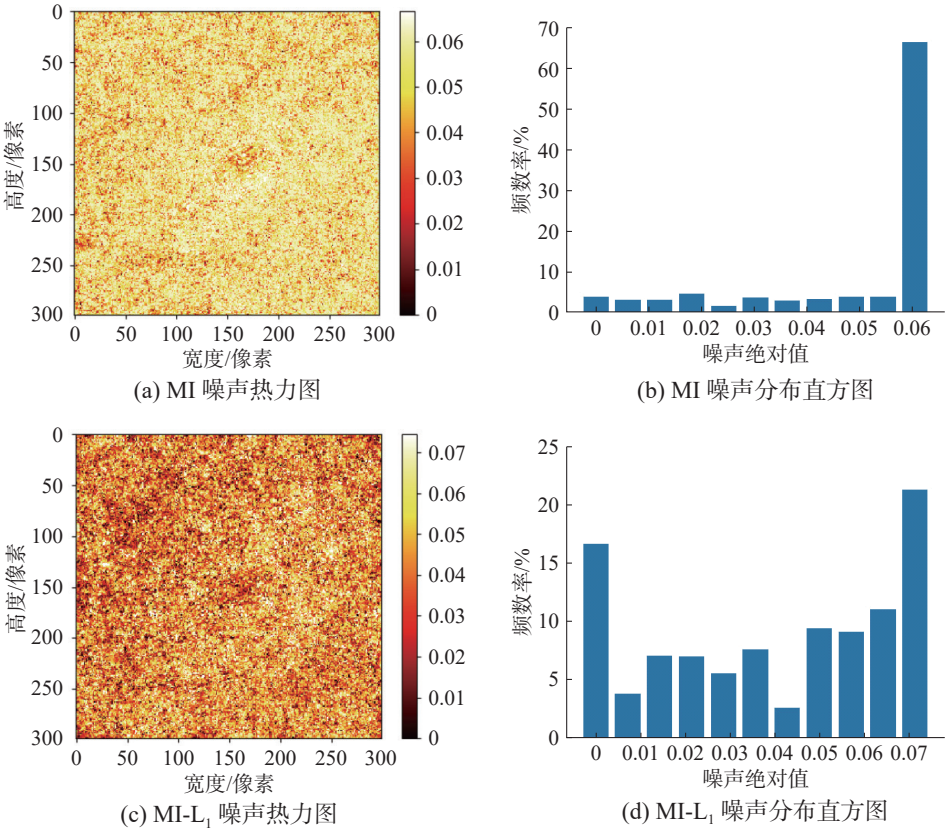
4) 更新动量 $\boldsymbol{g}_{t+1}=\mu\cdot\boldsymbol{g}_t+\dfrac{\boldsymbol{g}}{\boldsymbol{g}_1}$；

5) 更新 $\boldsymbol{x}_{t+1}^*=\boldsymbol{x}_t^*+\alpha\cdot\operatorname{sign}(\boldsymbol{g}_{t+1})$；

6) 向 L_1 范数球投影 $\boldsymbol{x}_{t+1}^*=P_{\boldsymbol{x},\varepsilon_1}\left\{\boldsymbol{x}_{t+1}^*\right\}$。

图 2 给出了图 1 中甲虫图片在 MI 和 MI-L_1 产生对抗噪声的热力图和分布图，其中频数率指

某类像素占像素总数的比例。向无穷范数球投影后绝大部分像素的变化都在最大值附近，对不同重要性的特征进行了无差别的处理；向 L_1 范数球投影后，噪声幅值分布较为分散，部分像素变化为零，部分像素变化值较大，一定程度上实现了特征的差异化处理。但目标区域和背景区域噪声差别并不明显，并没有使敏感的目标区域获得更大的改变而背景区域获得较小的改变。上述分析表明，L_1 范数约束与梯度符号方法结合产生的特征选择效果不够明显，究其原因，可能是梯度符号方法使显著性不同的像素产生了相同的改变量，这些无差别的改变量在向 L_1 范数球投影过程中难以产生差别化作用，导致特征选择的效果没有充分显露。



(a) MI 噪声热力图　(b) MI 噪声分布直方图

(c) MI-L_1 噪声热力图　(d) MI-L_1 噪声分布直方图

图 2　对抗噪声分析

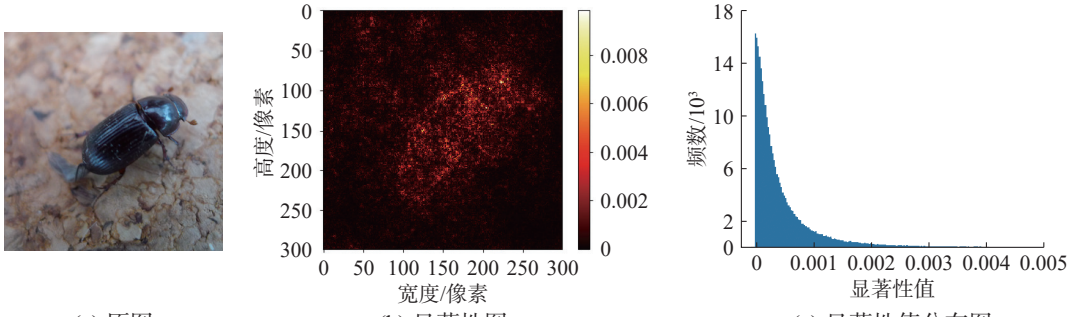
Fig. 2　Adversarial noise analysis

2.2　L_1-mask 约束方法

为充分利用目标区域和背景区域所蕴含的信息量差异，本研究借鉴深度神经网络可解释性分析方法，对图片进行了显著性分析。图 3 为图片像素对 ResNet-152 模型的显著性分析。显著性图由各像素对损失函数求偏导数取绝对值后绘制，其中显著性计算方法为

$$s_i = \left|\frac{\Delta J}{\Delta x_i}\right| = \left|\frac{\partial J}{\partial x_i}\right| \tag{6}$$

该图显示模型对图片中的甲虫所在区域敏感，即甲虫所在区域像素点显著更高，而背景区域像素显著性低，对模型推理作用小；图 3(c) 为显著性值分布图，该图显示显著性值呈指数分布，即大部分像素显著性值在 0 附近，只有少数像素显著性较大，对模型推理贡献较大。



(a) 原图　(b) 显著性图　(c) 显著性值分布图

图 3　显著性值分析

Fig. 3　Saliency analysis

由上述分析可得，像素的显著性对特征选择提供了重要依据，即在采用 L_1 范数约束的同时，把显著性较低的像素遮盖 (保持其不变)，并将有限扰动量分配给显著性高的像素，以降低背景区域的低效率改变，从而获得更好的攻击针对性。该方法称为 L_1-mask 约束方法。令 $\boldsymbol{m}$ 表示遮盖向量，其形状与输入图片相同，维度为 $H\times W\times C$ (高度×宽度×通道)，其元素 q_i 计算规则为

$$q_i=\begin{cases}1, & s_i\geqslant\tau\\ 0, & s_i<\tau\end{cases} \tag{7}$$

式中：τ 为阈值，τ 的大小由未被遮盖的像素所占比例 λ 确定，称之为 mask 系数。那么，基于 L_1-mask 约束的对抗攻击，更新公式只需要将式 (5) 替换为

$$\boldsymbol{x}_{t+1}^{*}=P_{x,\varepsilon_1}\{\boldsymbol{x}_t+\alpha\cdot\mathrm{sign}(\boldsymbol{g}_{t+1})\odot\boldsymbol{m}\} \tag{8}$$

该算法实现方法如算法 2。

算法 2　MI-L_1-mask

输入　模型 f，图片样本 $\boldsymbol{x}$，标记 y，扰动 ε_1，步长 α，mask 系数 λ；默认参数：损失函数 J 为交叉熵，迭代步数 T=10，动量系数 $\mu=1$。

输出　对抗样本 $\boldsymbol{x}^*$，满足 $\|\boldsymbol{x}^*-\boldsymbol{x}\|_1\leqslant\varepsilon_1$。

1) 初始化：$\boldsymbol{g}_0=0$；$\boldsymbol{x}_0^*=0$；

2) 循环 t=0 至 T–1：

3) 计算梯度 $\boldsymbol{g}=\nabla_x J\left(\boldsymbol{x}_t^*,y\right)$；

4) 如果 t=0：

5) 由式 (6) 计算显著性并排列 $s_1<s_2<\cdots<s_d$；

6) 计算阈值 $\tau=s_{d\times(1-\lambda)}$，$d$ 为输入图片维数；

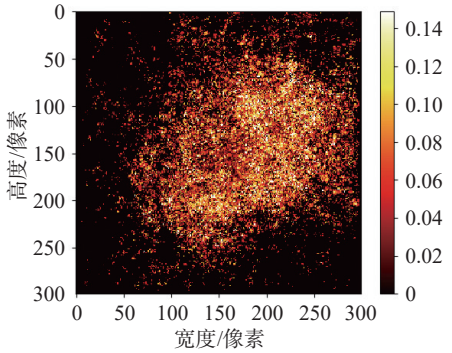
7) 由式 (7) 计算 $\boldsymbol{m}$；

8) 更新动量 $\boldsymbol{g}_{t+1}=\mu\cdot\boldsymbol{g}_t+\dfrac{\boldsymbol{g}}{\boldsymbol{g}_1}$；

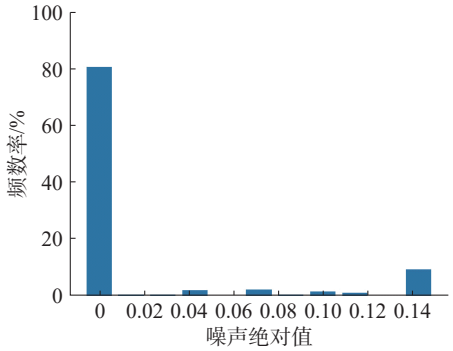
9) 更新 $\boldsymbol{x}_{t+1}^*=\boldsymbol{x}_t^*+\alpha\cdot\mathrm{sign}(\boldsymbol{g}_{t+1})\odot\boldsymbol{m}$；

10) 向 L_1 范数球投影 $\boldsymbol{x}_{t+1}^*=P_{x,\varepsilon_1}\{\boldsymbol{x}_{t+1}^*\}$。

图 4 为甲虫图片经 MI-L_1-mask 算法生成的对抗样本进行噪声分析的情况。其中，图 4(a) 为各像素点扰动幅值热力图，图 4(b) 为各像素点扰动幅值分布图。由图 4 可知，背景区域像素未发生改变，说明 L_1-mask 约束能够有效过滤背景无用信息，使噪声更加集中地攻击目标区域，同时降低无效攻击对不可察觉性的影响。



(a) 噪声热力图



(b) 噪声分布直方图

图 4　L_1-mask 约束的对抗噪声分析

Fig. 4　Noise analysis based on L_1-mask constraint method

3　实验及分析

为验证 L_∞、L_2、L_1、L_1-mask 4 种约束方法对黑盒攻击不可察觉性及迁移性的影响，选取经典的 MI-FGSM 和 APGD 算法作为基准算法进行实验。对比实验可以无障碍拓展至 PGD、FAB 等其他基于梯度的对抗攻击算法，因篇幅限制在此不做赘述。

3.1　实验设置

数据集。按照对抗训练的通用做法，本文在 ImageNet-Compatible 数据集上验证不同约束方法的攻击效能。该数据集由 1 000 张图片组成，每张图片的尺寸为 299×299×3。

模型。本文选用 5 种卷积神经网络模型和 1 种视觉 Transformer(vision Transformer, ViT) 模型。其中，5 种卷积神经网络模型分别是 Inception-v3(Inc-v3)[29]、GoogleNet[30]、ResNet-152[31]、VGG-16 [32] 和 MobileNet-v2 (Mob-v2) [33]，它们采用 torchvision 库提供的模型框架和预训练参数；ViT 模型选用 Vit_base_patch16_224(ViT-B)[34]，采用 timm 库提供的模型和预训练参数。本文以 Inc-v3 和 ResNet-152 为黑盒攻击替代模型，其余 4 种模型为黑盒攻击目标模型。

对比方法。本文以 MI 和 APGD 方法在 L_∞、L_2、L_1、L_1-mask 4 种约束下的表现为例，验证 4 种约束的效能，可以无障碍地推广至所有基于梯度的攻击方法。其中无穷范数约束方法参见文献 [12]，L_2 范数约束方法如算法 3。

算法 3　MI-L_2

输入　模型 f，图片样本 $\boldsymbol{x}$，标记 y，扰动 ε_2，步长 α；默认参数：损失函数 J 为交叉熵，迭代步数 T=10，动量系数 $\mu=1$。

输出　对抗样本 $\boldsymbol{x}^*$，满足 $\|\boldsymbol{x}^*-\boldsymbol{x}\|_2\leqslant\varepsilon_2$。

1) 初始化：$\boldsymbol{g}_0=0;\boldsymbol{x}_0^*=0$；

2) 循环 t=0 至 T−1:

3) 计算梯度 $\boldsymbol{g}=\nabla_x J\left(\boldsymbol{x}_t^*,y\right)$;

4) 更新动量 $\boldsymbol{g}_{t+1}=\mu\cdot\boldsymbol{g}_t+\dfrac{\boldsymbol{g}}{\|\boldsymbol{g}\|_1}$;

5) 更新 $\boldsymbol{x}_{t+1}^*=\boldsymbol{x}_t^*+\alpha\cdot\mathrm{sign}(\boldsymbol{g}_{t+1})$;

6) 向 L_2 范数球投影 $\boldsymbol{x}_{t+1}^*=P_{x,\varepsilon_2}\left\{\boldsymbol{x}_{t+1}^*\right\}$.

实现细节。在后续的实验研究中, 所有实验均在 PyTorch[35] 框架下编程完成; 对抗攻击方法调用 torchattacks[36] 库或在此基础上对约束方法进行改进; FID 引用 pytorch_fid 库进行计算; 所有实验在 RTX 3060 GPU 上完成。

评价标准。区分为迁移性和不可察觉性: 迁移性采用黑盒攻击成功率 R_{AS}(attack success rate, ASR) 度量, ASR 定义为

$$R_{\mathrm{AS}}=\frac{\text{攻击成功样本数}}{\text{总样本数}}\times 100\%$$

相同条件下 ASR 值越大迁移性越好, 反之则说明迁移性越差。不可察觉性采用 FID[37] 度量, FID 值越小则说明不可察觉性越好, 反之则越差。

3.2　参数设置

按照对抗攻击通常做法, 本文迭代步数取 10 次。

1) 约束大小 ε。为研究对抗攻击成功率与 FID 之间的关系 (FID 只能通过 ε 间接调节), 需要估计不同范数扰动值 ε 之间的对应关系, 以使不同方法的 FID 在大致相当范围内变化, 进一步研究约束方法对黑盒攻击成功率的影响。对于 299×299×3 的图片而言, 以无穷范数为基准, 讨论其他两类约束的对应范围: 假设图片的每个像素都产生了一个单位变化, 则对应的 L_1 范数变化约为 $299\times299\times3\approx27\times10^4$; 对应的 L_2 范数变化约为 $\sqrt{299\times299\times3}\approx518$, 以此为依据设置约束值 ε 的变化范围, 即:

$$\frac{\varepsilon_1}{\varepsilon_\infty}\approx27\times10^4,\quad \frac{\varepsilon_2}{\varepsilon_\infty}\approx518$$

2) 步长 α。无穷范数采用 torchattacks 库中 MI-FGSM 默认的设置方法 $\alpha=\dfrac{\varepsilon_\infty}{T}\times2.5$。其中, T 是迭代步数; 2.5 是步长系数, 表示约 40% 的迭代次数在约束空间内部寻优, 约 60% 的迭代次数在约束空间边界寻优。其他约束方式的步长系数用 β 表示。则对于 L_2 范数, 有

$$\sqrt{\alpha^2\cdot d}=\frac{\varepsilon_2}{T}\cdot\beta_2$$

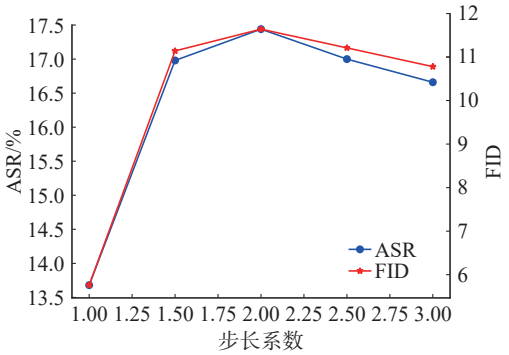
对于 L_1 范数, 有

$$\alpha\cdot d=\frac{\varepsilon_1}{T}\cdot\beta_1$$

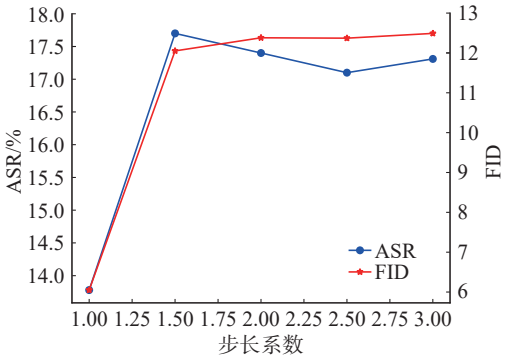
由图 5 可知 β_2=1.5, β_1=2 时, 综合性能最佳。对于 L_1-mask, 还有 1 个参数 mask 系数 λ 需要确定。显然, λ 应与 ε_1 正相关, 步长 α 也应与 ε_1 正相关, 假设 α 和 λ 在各自变化区间内等比例增减。RGB 图片所有像素在 [0,255] 变化, 因为迭代步数 T=10, 假设 α 在 [0,25.5] 变化, mask 的比例 λ 在 [0,1] 变化。显然, α 和 λ 正相关, 设 λ=α/25.5, 那么,

$$\alpha\cdot d\cdot\alpha/25.5=\frac{\varepsilon_1}{T}\cdot\beta_1$$

经求解得步长 α 的设置如表 1 所示。通过该关系, 步长 α、mask 系数 λ 均可以由 ε 来表示, ε 成为唯一的超参数, 有效提升了实验效率。



(a) L_2 范数约束



(b) L_1 范数约束

图 5　步长系数影响

Fig. 5　Influence of step size constraints

表 1　参数关系

Table 1　Parameter relationship

约束方式	扰动ε	步长α	mask系数λ
L_∞	ε_∞	$\dfrac{\varepsilon_\infty}{T}\times2.5$	—
L_2	ε_2	$\dfrac{\varepsilon_2}{T\cdot\sqrt{d}}\times1.5$	—
L_1	ε_1	$\dfrac{\varepsilon_1}{T\times d}\times2$	—
L_1-mask	ε_1	$\sqrt{\dfrac{51\varepsilon_1}{T\cdot d}}$	$\sqrt{\dfrac{\varepsilon_1}{12.75\cdot T\cdot d}}$

3.3　实验结果及分析

3.3.1　基于 MI 的实验

表 2 给出了 Inc-v3 和 ResNet-152 作为替代模型时, 4 种约束方式 ASR 与 FID 关系。

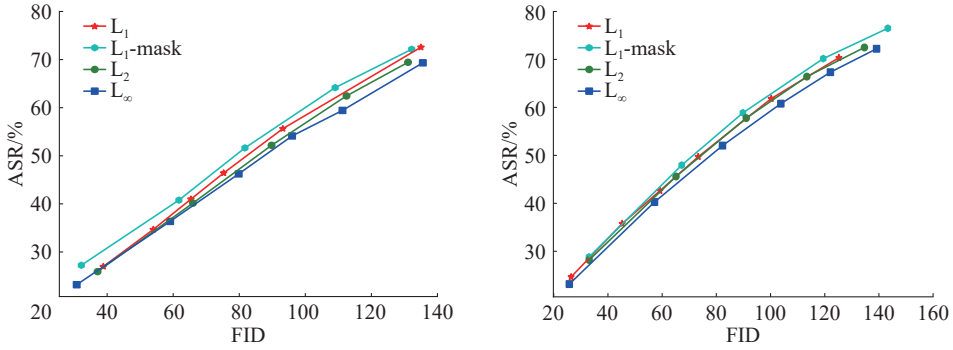
表 2　MI 不同约束方式 ASR 与 FID 对比

Table 2　Comparison of ASR and FID under different constraint methods of MI

替代模型	约束范数	扰动ε	ASR/%							FID
			Inc-v3	GoogleNet	VGG-16	ResNet-152	Mob-v2	ViT-B	黑盒平均	
Inc-v3	L_∞	4	98.8	28.4	27.1	18.9	31.6	10.0	23.20	30.78
		8	100.0	44.8	42.7	31.9	47.6	14.8	36.36	59.07
		12	100.0	55.8	53.0	42.5	62.1	17.8	46.24	79.83
		16	100.0	63.1	61.9	52.4	71.3	21.9	54.12	95.99
		20	100.0	69.4	66.2	58.7	77.7	25.2	59.44	111.26
		24	100.0	81.1	77.4	71.8	85.1	31.2	69.32	135.59
	L_2	2×10^3	99.3	30.8	32.0	20.6	34.8	11.4	25.92	37.11
		4×10^3	100.0	48.5	47.2	36.0	52.9	16.2	40.16	65.99
		6×10^3	100.0	63.5	60.1	50.2	66.0	21.1	52.18	89.78
		8×10^3	100.0	75.4	68.6	62.2	78.1	27.9	62.44	112.47
		10×10^3	100.0	80.3	77.3	71.2	85.4	33.0	69.44	131.10
	L_1	8×10^5	99.8	32.7	32.3	22.5	36.5	10.9	26.98	38.85
		12×10^5	100.0	42.8	39.4	30.3	46.6	14.3	34.68	53.98
		16×10^5	100.0	50.0	47.3	37.5	53.3	16.8	40.98	65.36
		20×10^5	100.0	57.3	53.5	43.0	58.7	19.6	46.42	75.28
		30×10^5	100.0	66.2	62.3	54.5	70.5	24.7	55.64	93.19
		40×10^5	100.0	85.1	80.8	75.3	86.5	35.1	72.56	135.00
	L_1-mask	3×10^5	96.7	29.2	31.7	31.3	32.2	11.8	**27.24**	**32.19**
		9×10^5	99.2	48.7	47.5	38.2	50.7	18.8	**40.78**	**61.72**
		15×10^5	100.0	61.7	59.4	50.6	61.1	25.4	**51.64**	**81.72**
		25×10^5	100.0	73.8	73.3	63.3	75.1	35.3	**64.16**	**109.03**
		35×10^5	100.0	82.1	80.0	71.8	84.0	42.8	**72.14**	**132.20**
ResNet-152	L_∞	4	14.6	24.2	31.7	96.8	36.1	9.2	23.16	25.74
		8	32.3	44.7	53.0	99.6	59.5	11.8	40.26	57.18
		12	46.4	61.4	66.0	100.0	72.8	13.6	52.04	82.27
		16	56.8	72.1	74.9	100.0	81.4	18.7	60.78	103.79
		20	66.9	80.1	80.3	100.0	86.7	22.7	67.34	122.10
		24	72.9	85.5	85.8	100.0	90.3	26.6	72.22	139.12
	L_2	2×10^3	19.9	30.1	36.5	98.0	44.3	10.1	28.18	33.06
		4×10^3	40.3	52.1	58.5	99.6	64.6	12.5	45.60	65.12
		6×10^3	55.9	68.3	71.2	99.9	76.8	16.6	57.76	91.02
		8×10^3	67.7	78.8	79.8	100.0	84.0	21.8	66.42	113.41
		10×10^3	74.0	86.8	83.9	100.0	89.7	28.1	72.50	134.65
	L_1	8×10^5	16.8	26.1	31.9	96.7	39.2	9.2	24.64	26.41
		12×10^5	28.2	40.7	47.2	99.1	51.9	10.8	35.76	45.26
		16×10^5	35.7	47.3	56.8	99.5	61.1	12.2	42.62	59.20
		20×10^5	46.2	57.4	63.3	99.9	68.4	13.2	49.70	73.31
		30×10^5	61.8	73.3	74.4	100.0	80.6	19.0	61.82	100.24
		40×10^5	71.9	83.7	82.1	100.0	87.8	26.3	70.36	125.22
	L_1-mask	3×10^5	25.6	32.2	36.1	88.4	39.0	11.1	**28.80**	**33.04**
		9×10^5	49.3	55.5	57.8	98.5	60.2	17.0	**47.96**	**67.22**
		15×10^5	61.6	69.6	69.9	99.0	70.5	22.7	**58.86**	**89.81**
		25×10^5	75.0	83.4	79.9	99.4	82.2	30.4	**70.18**	**119.44**
		35×10^5	80.3	88.3	85.5	99.7	87.1	41.3	**76.50**	**143.28**

注: 加黑表示本文方法在对比实验中效果最好。

根据表 2 的实验数据绘制 ASR 与 FID 关系图, 如图 6 所示。从图中可以清晰地看到, 在相同的 FID 情况下 L_1-mask 约束对抗攻击成功率最高, L_1 次之, L_2 再次之, L_∞最差。该结论与理论分析一致。



(a) Inc-v3 替代模型　　(b) ResNet-152 替代模型

图 6　4 种约束方式 ASR 与 FID 关系对比

Fig. 6　Comparison of ASR and FID among four constraints

用线性插值法计算 ASR 为 30%、40%、50%、60%、69% 时 4 种约束方法的平均 FID (表 3)。由表 3 可知, 在相同 ASR 下, 以无穷范数约束为参考, L_2 方法的 FID 值下降了约 3.7%, L_1 方法的 FID 值下降了约 5.7%, L_1-mask 方法的 FID 值下降了约 9.5%。实验数据证明了 L_1 方法和 L_1-mask 方法在改善不可察觉性方面的有效性。

表 3　线性插值后相同 ASR 对应 FID 对比

Table 3　Comparison of FID corresponding to the same ASR after linear interpolation

约束方法	ASR插值点的FID值					较L_∞ FID下降率/%
	30%	40%	50%	60%	69%	
L_∞	41.86	61.71	82.73	107.25	131.35	—
L_2	40.90	60.24	79.98	101.94	126.18	3.7
L_1	40.14	58.73	78.11	100.08	123.72	5.7
L_1-mask	36.70	56.52	75.07	96.37	119.72	9.5

图 7 给出选取 3 张图片和黑盒攻击成功率相当的对抗样本时, 4 种范数约束的不可察觉性差别对比。可以看出, 无穷范数的图片背景噪声最大, 不可察觉性最差, L_2 范数约束其次, L_1 范数再次, L_1-mask 的图片背景噪声最小, 不可察觉性最好, 实验结果证明了提出的理论假设。

(a) 干净样本　(b) L_1　(c) L_1-mask　(d) L_2　(e) L_∞

图 7　不同约束方式对抗样本实例

Fig. 7　Different constraint methods counteract sample instances

3.3.2　基于 APGD 的实验

为进一步验证 L_∞、L_2、L_1、L_1-mask 4 种约束方法对黑盒攻击不可察觉性及迁移性的影响, 选取 APGD 为基准算法进行实验。其中 APGD-L_∞、APGD-L_2 的实验方法参见文献 [18], APGD-L_1、Autoattack-L_1 的实现方法参见文献 [22], APGD-L_1-mask 的实现方法为在 APGD-L_1 算法基础上加上式 (6) ~ (8)。

实验结果如表 4 所示。该表给出了 ResNet-152 作为替代模型时，APGD 算法 4 种约束方式及 Autoattack-L_1 的 ASR 与 FID 关系。

根据表 4 的实验数据绘制黑盒攻击成功率 ASR 与 FID 距离关系图，如图 8 所示。从图中可以清晰地看到，相同的 FID 情况下 L_1-mask 约束对抗攻击成功率最高，L_1 次之，L_2 再次之，L_∞最差；相比于 Autoattack-L_1 方法也有较大优势。该结果说明 L_1-mask 方法能够有效提升对抗攻击方法的不可察觉性和对抗攻击迁移性。

表 4　APGD 不同约束方式 ASR 与 FID 对比

Table 4　Comparison of ASR and FID under different constraint methods of APGD

攻击方法	扰动ε	ASR/%							FID
		Inc-v3	GoogleNet	VGG-16	ResNet-152	Mob-v2	ViT-B	黑盒平均	
APGD-L_∞	4	6.2	13.5	21.7	68.3	23.3	7.8	14.50	12.90
	8	12.7	21.3	32.1	70.9	37.2	8.4	22.34	29.08
	12	18.8	31.6	43.3	71.1	47.5	8.2	29.88	43.80
	16	23.9	39.5	48.5	71.1	55.1	9.2	35.24	53.57
	20	29.8	47.6	52.8	71.1	59.5	11.3	40.20	65.36
	24	34.9	53.2	55.3	71.1	63.1	11.5	43.60	73.90
	30	41.6	58.0	60.4	71.1	67.4	12.9	48.06	88.55
APGD-L_2	1×10^3	7.7	14.4	23.8	66.9	23.6	7.3	15.36	13.29
	2×10^3	12.9	22.8	36.5	70.8	38.4	7.7	23.66	29.42
	4×10^3	28.7	43.1	54.2	71.1	56.8	9.8	38.52	58.07
	6×10^3	41.1	55.8	62.3	71.1	64.5	12.8	47.30	81.26
	8×10^3	50.5	63.3	66.5	71.7	69.5	17.3	53.42	99.19
APGD-L_1	3×10^5	10.5	21.2	30.9	69.0	30.0	7.6	20.04	20.80
	8×10^5	18.1	32.5	41.4	71.0	43.4	7.8	28.64	37.64
	12×10^5	26.5	42.0	49.1	71.1	51.0	9.1	35.54	50.99
	16×10^5	33.0	51.8	54.6	71.1	59.8	10.5	41.94	65.18
	20×10^5	41.0	55.2	59.7	71.1	63.2	12.5	46.32	76.64
	30×10^5	53.1	63.0	65.5	71.1	69.2	18.2	53.80	99.76
Autoattack-L_1	4×10^5	12.9	23.8	34.6	69.9	33.8	7.5	22.52	26.98
	8×10^5	18.9	33.0	43.4	70.9	44.7	8.6	29.72	40.48
	16×10^5	33.9	51.2	55.3	71.1	60.3	10.3	42.20	66.88
	30×10^5	51.0	64.4	65.9	71.1	69.6	16.4	53.46	105.02
APGD-L_1-mask	1×10^5	15.0	22.7	27.7	53.4	26.4	8.3	**20.02**	**20.45**
	3×10^5	24.3	35.9	39.4	66.8	39.2	9.9	**29.74**	**38.72**
	9×10^5	40.6	52.8	57.1	70.7	55.7	14.5	**44.14**	**64.77**
	15×10^5	48.9	60.8	63.3	71.0	60.5	18.3	**50.36**	**81.85**
	25×10^5	55.7	65.1	68.9	71.1	68.6	25.9	**56.84**	**106.23**

注：加黑表示效果最好。

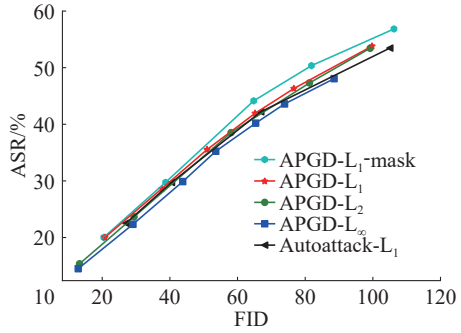


图 8　APGD 算法 4 种约束方式 ASR 与 FID 关系对比

Fig. 8　Comparison of ASR and FID among four constraint methods of APGD

4　结束语

本文提出了一种基于 L_1 范数约束的对抗攻击方法，并在此基础上通过显著性分析设计了基于 L_1-mask 约束的对抗攻击方法。通过实验证明，两种攻击方法在相同的黑盒攻击成功率下不可察觉性得到改善，性能优于基于无穷范数和 L_2 范数约束的对抗攻击方法。

在后续研究工作中，将在更大规模、更多样化的数据集上验证基于 L_1 范数及 L_1-mask 约束的对抗攻击算法的性能，同时考虑拓展至基于扩散模型的攻击方法上，以进一步探索和改进其适用性。

此外，L_1 范数具有稀疏性，与基于 L_0 范数约束的稀疏对抗密切相关，将探索以 L_1-mask 约束替代 L_0 范数约束，提升稀疏对抗可求解性。

另外，约束优化问题大多可以通过罚函数方法转化为无约束优化问题，而后通过近端梯度方法进行求解。将探索把多种约束混合使用，而后转化为正则化问题进行求解，提升对抗攻击的综合效能。

最后，数据增广技术是一种有效提升对抗攻击性能的手段，拟通过向图片增加随机噪声的方法增加原始图片的丰富性，再通过求平均值的方法获得对抗样本。

参考文献：

[1] 鲁思迪, 何元恺, 施巍松. 车计算: 自动驾驶时代的新型计算范式[J]. 计算机研究与发展, 2025, 62(1): 2−21.
LU Sidi, HE Yuankai, SHI Weisong. Vehicle computing: an emerging computing paradigm for the autonomous driving era[J]. Journal of computer research and development, 2025, 62(1): 2−21.

[2] 樊琳, 龚勋, 郑岑洋. 基于文本引导下的多模态医学图像分析算法[J]. 电子学报, 2024, 52(7): 2341−2355.
FAN Lin, GONG Xun, ZHENG Cenyang. A multi-modal medical image analysis algorithm based on text guidance [J]. Acta electronica sinica, 2024, 52(7): 2341−2355.

[3] GOODFELLOW I J, JONATHON S, SZEGEDY C. Explaining and harnessing adversarial examples[C]//International Conference on Learning Representations. Washington DC: ICLR, 2014.

[4] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial examples in the physical world[M]//Artificial Intelligence Safety and Security. First edition. Boca Raton: Chapman and Hall/CRC, 2018: 99−112.

[5] 纪守领, 杜天宇, 邓水光, 等. 深度学习模型鲁棒性研究综述[J]. 计算机学报, 2022, 45(1): 190−206.
JI Shouling, DU Tianyu, DENG Shuiguang, et al. Robustness certification research on deep learning models: a survey[J]. Chinese journal of computers, 2022, 45(1): 190−206.

[6] DONOHO D L. Compressed sensing[J]. IEEE transactions on information theory, 2006, 52(4): 1289−1306.

[7] CANDES E J, WAKIN M B. An introduction to compressive sampling[J]. IEEE signal processing magazine, 2008, 25(2): 21−30.

[8] BAEHRENS D, SCHROETER T, HARMELING S, et al. How to explain individual classification decisions[EB/OL]. (2019−12−06)[2024−01−01]. https://arxiv.org/abs/0912.1128v1.

[9] DOSHI-VELEZ F, KIM B. Towards a rigorous science of interpretable machine learning[EB/OL]. (2017−03−02)[2024−01−01]. https://arxiv.org/abs/1702.08608v2.

[10] SIMONYAN K, VEDALDI A, ZISSERMAN A. Deep inside convolutional networks: visualising image classification models and saliency maps[EB/OL]. (2013−12−20)[2024−01−01]. https://arxiv.org/abs/1312.6034v2.

[11] SMILKOV D, THORAT N, KIM B, et al. SmoothGrad: removing noise by adding noise[EB/OL]. (2017−06−12)[2024−01−01]. https://arxiv.org/abs/1706.03825v1.

[12] DONG Yinpeng, LIAO Fangzhou, PANG Tianyu, et al. Boosting adversarial attacks with momentum[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 9185−9193.

[13] LIN Jiadong, SONG Chuanbiao, HE Kun, et al. Nesterov accelerated gradient and scale invariance for adversarial attacks[EB/OL]. (2020−02−08)[2024−01−01]. https://arxiv.org/abs/1908.06281.

[14] DONG Yinpeng, PANG Tianyu, SU Hang, et al. Evading defenses to transferable adversarial examples by translation-invariant attacks[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 4307−4316.

[15] XIE Cihang, ZHANG Zhishuai, ZHOU Yuyin, et al. Improving transferability of adversarial examples with input diversity[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 2725−2734.

[16] GAO Lianli, ZHANG Qilong, SONG Jingkuan, et al. Patch-wise attack for fooling deep neural network[C]// Computer Vision–ECCV 2020. Cham: Springer International Publishing, 2020: 307−322.

[17] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks [EB/OL]. (2019−09−04)[2024−01−01]. https://arxiv.org/abs/1706.06083.

[18] CROCE F, HEIN M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks[EB/OL]. (2020−08−04)[2024−01−01]. https://arxiv.org/abs/2003.01690v2.

[19] CROCE F, HEIN M. Minimally distorted adversarial examples with a fast adaptive boundary attack[EB/OL]. (2020−07−20)[2024−01−01]. https://arxiv.org/abs/1907.02044.

[20] 陶卿, 高乾坤, 姜纪远, 等. 稀疏学习优化问题的求解综述[J]. 软件学报, 2013, 24(11): 2498−2507.
TAO Qing, GAO Qiankun, JIANG Jiyuan, et al. Survey of solving the optimization problems for sparse learning [J]. Journal of software, 2013, 24(11): 2498−2507.

[21] ANDRIUSHCHENKO M, CROCE F, FLAMMARION N, et al. Square attack: a query-efficient black-box adversarial attack via random search[M]//Computer Vision–ECCV 2020. Cham: Springer International Publishing, 2020: 484−501.

[22] CROCE F, HEIN M. Mind the box: l_1-APGD for sparse adversarial attacks on image classifiers[EB/OL]. (2023−11−24)[2024−01−01]. https://arxiv.org/abs/2103.01208v3.

[23] CHEN Pinyu, SHARMA Y, ZHANG Huan, et al. EAD: elastic-net attacks to deep neural networks via adversarial examples[J]. Proceedings of the AAAI conference on artificial intelligence, 2018, 32(1): 10−17.

[24] SU Jiawei, VARGAS D V, SAKURAI K. One pixel attack for fooling deep neural networks[C]//IEEE Transactions on Evolutionary Computation. [S.l.]: IEEE, 2019: 828−841.

[25] MODAS A, MOOSAVI-DEZFOOLI S M, FROSSARD P. SparseFool: a few pixels make a big difference[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 9079−9088.

[26] POMPONI J, SCARDAPANE S, UNCINI A. Pixle: a fast and effective black-box attack based on rearranging pixels[C]//2022 International Joint Conference on Neural Networks. Padua: IEEE, 2022: 1−7.

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

[28] DUCHI J, SHALEV-SHWARTZ S, SINGER Y, et al. Efficient projections onto the l1-ball for learning in high dimensions[C]//Proceedings of the 25th International Conference on Machine Learning. Helsinki: ACM, 2008: 272−279.

[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. Las Vegas: IEEE, 2016: 2818−2826.

[30] SZEGEDY C, LIU Wei, JIA Yangqing, et al. Going deeper with convolutions[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015: 1−9.

[31] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770−778.

[32] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2015−04−10)[2024−01−01]. https://arxiv.org/abs/1409.1556v6.

[33] SANDLER M, HOWARD A, ZHU Menglong, et al. MobileNetV2: inverted residuals and linear bottlenecks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 4510−4520.

[34] ALEXEY D, LUCAS B, ALEXACDER K, et al. An image is worth 16×16 words: transformers for image recognition at scale[C]//International Conference on Learning Representations. Washington DC: ICLR, 2021.

[35] PASZKE, ADAM, SAM G, et al. PyTorch: an imperative style, high- performance deep learning library[EB/OL]. (2019−12−03)[2024−01−01]. https://arxiv.org/abs/1912.01703.

[36] KIM H. Torchattacks: a PyTorch repository for adversarial attacks[J]. (2021−02−19)[2024−01−01]. https://arxiv.org/abs/2010.01950.

[37] 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.

作者简介：

周强，硕士研究生，主要研究方向为机器学习、数学优化。E-mail：1071391319@qq.com。

陈军，硕士研究生，主要研究方向为机器学习、数学优化。E-mail：1358749376@qq.com。

陶卿，教授，博士生导师，博士，中国计算机学会高级会员，主要研究方向为机器学习、模式识别、应用数学。E-mail：taoqing@gmail.com。