计算机科学
COMPUTER SCIENCE

分区稀疏攻击:一种更高效的黑盒稀疏对抗攻击

温泽瑞, 姜天, 黄子健, 崔晓晖

引用本文

温泽瑞, 姜天, 黄子健, 崔晓晖. 分区稀疏攻击:一种更高效的黑盒稀疏对抗攻击[J]. 计算机科学, 2026, 53(1): 323-330.

WEN Zerui, JIANG Tian, HUANG Zijian, CUI Xiaohui. Section Sparse Attack:A More Powerful Sparse Attack Method [J]. Computer Science, 2026, 53(1): 323-330.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.241200002

分区稀疏攻击:一种更高效的黑盒稀疏对抗攻击

温泽瑞　姜　天　黄子健　崔晓晖
武汉大学国家网络安全学院空天信息安全与可信计算教育部重点实验室　武汉 430000
(zeruiwen2022@whu.edu.cn)

摘　要　深度神经网络长期以来受到对抗样本的攻击威胁,特别是黑盒攻击分类下的稀疏攻击,这类攻击依靠目标模型的输出结果来指导生成对抗样本,通常只需扰动少量像素即可达到欺骗图片分类器的目的。然而现有的稀疏攻击方法采用固定步长和欠佳的初始化策略,使得对扰动的利用率较低,导致整体攻击效率不佳。为解决这些问题,分区稀疏攻击(SSA)方法[1)]应运而生。不同于其他方法使用的固定步长策略,SSA利用历史搜索信息来自适应调整步长,从而加速对抗样本的发现过程。此外,针对不同稀疏攻击在黑盒环境中都倾向于扰动高重要性像素的共性,设计了一种基于类激活图(CAM)可解释性方法的初始化策略,使得SSA能够快速识别并初始化具有高重要性像素的种群。最后,为了在随机搜索过程中将扰动限制在关键区域内并提升扰动的利用率,提出了分区搜索策略以进一步缩小SSA的搜索空间。实验结果表明,SSA在攻击传统卷积网络和视觉Transformer模型时均表现出色。与现有的先进方法相比,SSA能够将攻击成功率提高2%～8%,效率提升近30%。

关键词:人工智能安全;对抗样本;可解释性;稀疏攻击;随机搜索
中图分类号　TP389.1

Section Sparse Attack:A More Powerful Sparse Attack Method

WEN Zerui,JIANG Tian,HUANG Zijian and CUI Xiaohui
Key Laboratory of Aerospace Information Security and Trusted Computing,Ministry of Education,School of Cyber Science and Engineering,Wuhan University,Wuhan 430000,China

Abstract　Deep neural networks(DNNs) have long been threatened by adversarial attacks,particularly sparse attacks in black-box attacks. These attacks rely on the target model's output to guide the generation of adversarial examples and typically deceive image classifiers by perturbing only a few pixels. However,existing sparse attack methods suffer from inefficiencies due to the use of fixed step-size strategies and poor initialization approaches,which fail to fully exploit perturbations. To address these issues,SSA is proposed. Unlike other methods that use fixed step sizes,SSA adapts the step size based on historical search information,thus accelerating the discovery of adversarial examples. Additionally,recognizing that sparse attacks in black-box settings tend to perturb high-importance pixels,SSA uses an initialization strategy based on the CAM,interpretability method,to quickly identify and initialize populations of high-importance pixels. Finally,to confine perturbations within critical sections and maximize their effectiveness during the search process,SSA adopts a section search strategy to reduce the search space. Experimental results demonstrate that SSA outperforms the SOTA(State-of-the-Art) methods,in attacking traditional convolutional networks and Vision Transformer(ViT) models. Specifically,SSA achieves a 2%～8% improvement in attack success rates and approximately a 30% enhancement in efficiency.

Keywords　Artificial intelligence security,Adversarial examples,Interpretability,Sparse attack,Random search

1　引言

近年来,基于深度神经网络(DNN)的人工智能(AI)技术在自动驾驶[1]、金融[2]和医疗[3]等多个领域取得了广泛应用。然而,DNN并非总是可靠的,因为它们容易受到对抗攻击的威胁[4]。根据攻击者所掌握的信息,对抗攻击可以分为白盒攻击和黑盒攻击。其中黑盒攻击更贴近实际应用场景,因为在现实场景中通常无法像白盒攻击一样完全掌握目标模型的梯度、架构以及训练数据等信息。因此,本文主要关注黑盒查询攻击,更确切地说是稀疏查询攻击。不同于在l_2或l_∞范数下的稠密攻击,关于稀疏攻击的研究较少。目前,大多数基于查询的黑盒攻击依赖于种群搜索算法[5-7],并且评估这

1) https://github.com/huge-fish/IASA-attack

到稿日期:2024-12-02　返修日期:2025-02-22
基金项目:国家重点研发计划(2024YFE0199500)
This work was supported by the National Key Research and Development Program of China(2024YFE0199500).
通信作者:崔晓晖(xcui@whu.edu.cn)

些算法的核心指标是有效性(攻击成功率)和效率(平均查询次数)。本文选择了当前效率最高的随机搜索算法,并针对现有方法的不足进行了改进,从而显著提升了攻击成功率和效率。

传统方法通常采用简单的步长调整策略[8-9],未能充分考虑历史攻击信息和稀疏对抗攻击的特性,从而导致性能受限。SSA(Section Sparse Attack)采用了一种自适应步长调整策略,根据当前迭代和历史损失信息动态调整步长,从而减少查询次数。此外,SSA 使用类激活图(Class Activation Map, CAM)[10]来选择初始种群。而许多现有方法仅随机选择初始种群或只是简单地对图像进行分割,忽略了稀疏攻击的本质,导致最终的攻击效果不佳。

回顾 JSMA(Jacobian Saliency Map)[11]、单像素攻击[12]、SparseFool[13]等经典的白盒和黑盒稀疏攻击方法,它们通常会找到与分类结果密切相关的像素,扰动这些像素能够显著提高攻击成功率。由于在黑盒攻击的条件下无法获取模型梯度,因此直接找到关键像素并不现实。然而,图像可解释性方向的相关实验结果表明,不同模型对同一图片生成的类激活图之间具有相似性[14],这为本文通过替代模型来识别关键像素提供了有利条件,进而可以实现在无需访问目标模型的情况下进一步提高攻击成功率。以图 1 中的示例来解释说明。图 1(a)—图 1(c)分别展示了目标模型在单像素攻击、JSMA 和 SparseFool 攻击下将图片中的黄貂鱼错误分类为电鳐、海胆和电鳐,而图 1(d)为基于 ResNet50[15]使用 Grad-CAM[16]生成的注意力图,用于展示目标模型在分类黄貂鱼时关注的区域。可以观察到,无论使用哪种攻击方法,大多数扰动都集中在黄貂鱼本身,再结合图 1(d)中的注意力结果可以发现,扰动的位置主要集中在类激活图中的高热度区域。

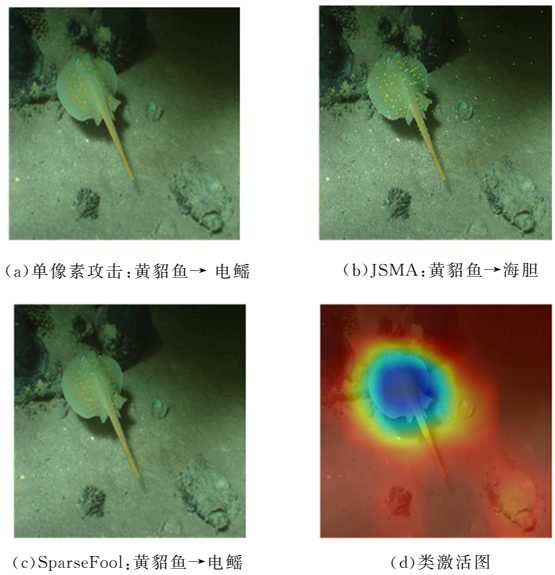


图 1 针对图像分类的l_0稀疏攻击与类激活图

Fig. 1 Sparse l_0 attack against image classification and the class activation map

最后,SSA 使用分区搜索方法将图像划分为不同区域,并将扰动集中于关键区域,进而在每个区域内分别搜索最优解。图 2 展示了 SSA 具体的实现过程,包括如何进行分区搜索、初始化种群,以及最终生成的对抗样本。得益于这些改进,与 SparseRS 相比,SSA 能够在扰动预算更为有限的情况下将攻击成功率提升 5%,同时显著降低平均查询次数和中位查询次数。

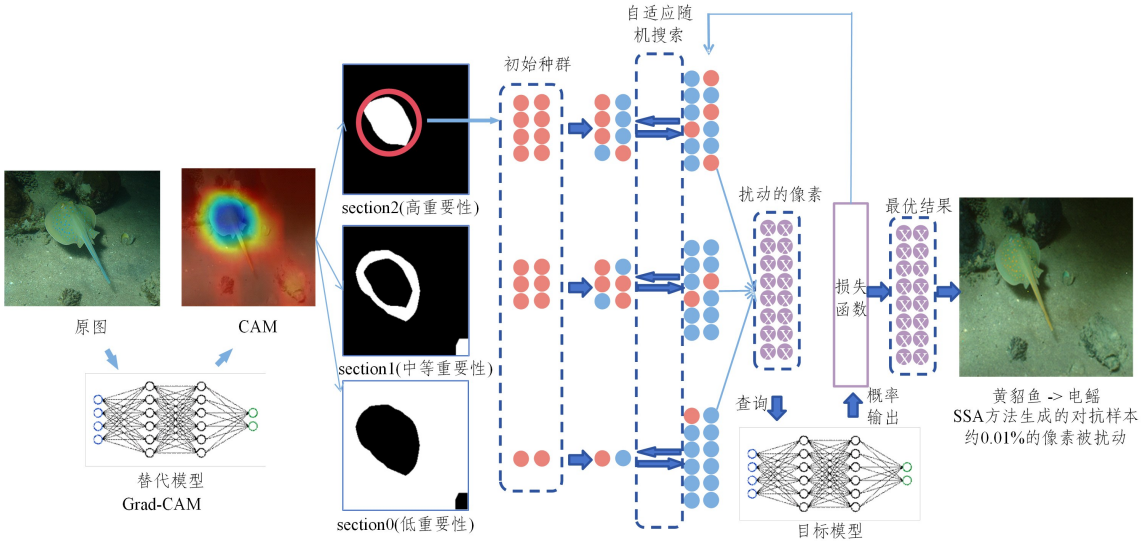


图 2 SSA 流程图

Fig. 2 Flowchart of SSA

综上所述,本文的主要贡献总结如下:

1)设计了一种自适应策略,能够根据当前搜索状态自动调整步长大小,从而显著提升攻击效率。

2)提出了一种基于l_0稀疏攻击共性特征的初始化策略,利用可解释的方法选择初始化种群,有效提升了攻击的成功率和效率。

3)为避免随机搜索过程中查询次数的浪费,进一步提出了一种基于类激活图的分区搜索方法,确保扰动得以充分利用。与传统方法相比,SSA 在扰动预算有限的情况下实现了更高的攻击成功率和效率。

4)在 ImageNet 数据集上对传统卷积神经网络和 ViT 模型进行了广泛实验。实验结果表明,与现有先进方法相比,

SSA 不仅将攻击成功率提升了 2%～8%，还将攻击效率提升了近 30%。

2 相关工作

目前许多攻击方法[17-18]基于l_2，l_1和l_∞范数。相比之下，基于 l_0范数的稀疏攻击研究相对较少。l_0范数下的攻击在测试模型鲁棒性方面同样具有重要意义，并且属于 NP-Hard 问题，更具挑战性[19]。JSMA 和单像素攻击是经典的稀疏攻击方法。JSMA 利用显著图识别最能影响模型输出的像素点并进行攻击。单像素攻击使用差分进化算法生成对抗样本，并且只需攻击图像中的少数甚至一个像素点，因而具有高度的隐蔽性。Croce 等[20]提出了 CornerSearch(CS)方法，他们在图像中搜索所有像素以找到最重要的点，然后对其进行扰动。此外，还将 PGD 攻击推广到了l_0范数。为了提高效率，Croce 等提出了 SparseRS 方法，利用随机搜索快速扰动图像。该方法也可用于生成对抗补丁和对抗帧。结果表明，SparseRS 能够实现良好的攻击成功率和效率。

一些最新的工作采用多目标优化方法[21-23]，以在隐蔽性和攻击成功率之间取得平衡。由于他们的目标是使对抗样本对不易被人眼察觉，因此在 l_0和l_2距离上限制扰动。这些方法的攻击成功率接近 SparseRS，但需要更多的查询次数。

2024 年的 BruSLeAttack[24]是最新的相关工作，其通过贝叶斯框架搜索具有高重要性的点，虽然能达到较高的攻击成功率，但需要较高的计算资源。

SSA 旨在在有限预算的情况下实现更高的效率和性能。有限预算意味着使用更少的像素点和更少的查询次数。更少的像素点意味着更难以察觉，更少的查询使得攻击更难被检测到。由于与目标模型的一系列密集查询交互会使攻击更易被发现，因此有必要限制查询次数。

3 算法设计

SSA 采用自适应随机搜索方法来生成对抗样本，其目标是满足查询效率的需求，更快速地找到一个能够攻击成功的对抗样本。在初始阶段，SSA 基于 Grad-CAM 和 Grad-CAM++[25]等算法生成的类激活图，将图像划分为不同的区域。随后，为了将扰动集中于重要性较高的区域，分别在每个区域中进行搜索。本章将详细介绍问题的形式化定义及 SSA 的具体实现细节。

3.1 问题定义和随机搜索方法

设 $f:X\in\mathbb{R}^d\to\mathbb{R}^k$为一个分类器，其将输入图像 x 分类为类别 y，θ 为对抗扰动。黑盒稀疏攻击的目标是仅通过目标模型的概率输出，找到一个扰动 θ，使得$f_i(x+\theta)\neq y$。可以通过求解以下优化问题找到 θ：

$$\begin{aligned}&\min_{\theta}\mathcal{L}(f;x+\theta)\\&\text{s.t.}\quad\|\theta\|_0\leqslant\epsilon\ \text{and}\ x+\theta\in X\end{aligned}\tag{1}$$

其中，$\mathcal{L}(\cdot)$为损失函数，$\|\theta\|_0$表示 l_0范数约束扰动，ϵ 是扰动预算。在稀疏攻击中，ϵ 表示可以被扰动的像素点的最大数量。

为高效求解式(1)，SSA 采用了类似于 SparseRS 和 Square Attack[26]的随机搜索方法。随机搜索是一种高效的算法，符合本文开发更强大且高效攻击方法的目标。SSA 中基础随机搜索方法的实现参考了 SparseRS。具体步骤如下：

1)选择初始像素集合。对于一个输入 $x\in\mathbb{R}^3$，即一张彩色图像，x 包含高乘以宽个像素。设 U 表示所有像素的集合，P 表示被扰动的像素集合。现有方法通常通过随机选择 $P\subset U$ 来生成初始像素集合。

2)扰动像素并计算损失函数。对集合 P 中的像素进行扰动，将这些像素的值分配到颜色立方体$[0,1]^d$的某个角点。损失函数采用边际损失，这一方法常用于基于评分的攻击。

$$\mathcal{L}_m(f(x+\theta),y)=f_y(x+\theta)-\max_{r\neq y}f_r(x+\theta)\tag{2}$$

式(2)表示边际损失，其意义在于，当扰动 θ 使分类器以比原始类别更高的概率将 $x+\theta$ 分类为新类别 r 时，即被视为成功攻击。

3)生成新的像素集合并评估损失改进。随机选择 $A\subset P$ 和 $B\subset U\backslash P$，其中 A 表示从已扰动集合中选择的像素子集，B 表示从未扰动像素集合中选择的像素子集。集合 A 和 B 的大小由当前迭代次数 it 和步长s_{it}决定。

可扰动像素的最大数量为 k，其作用类似于式(2) 中提到的ϵ 。因此，集合 A 和 B 的大小为 $s_{it}*k$。新的扰动像素集合定义为$P'=(P\backslash A)\cup B$。如果使用新集合后损失值有所提升，则采用新的像素集合P'。此外，基础随机搜索方法重复这一过程最多 Q 次。

3.2 自适应步长策略

先前的方法如 SparseRS 通常采用固定步长或随迭代次数衰减的策略，但在实际应用中攻击效率欠佳。因此，设计了一种基于历史搜索信息的自适应策略，以提高搜索效率。

$$\begin{aligned}&\mathcal{L}_c=\sum_{i=it-n}^{it}\|\mathcal{L}_i-\mathcal{L}_{i-1}\|\\&s_{it}=s_{it-1}-\beta\frac{\mathcal{L}_c}{\sqrt{\sum_{i=1}^{it}\mathcal{L}_i^2}+\sigma}+rs\end{aligned}\tag{3}$$

其中，$\mathcal{L}_c$反映了最近 n 步内损失函数的变化。$\mathcal{L}_c$值越大，说明在这 n 步内成功找到的对抗样本越多，这表明可以通过减小步长来缩小搜索空间，从而探索更具挑战性的解。为了防止分母过小，引入了 σ，以确保计算的稳定性。重启机制 rs 允许在搜索过程中重新开始。当计算资源充足时，这一机制可以在预算范围内探索一些难以找到的对抗样本，从而提高攻击的全面性。β为超参数。

3.3 初始种群

SSA 利用如 Grad-CAM 的可解释性方法来快速识别关键区域，并以这些区域内的像素作为初始种群，从而提升攻击的效率和效果。由于不同模型的类激活图之间具有相似性[27]，因此 SSA 可以通过替代模型生成的类激活图来识别高重要性区域，而无须访问目标模型。

使用类激活映射图可以获得一个掩码 M，M 由 0 和 1 组成，其中 $M=1$ 的区域对应重要性较高的像素，而 $M=0$ 的区域对应重要性相对较低的像素。

$$M=\begin{cases}1, & cam(x,c,f_s)>T\\0, & cam(x,c,f_s)\leqslant T\end{cases}\tag{4}$$

其中，T 是阈值，决定哪些像素点比较重要；c 为指定类别。通过 M，可以以不同的比例从高重要性和低重要性区域中选

择像素形成初始像素集 P。

$$P=s_0\eta(x\odot M)\bigcup s_0\gamma(x\backslash x\odot M),\eta+\gamma=1 \tag{5}$$

其中，$\odot$是哈达玛积。由于 P 的大部分像素点是从高重要性区域中选择的，因此设置 $\eta\gg\gamma$。

3.4 分区搜索

尽管上述改进显著提升了攻击的效率和效果，但部分扰动仍出现在图像中不重要的背景区域，这表明扰动并未被充分利用。为解决这一问题并最大化扰动的有效性，SSA 采用了一种新的搜索策略：将图像按重要性划分为多个区域，并在每个区域内分别进行独立的搜索操作。这种分区搜索策略根据不同区域的重要性分配扰动预算，确保扰动集中于指定区域内并缩小搜索空间，从而在有限扰动预算的情况下最大化使用效率。区域可以通过以下方法划分得到：

$$R=\begin{cases}2, & cam(x,c,f_s)>T_1\\ 1, & T_2\leqslant cam(x,c,f_s)\leqslant T_1\\ 0, & cam(x,c,f_s)<T_2\end{cases} \tag{6}$$

其中，T_1 和 T_2 为阈值。阈值 T_1 决定了高重要性区域的大小；而阈值 T_1 和 T_2 共同决定了中等重要性区域的大小；剩余的区域为低重要性区域。

综合上述策略，SSA 的完整流程如算法 1 所示。

算法 1 SSA 算法

输入：原图 x，标签 y，损失函数 L，最大扰动像素数量 k，最大查询次数 Q，替代模型 f_s

输出：对抗样本 x_p

1. AM=cam(x,c,f_s) //获取类激活图 AM
2. R_0,R_1,R_2⇐AM //根据 AM 将原始图片划分成 3 个区域
3. P⇐$\alpha_0R_0\cup\alpha_1R_1\cup\alpha_2R_2$//生成初始点集
4. Δ⇐$[0,1]^d$//要添加的扰动
5. z⇐x,z_p⇐Δ
6. $\mathcal{L}_m=\mathcal{L}(z)$,i⇐0 //初始化损失
7. while i<Q and attack not success do
8. for R_j in $R_{0,1,2}$ do
9. //分区搜索
10. A⊏P∩R_j,B⊏(U\P)∩R_j
11. $|A|=|B|=\alpha_j*\left(s_{i-1}-\beta\frac{\mathcal{L}_c}{\sqrt{\sum_{i=1}^{i}\mathcal{L}_i^2+\sigma}}\right)*k$ //自适应步长
12. P_j=((P∩R_j)\A)∪B
13. end for
14. P′=$P_0\cup P_1\cup P_2$,Δ′⇐$[0,1]^d$
15. z⇐x,$z_{p'}$⇐Δ′
16. if L(z)<L_m then
17. L_m⇐L(z),P⇐P′,Δ⇐Δ′ //如果损失函数减小则更新解
18. end if
19. i=i+1
20. end while
21. x_p=x+Δ
22. return x_p

4 实验

4.1 实验设计

本文所使用的数据集为 ImageNet ILSVRC2012[28]，该数据集包含 1000 个类别，图像分辨率为 224×224。在攻击实验中，从测试集中选取了 10000 张图像作为实验对象。

使用 3 种卷积神经网络模型（ResNet50，VGG16[29]，Inceptionv3[30]）和 1 种视觉 Transformer 模型（T2T-ViT[31]）。引入 ViT 模型的原因在于，现有文献中关于 ViT 模型稀疏黑盒攻击的讨论较少。此外，与传统卷积架构相比，ViT 模型架构表现出显著的差异性，导致基于迁移的攻击很难在两种模型间迁移。因此，本文将 T2T-ViT 模型作为目标模型之一，以探索在 ViT 模型上的稀疏黑盒攻击性能。

本文采用攻击成功率（Attack Success Rate，ASR）、平均查询次数（Average Number of Queries）、查询中位数（Median Number of Queries）来评估攻击性能。攻击成功率指在模型原始分类正确的样本中，攻击成功的比例。平均查询次数指在攻击过程中对目标模型进行查询的平均次数。查询中位数指查询次数的中位数，中位数越低，表示有一半的对抗样本是在较短的时间内生成的，表明攻击效率更高。此外，采用 l_1 和 l_2 距离（Distance）来衡量对抗样本的隐蔽性，其中 l_1 和 l_2 距离的计算式分别为：

$$l_1(x,x_{adv})=\sum_p|x^p-x_{adv}^p| \tag{7}$$

$$l_2(x,x_{adv})=\sqrt{\sum_p(x^p-x_{adv}^p)^2} \tag{8}$$

其中，x_{adv} 为对抗样本；p 为维度，对于彩色图片，$p=3$；l_1 和 l_2 越大，代表隐蔽性越差。

本文方法主要与 l_0 稀疏黑盒攻击方法进行对比。本文方法可以直接在目前的一些搜索算法上使用，使这些算法得以在不同重要性的区域中独立执行搜索。本文选择目前攻击效率较为出色的 SparseRS 作为主要的基线方法。虽然 BruSLeAttack是最新的研究，但其计算较 SparseRS 更繁琐，并且本文方法也可以直接在 BruSLeAttack 中使用。此外，Pixle[32] 也是最近提出的一种攻击方法，该方法通过重新排列像素位置实现高效攻击。SparseRS 的实现基于其官方提供的代码。Pixle 方法的实现基于 torchattack[33] 库。

在实验中，卷积神经网络模型使用 PyTorch 提供的预训练模型，T2T-ViT 模型使用官方预训练模型。各模型的 Top-1 准确率分别为 75.42%（ResNet50），72.71%（VGG16），68.33%（Inceptionv3）和 79.90%（T2T-ViT），其中 T2T-ViT 准确率最高。SSA 使用 Grad-CAM 生成类激活图，自适应步长参数设置为 $\beta=0.01$，$n=5$。初始种群参数为 $T=0.6$，$\eta=0.8$，$\gamma=0.2$。其他参数包括 $T_1=0.5$，$T_2=0.3$，$\alpha_0=0.1$，$\alpha_1=0.2$，$\alpha_2=0.7$，$\sigma=0.001$。对于 Pixle 攻击，不同查询次数下的重启次数分别为 4（查询次数 100）、8（查询次数 500）和 80（查询次数 5000）。实验统一设置随机种子为 0，以确保结果的公平性和可复现性。

4.2 实验结果

表 1 列出了对在 ImageNet 数据集上训练的 VGG 和 ResNet50 模型的攻击结果。表 2 列出了对在 ImageNet 数据集上训练得到的 T2T-ViT 模型的攻击结果。其中，“adv.”表示仅使用自适应策略，“cam.”表示仅更改初始化策略，“cam.+adv.”表示结合自适应策略与初始化策略。指标中的↑表示数值越大越好，↓表示数值越小越好。表 1 设置可扰动像素点数为 50 和 100，对应最大查询次数为 100 和 500。

由表 1 可知，无论扰动预算如何，本文方法在成功率方面

始终优于 SparseRS 和 Pixle 方法，成功率平均提高约 6%，而平均查询次数和查询中位数减少约 8%。尤其是在扰动预算设置为 100 个像素点、最大查询次数为 500 时，成功率相对较高，此时攻击效率的对比更加明显。从查询中位数来看，SSA 能更快速地找到潜在解，这与 SSA 的设计初衷一致。在有限预算内，更高的效率能够实现更高的攻击成功率。

表 2 列出了在不同预算下对 T2T-ViT 的攻击效果，最大可扰动像素点数为 50，最大查询次数分别为 100，500 和 5000。从结果可以看出，T2T-ViT 的对抗鲁棒性显著优于卷积神经网络(CNN)。例如，在 $K=50$ 和 $Q=100$ 时，攻击成功率仅约为 15%。因此，主要在最大查询次数为 5000 的条件下进行对比。

表 1 在 4 种不同预算条件下攻击 ResNet 和 VGG 网络的攻击成功率、平均查询次数和中位查询次数

Table 1 Attack success rate, the average number of queries, and median number of queries for attacking ResNet and VGG networks under four different budget conditions

Setting	Attack	VGG			ResNet		
		ASR↑	Avg. query↓	Med. query↓	ASR↑	Avg. query↓	Med. query↓
$K=50$, $Q=100$	SparseRS	46.63	31.7	24	40.71	31.5	23
	Pixle	33.87	33.6	27	29.16	31.3	25
	adv. (Ours)	47.74	29.5	19	41.22	29.2	19
	cam. (Ours)	49.77	29.2	20	44.61	29.5	19
	cam. +adv. (Ours)	49.78	**27.8**	**16**	43.48	**26.8**	**14**
	SSA(Ours)	**52.18**	**27.8**	17	**46.19**	30.2	21
$K=50$, $Q=500$	SparseRS	78.90	113.6	61	70.82	116.8	65
	Pixle	59.50	154.1	110	50.14	148.3	108
	adv. (Ours)	79.59	113.3	63	71.76	121.5	72
	cam. (Ours)	79.47	105.8	**53**	72.58	112.3	**57**
	cam. +adv. (Ours)	79.76	106.2	56	72.28	113.6	61
	SSA(Ours)	**81.31**	**100.7**	**53**	**74.64**	**111.9**	63
$K=100$, $Q=100$	SparseRS	62.66	28.1	18	54.78	28.3	19
	Pixle	33.87	33.6	27	29.16	31.3	25
	adv. (Ours)	63.49	26.1	14	54.75	26.2	14
	cam. (Ours)	66.72	24.7	13	59.58	25.9	15
	cam. +adv. (Ours)	65.83	23.8	**11**	58.26	**24.2**	**11**
	SSA(Ours)	**68.81**	**23.6**	**11**	**59.66**	24.4	12
$K=100$, $Q=500$	SparseRS	89.09	87.9	37	82.73	98.9	43
	Pixle	59.50	154.1	110	50.14	148.3	108
	adv. (Ours)	90.63	85.7	37	84.70	99.6	48
	cam. (Ours)	89.73	79.9	**28**	84.10	91.5	**33**
	cam. +adv. (Ours)	91.29	80.0	30	85.73	89.6	38
	SSA(Ours)	**92.08**	**74.1**	**28**	**87.24**	**88.8**	37

总体来看，本文方法在大多数情况下，在攻击成功率、平均查询次数和查询中位数方面均优于 SparseRS 和 Pixle。虽然在扰动预算特别有限的情况下，SSA 可能在查询次数方面略逊色于 SparseRS，但攻击成功率有所提升。这种现象可能归因于 T2T-ViT 的鲁棒性促使搜索在有限的最大查询次数附近更深入地寻找更多对抗样本。当最大查询次数设置为 5000 时，这一假设得到验证，攻击成功率显著提高了 2%。值得注意的是，由于 T2T-ViT 更加鲁棒，因此在查询次数有限(如$Q=100$)的情况下，无论是 SSA 还是基线方法都难以找到大量攻击成功的对抗样本，故它们在攻击成功率和效率指标上差距较小。但随着查询次数的增加，SSA 的攻击成功率和攻击效率逐渐与基线拉开差距，分区搜索策略的优势也因此得以体现。

表 2 在 4 种不同预算条件下攻击 T2T-ViT 的攻击成功率、平均查询次数和中位查询次数

Table 2 Attack success rate, average number of queries, and median number of queries for attacking T2T-ViT under four different budget conditions

Attack	$K=50, Q=100$			$K=50, Q=500$			$K=50, Q=5000$		
	ASR↑	Avg. query↓	Med. query↓	ASR↑	Avg. query↓	Med. query↓	ASR↑	Avg. query↓	Med. query↓
SparseRS	14.86	34.0	24	35.59	173.5	129	79.40	1337.6	864
Pixle	12.13	33.9	24	33.64	197.4	163	59.97	1560.1	1 210
adv. (Ours)	14.21	**29.2**	21	35.98	184.6	172	**81.10**	1137.9	662
cam. (Ours)	16.30	35.6	29	36.38	**171.8**	**115**	80.57	1340.3	886
cam. +adv. (Ours)	15.25	31.4	**20**	36.64	183.7	151	80.70	**1053.7**	596
SSA(Ours)	**16.95**	35.1	28	**38.20**	177.8	142	**81.10**	1058.7	**529**

尽管对 ViT 的改进幅度不如对卷积神经网络的改进显著，但平均查询次数和查询中位数减少了 30%～40%，总体攻击成功率也超过了 80%，展现了本文方法在 ViT 模型上的良好性能。

4.3 各策略分析实验

为验证各策略对最终结果的影响，进行了消融实验(结果见表 1 和表 2)。从实验结果可以看出，自适应策略能够加速搜索速度并提高攻击成功率。这一优势在最大查询次数较为充裕时尤为显著。例如，在攻击 T2T-ViT 的过程中，当最大查询次数分别为 500 和 5000 时，加入“adv.”策略后，攻击成功率较 SparseRS 方法有所提升，同时平均查询次数从 1337.6 降至 1137.9，查询中位数从 864 降至 662。此外，可以看到“cam. + adv.”在效率方面明显优于仅使用初始化策略的情况。

尽管初始化策略可以提高攻击成功率，但也可能导致效率下降。从查询中位数的变化可以看出，引入初始化策略通常会使查询中位数增大。在某些情况下，初始化策略的查询中位数可能较小甚至是最小的，但这些情况下攻击成功率的提升幅度并不显著。

在同时引入初始化策略和自适应策略后，各项指标对应的性能得到进一步提升，但在攻击 ResNet 时性能略有下降。这可能归因于扰动未被充分利用，表明有必要引入分区搜索策略。当引入该策略后，本文方法在绝大部分情况下均取得了最高的攻击成功率和效率。

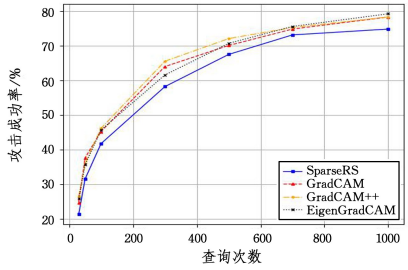
在一些特殊情况下，当查询次数接近其最大值时，本文方法有一定概率成功攻击其他方法未能攻击的样本，这可能对效率指标造成影响，但影响较小。这表明，本文提出的 3 种策略（自适应、初始化和分区搜索）是相辅相成的，每一种策略都能发挥其作用，并且只有将它们结合起来，才能充分发挥 SSA 的优势。

4.4 可解释性方法分析

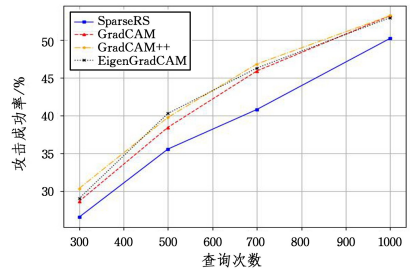
本节将讨论不同可解释方法生成的类激活图对最终攻击效果的影响。由于类激活图的生成方法很多且生成的结果各不相同，因此引出了一个问题：更先进的生成方法是否会带来更好的攻击效果？为此，选择了 3 种可解释算法进行比较，分别是 GradCAM，GradCAM++和 EigenGradCAM[29]。GradCAM++是 GradCAM 的改进版，能够更好地解释目标定位，特别是在单张图像中存在多个目标实例的情况下。EigenGradCAM 生成的结果与 GradCAM 类似，但图像更加干净。本文选用了 T2T-ViT 和 Inceptionv3 作为目标模型，并将攻击预算限制为 50 个像素点。攻击结果如图 3 所示。

从结果可以看出，任何一种注意力图生成方法的表现都显著优于 SparseRS 基准方法，即攻击成功率曲线均位于 SparseRS 之上，展现了更高的攻击效率，进一步突显了 SSA 方法的有效性。

进一步分析 3 种不同注意力图生成方法的结果发现，它们之间的差异并不显著，最优结果与最差结果的差距不超过 2%。这表明，更先进的注意力图生成方法不太可能带来显著的性能提升，即使生成的类激活图更干净，也不能显著提升攻击效果。



(a)利用不同的可解释方法对 Inceptionv3 进行攻击

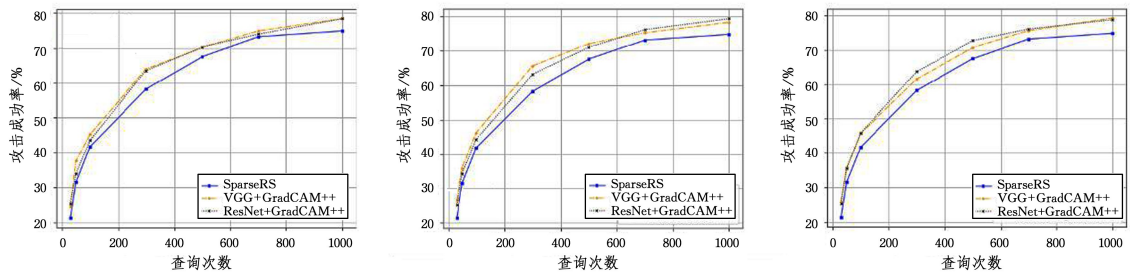


(b)利用不同的可解释方法对 T2T-ViT 进行攻击

图 3 利用不同的可解释方法对两种模型进行攻击

Fig. 3 Attacks on two models using different interpretable methods

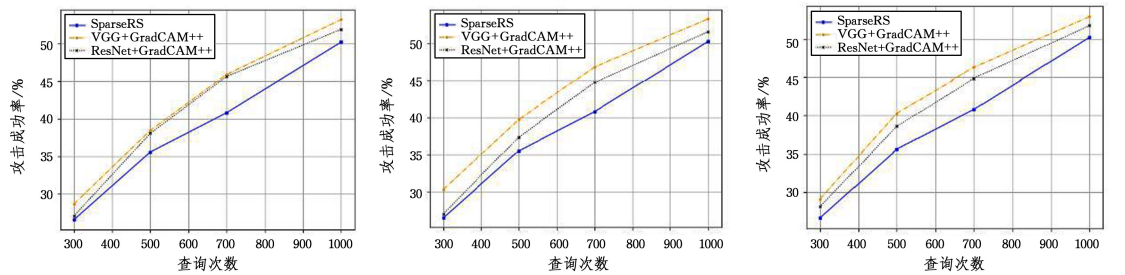
接下来探讨由不同模型生成的注意力图对最终结果的影响。选用 VGG16 和 ResNet50 作为类激活图生成模型，并将 Inceptionv3 和 T2T-ViT 作为攻击目标模型，相关结果分别如图 4 和图 5 所示。目标模型和用于类激活图生成的替代模型不一样，因此仍然保留了攻击的黑盒属性。



(a)GradCAM (b)EigenGradCAM (c) GradCAM++

图 4 利用不同模型生成的类激活图对 Inceptionv3 进行攻击

Fig. 4 Attack Inceptionv3 using class activation maps generated by different models



(a)GradCAM (b)EigenGradCAM (c) GradCAM++

图 5 利用不同模型生成的类激活图对 T2T-ViT 进行攻击

Fig. 5 Attack T2T-ViT using class activation maps generated by different models

可以观察到，无论类激活图是由 ResNet 还是 VGG 生成，最终结果均优于基准方法。此外，在大多数情况下，使用由 VGG 网络生成的类激活图的最终表现优于使用 ResNet 生成的。

4.5 时间分析

本文还对 SSA 方法的攻击耗时进行分析。时间消耗主要包括两部分：一部分是在准备阶段，根据替代模型生成的类激活图来对图像进行分区（Partition Time）；另一部分是生成对抗样本的耗时（Generation Time）。耗时分析结果如表 5 所列。时间单位为秒每批，批处理大小为 32，时间消耗为 100 个批次的平均值。

表 5 时间分析比较

Table 5 Time analysis and comparison

Attack	Partition Time	Generation Time
SparseRS	—	22.3
SSA(Ours)	0.35	26.3

可以看到，SSA 方法平均需要 0.35 秒的时间来对图片进行区域划分，其对每 32 张图片的平均搜索时间比基线方法慢 4 秒，且生成单个对抗样本的平均速度比基线方法慢 1.2 倍，即 0.14 秒。然而从整体上来看，SSA 方法的实际时间消耗仍在可接受范围内。

4.6 隐蔽性分析

最后分析对抗样本的隐蔽性。实验选取了最大扰动像素（K）为 8 和 25 两种情况，基线方法为 SparseRS 和 BruSLeAttack，并且两种基线方法与 SSA 一样均能自定义最大扰动像素。实验结果如表 6 所列，可以看到，BruSLeAttack 方法生成的样本与原始样本更接近（取得最小的 l_1 距离），但样本间的 l_2 距离偏大。总的来说，3 种方法的 l_1 和 l_2 距离结果都在同一数量级，因此 SSA 的隐蔽性仍在可接受范围内。

表 6 隐蔽性分析比较

Table 6 Stealthiness analysis and comparison

K	Attack	Distance	
		l_1	l_2
8	SparseRS	2.04	0.23
	BruSLeAttack	**1.64**	1.03
	SSA(Ours)	1.98	**0.21**
25	SparseRS	8.71	**0.46**
	BruSLeAttack	**8.23**	2.40
	SSA(Ours)	9.31	0.47

结束语 本文提出了一种更加高效的稀疏黑盒对抗攻击算法（SSA）。SSA 利用自适应策略，根据当前搜索状态动态调整步长，从而加速搜索过程。此外，SSA 分析了现有基于搜索的对抗攻击方法的特点，并采用基于可解释的方法快速提取高重要性点集，这些像素点也被用作初始化种群。相较于随机初始化方法，这种做法显著提升了攻击效率和效果。

为充分利用所有扰动，SSA 根据可解释性方法输出的结果将图像划分为不同重要性的区域，并分别在每个区域内进行搜索。实验结果显示，无论是攻击传统卷积神经网络还是 ViT，SSA 均取得了显著的效果。在攻击成功率和查询效率方面，SSA 均优于基线方法。在攻击耗时和隐蔽性方面，SSA 带来的额外时间消耗在可接受范围内并且隐蔽性与基线方法保持一致。

参 考 文 献

[1] LIU J H, GUANG J C, FANG H Q, et al. Efficient View Transformation for Autonomous Driving[J]. Computer Systems and Applications, 2025, 34(2): 246-253.

[2] CHENG C Z. Financial Time Series Prediction Based on Deep Learning[D]. Chengdu: University of Electronic Science and Technology of China, 2021.

[3] WANG K. Research on Medical Image Classification and Segmentation Based on Deep Learning[D]. Changsha: National University of Defense Technology, 2022.

[4] SZEGEDY C. Intriguing properties of neural networks [J]. arXiv: 1312.6199, 2013.

[5] LI Z, CHENG H, CAI X, et al. Sa-es: Subspace activation evolution strategy for black-box adversarial attacks[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2022, 7(3): 780-790.

[6] WILLIAMS P N, LI K. Black-box sparse adversarial attack via multiobjective optimisation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 12291-12301.

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

[8] CROCE F, ANDRIUSHCHENKO M, SINGH N D, et al. Sparse-rs: a versatile framework for query-efficient sparse black-box adversarial attacks[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2022: 6437-6445.

[9] JI S H, HU L, ZHANG P C, et al. Adversarial Example Generation Method Based on Sparse Perturbation[J]. Journal of Software, 2023, 34(9): 4003-4017.

[10] ZHOU B, KHOSLA A, LAPEDRIZA A, et al. Learning deep features for discriminative localization[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2921-2929.

[11] 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 (EuroS&P). IEEE, 2016: 372-387.

[12] SU J, VARGAS D V, SAKURAI K. One pixel attack for fooling deep neural networks[J]. IEEE Transactions on Evolutionary Computation, 2019, 23(5): 828-841.

[13] MODAS A, MOOSAVI-DEZFOOLI S M, FROSSARD P. Sparsefool: a few pixels make a big difference[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 9087-9096.

[14] WU W, SU Y, CHEN X, et al. Boosting the transferability of adversarial samples via attention[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 1161-1170.

[15] HE K, ZHANG X, REN S, et al. Deep residual learning for

image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:770-778.
[16] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization [C]// Proceedings of the IEEE International Conference on Computer Vision. 2017:618-626.
[17] LI W T, XIAO R, YANG X. Improving Transferability of Adversarial Samples Through Laplacian Smoothing Gradient[J]. Computer Science, 2024, 51(S1):938-943.
[18] CHEN J Y, CHEN Y Q, ZHENG H B, et al. Black-box Adversarial Attack Against Road Sign Recognition Model via PSO[J]. Journal of Software, 2020, 31(9):2785-2801.
[19] DONG X, CHEN D, BAO J, et al. Greedyfool: Distortion-aware sparse adversarial attack[J]. Advances in Neural Information Processing Systems, 2020, 33:11226-11236.
[20] CROCE F, HEIN M. Sparse and imperceivable adversarial attacks[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019:4724-4732.
[21] BAI Z X, WANG H J. Adversarial Example Generation Method Based on Improved Genetic Algorithm[J]. Computer Engineering, 2023, 49(5):139-149.
[22] LI Z, CHENG H, CAI X, et al. Sa-es: Subspace activation evolution strategy for black-box adversarial attacks[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2022, 7(3):780-790.
[23] WANG H, ZHU C, CAO Y, et al. ADSAttack: An Adversarial Attack Algorithm via Searching Adversarial Distribution in Latent Space[J]. Electronics, 2023, 12(4):816.
[24] VO V Q, ABBASNEJAD E, RANASINGHE D C. BruSLeAttack: A Query-Efficient Score-Based Black-Box Sparse Adversarial Attack[J]. arXiv:2404.05311, 2024.
[25] CHATTOPADHAY A, SARKAR A, HOWLADER P, et al. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks[C]// 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018:839-847.
[26] ANDRIUSHCHENKO M, CROCE F, FLAMMARION N, et al. Square attack: a query-efficient black-box adversarial attack via random search[C]// European Conference on Computer Vision. Cham: Springer, 2020:484-501.
[27] LIN C, HAN S, ZHU J, et al. Sensitive region-aware black-box adversarial attacks[J]. Information Sciences, 2023, 637:118929.
[28] DENG J, DONG W, SOCHER R, et al. Imagenet: A large-scale hierarchical image database[C]// 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009: 248-255.
[29] SIMONYAN K. Very deep convolutional networks for large-scale image recognition[J]. arXiv:1409.1556, 2014.
[30] 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.
[31] YUAN L, CHEN Y, WANG T, et al. Tokens-to-token vit: Training vision transformers from scratch on imagenet[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021:558-567.
[32] 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 (IJCNN). IEEE, 2022:1-7.
[33] KIM H. Torchattacks: A pytorch repository for adversarial attacks[J]. arXiv:2010.01950, 2020.
[34] BANY MUHAMMAD M, YEASIN M. Eigen-CAM: Visual explanations for deep convolutional neural networks[J]. SN Computer Science, 2021, 2(1):47.

WEN Zerui, born in 2000, postgraduate, is a member of CCF(No. L7493G). His main research interests include artificial intelligence security and adversarial attack.

CUI Xiaohui, born in 1971, Ph.D, professor, Ph.D supervisor, is a member of CCF(No. 36210S). His main research interests include big data, blockchain technology, food safety and high performance computing.

(责任编辑:何杨)