

基于改进遗传算法的对抗样本生成方法

白祉旭,王衡军

(战略支援部队信息工程大学,郑州 450001)

摘要: 对抗样本是评估模型安全性和鲁棒性的有效工具,对模型进行对抗训练能有效提升模型的安全性。现有对抗攻击按主流分类方法可分为白盒攻击和黑盒攻击两类,其中黑盒攻击方法普遍存在攻击效率低、隐蔽性差等问题。提出一种基于改进遗传算法的黑盒攻击方法,通过在对抗样本进化过程中引入类间激活热力图解释方法,并对原始图像进行区域像素划分,将扰动进化限制在图像关键区域,以提升所生成对抗样本的隐蔽性。在算法中使用自适应概率函数与精英保留策略,提高算法的攻击效率,通过样本初始化、选择、交叉、变异等操作,在仅掌握模型输出标签及其置信度的情况下实现黑盒攻击。实验结果表明,与同是基于遗传算法的POBA-GA黑盒攻击方法相比,该方法在相同攻击成功率下生成的对抗样本隐蔽性更好,且生成过程中模型访问次数更少,隐蔽性平均提升7.14%,模型访问次数平均降低6.43%。

关键词: 对抗样本;遗传算法;热力图;白盒攻击;黑盒攻击

开放科学(资源服务)标志码(OSID):



中文引用格式:白祉旭,王衡军.基于改进遗传算法的对抗样本生成方法[J].计算机工程,2023,49(5):139-149.

英文引用格式:BAI Z X, WANG H J. Adversarial example generation method based on improved genetic algorithm[J]. Computer Engineering, 2023, 49(5): 139-149.

Adversarial Example Generation Method Based on Improved Genetic Algorithm

BAI Zhixu, WANG Hengjun

(Strategic Support Force Information Engineering University, Zhengzhou 450001, China)

[Abstract] The adversarial example is an effective tool to evaluate the security and robustness of a model. Conducting antagonism training on the model can effectively improve the model's security. The mainstream classification methods divide the existing counterattacks into white-box attack and black-box attack. Black-box attack methods generally have problems of low attack efficiency and poor concealment. Thus, a black-box attack method based on an improved Genetic Algorithm (GA) is proposed. A Class Activation Mapping (CAM) interpretation method is introduced in the process of adversarial sample evolution. In addition, the original image is divided into regional pixels to restrict the perturbation evolution to essential regions of the image to improve the concealment of the generated adversarial examples. An adaptive probability function is introduced with an elite retention strategy to improve the algorithm's attack efficiency and achieve black-box attacks by sample initialization, selection, crossover, and mutation operations with only the model output labels and their confidence levels. The experimental results show that, compared with the POBA-GA black-box attack method that also uses a genetic algorithm, the adversarial examples generated by the improved genetic algorithm-based adversarial sample generation method have better steganography and fewer model visits for the same attack success rate. The average increase of 7.14% and the average decrease of 6.43% in the number of model visits validate the method's effectiveness.

[Key words] adversarial example; Genetic Algorithm (GA); heat map; white-box attack; black-box attack

DOI: 10.19678/j.issn.1000-3428.0065260

0 概述

随着深度学习的崛起,人工智能在计算机视觉、自然语言处理等^[1-3]领域发挥着越来越重要的作用。

在大数据的支撑下,深度神经网络在图像识别和目标检测领域中拥有惊人的表现,使得大量基于深度学习的应用得以投入使用,尤其在与安全相关的领域^[4-6],如人脸识别系统、自动驾驶汽车、无人机侦察

基金项目:国家重点研发计划(2017YFB0801904)。

作者简介:白祉旭(1992—),男,硕士研究生,主研方向为人工智能安全;王衡军,副教授、博士。

收稿日期:2022-07-15 修回日期:2022-08-22 E-mail: b347072272@163.com

等^[7-8]。深度神经网络的安全问题逐渐成为人工智能安全领域的研究重点。尽管深度神经网络在大多数任务中表现良好,但在面对对抗样本时显得十分脆弱。对抗样本是在数据集中通过故意添加细微扰动形成的一类能够诱导学习模型错误分类的样本,其存在给模型的安全形成了严重威胁。另一方面,对于模型设计者来说,对抗样本可以作为评估模型安全性和鲁棒性的有效工具,通过对抗训练,能有效提升模型的安全性。

对抗攻击按主流分类方法分为白盒攻击和黑盒攻击两类。虽然白盒攻击的效果更好、成果更多,但在现实攻击场景中,攻击者通常难以获得模型的内部结构和相关参数,因此面向黑盒攻击的研究更具现实意义^[9-11]。主流的黑盒攻击大致可以分为两类。一是基于模型迁移的攻击方法,该方法使用替代模型生成对抗样本,根据对抗样本的可迁移性,完成黑盒攻击。但不同模型的结构和参数通常存在差异,最终导致黑盒攻击成功率偏低。二是基于查询反馈的攻击方法,该类方法根据预先设计的算法生成输入样本,随后通过查询输出结果,调整输入,直至完成攻击。但由于此类攻击的模型访问次数过多,导致攻击效率不高,且对抗样本的隐蔽性较差。

针对上述黑盒攻击面临的挑战,本文提出一种基于改进遗传算法的对抗样本生成方法。通过掌握模型输出标签及其置信度,完成有效攻击。引入 Grad-CAM 解释方法,有效提升对抗样本的隐蔽性。通过使用自适应概率函数与精英保留策略,有效减少算法的模型访问次数。

1 相关工作

1.1 对抗样本的生成方法

对抗样本作为一类被用来恶意攻击深度神经网络的样本,能够有效利用模型自身的脆弱性对其进

行攻击,以诱导模型误分。目前该领域提出多种对抗样本生成方法,依据攻击预期、模型信息等不同标准可将其分为特定目标攻击与非特定目标攻击、白盒攻击与黑盒攻击,下面对其中较具代表性的攻击方法进行介绍。

SZEGEDY 等^[12]提出“对抗样本”这一概念,对抗样本利用难以察觉的扰动,使网络的预测误差最大化,导致网络进行错误分类。GOODFELLOW 等^[13]研究发现深度神经网络模型在高维空间中呈线性是对抗样本产生的主要原因,并提出快速梯度符号法(Fast Gradient Sign Method, FGSM),这是一种单步对抗攻击方法,为后续 FGSM 类方法奠定了基础。DONG 等^[14]在快速梯度符号法的基础上引入动量,提出 MI-FGSM 法,有效提升了所生成对抗样本的可迁移性,使黑盒攻击成功率明显提升。CARLINI 等^[15]提出 C&W 攻击方法,成功对防御蒸馏网络进行攻击。BRENDDEL 等^[16]提出 Boundary 方法,通过在决策边界附近搜索距离最近的对抗样本完成攻击。CHEN 等^[17]提出一种零阶优化的 ZOO 方法,通过与目标模型交互进行梯度估计,完成对抗攻击。TU 等^[18]在 ZOO 的基础上提出 AutoZOOM 方法,通过加入自编码的方法,减少算法的模型访问次数。DONG 等^[19]提出基于进化策略的 Face Attack 方法,通过与模型交互,对模型分类边界建模,估计对抗样本的移动方向,从而自适应调节算法参数,完成对抗攻击。XIE 等^[20]提出 DIM 方法,通过扩展输入数据的多样性,提升对抗样本的可迁移性。CHEN 等^[21]提出 POBA-GA 方法,使用遗传算法(Genetic Algorithm, GA)^[22]并引入新的扰动适应度函数实现高效的黑盒攻击。本文对比了8种具有代表性的对抗样本生成方法,如表1所示,其中“特/非”表示特定目标攻击与非特定目标攻击,*表示所生成对抗样本的攻击能力,*越多,表示能力越强。

表1 对抗样本生成方法的对比

Table 1 Comparison of adversarial example generation methods						
方法	分类	目标	方式	攻击能力	优点	缺点
FGSM 方法	白盒	特/非	单步	***	方法简单,攻击效率高	单步攻击,攻击成功率偏低
MI-FGSM 方法	白盒	特/非	迭代	****	梯度更新方向稳定,对抗攻击性强	可迁移性差,黑盒攻击成功率不高
C&W 方法	白盒	特/非	迭代	*****	攻击成功率高,隐蔽性好	可迁移性差,黑盒攻击成功率低
DIM 方法	黑盒	特/非	迭代	****	引入数据增强,可迁移性好	计算成本高,隐蔽性不强
Boundary 方法	黑盒	特/非	迭代	****	思路简单,隐蔽性好	攻击效率低,攻击成功率不高
ZOO 方法	黑盒	特/非	迭代	****	对抗性训练中表现好,更接近原图	攻击效率低,隐蔽性不强
AutoZOOM 方法	黑盒	特/非	迭代	*****	攻击成功率高,攻击效率高	隐蔽性差
Face Attack 方法	黑盒	特/非	迭代	*****	成功率高,隐蔽性好	时间成本高,计算复杂

1.2 遗传算法

遗传算法起源于对自然界中生物种群的计算机模拟研究,通过模仿种群的自然选择和自然遗传,实现待求解问题的随机全局搜索和优化。该算法的基本操作步骤包括种群初始化、个体适应度的计算、遵循适者生存原则对优良个体进行选择、将挑选出的优秀个体作为父代、对父代染色体的基因进行随机交叉与变异、生成下一代种群等,重复进行上述步骤,直到满足进化终止条件。遗传算法本质上是一种高效率的、能够并行运算的全局寻优方法,能够在搜索过程中学习并存储搜索空间中的有效信息,以自适应的方式操纵搜索过程,得到最优解。

遗传算法作为一种典型的优化工具被广泛应用于各种场景中,实践证明,该算法能取得良好的优化性能。对抗样本的生成过程本质上就是求解最优优化问题的过程,因此,可以很自然地使用遗传算法,在不完全掌握模型内部结构与相关参数的情况下,高效生成对抗样本。

1.3 Grad-CAM解释方法

ZHOU 等^[23]提出深度神经网络能够定位出图像中的关键区域,并将结果以类间激活热力图(Class Activation Mapping, CAM)的形式量化表示,用以帮助解释卷积神经网络(Convolutional Neural Network, CNN)的预测结果。但由于在较多常见的网络中并没有该方法使用的全局平均池化(Global Average Pooling, GAP)层,导致 CAM 方法应用受限。在 CAM 的基础上,SELVARAJU 等^[24]提出基于梯度的类激活热力图(Gradient-weighted CAM, Grad-CAM)。Grad-CAM 的生成过程如图 1 所示。

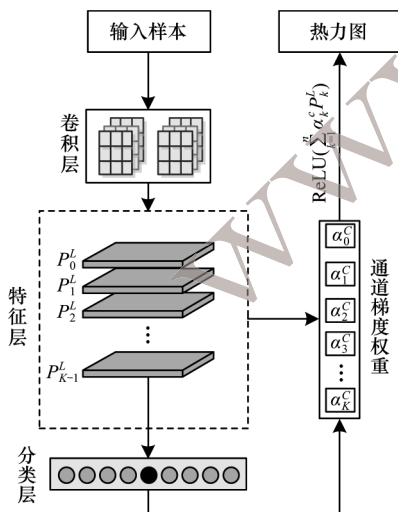


图 1 Grad-CAM 的生成过程示意图

Fig.1 Schematic diagram of the Grad-CAM generation process

由图 1 可知,Grad-CAM 的基本操作与 CAM 的操作相似,只是为了解除由 GAP 带来的限制。Grad-CAM 使用反向传播中获取的通道梯度均值作为通道权重,如式(1)所示:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{1}$$

其中: y^c 表示分类结果; A_{ij}^k 表示特征图 k 中坐标为 (i,j) 的值; $\frac{\partial y^c}{\partial A_{ij}^k}$ 表示分类结果对 CNN 最后一个卷积层进行求导运算;求和运算 $\sum_i \sum_j$ 表示求第 k 个通道中的梯度; Z 表示归一化因子。最后将特征图与对应的通道梯度权重逐点相乘,得到 CNN 中关注区域的热力图,如式(2)所示:

$$L_{\text{Grad-CAM}}^c = \text{ReLU} \left(\sum_{k=1}^n \alpha_k^c A_k^L \right) \tag{2}$$

其中:ReLU 函数的作用是将特征图中对分类结果影响的正值部分筛选出来。

2 对抗样本的生成

本文提出一种基于改进遗传算法的对抗样本生成方法(Adversarial Examples Generation Method based on Improved Genetic Algorithm, AEGM-IGA)。将 Grad-CAM 解释方法引入到种群进化中,根据图像的热力值对像素进行分组,将扰动的进化过程控制在图像中的关键区域,以减少图像中添加的扰动,提高对抗样本的隐蔽性。在种群进化过程中,引入自适应机制,根据种群适应度的动态变化,自动调整交叉、变异概率,加快种群的收敛速度,以减少算法的模型访问次数。最后在算法中加入精英保留机制,避免因交叉、变异操作引起的优秀基因丢失。AEGM-IGA 算法的流程如图 2 所示,其具体流程如下:

- 1)生成输入图像的热力图,并将其值域 M 等分。
- 2)选定扰动进化区域,在像素中随机添加扰动,生成初始种群。
- 3)计算种群中每个样本的适应度函数值。
- 4)使用选择算子选出高适应度的样本构成父代。
- 5)判断是否满足终止条件或达到最大迭代次数,如果是,则输出对抗样本;否,则执行步骤 6)。
- 6)对父代种群进行自适应交叉、变异操作,完成本轮种群进化。
- 7)判断是否进行精英保留,是则重复步骤 6),否则跳转步骤 3)。

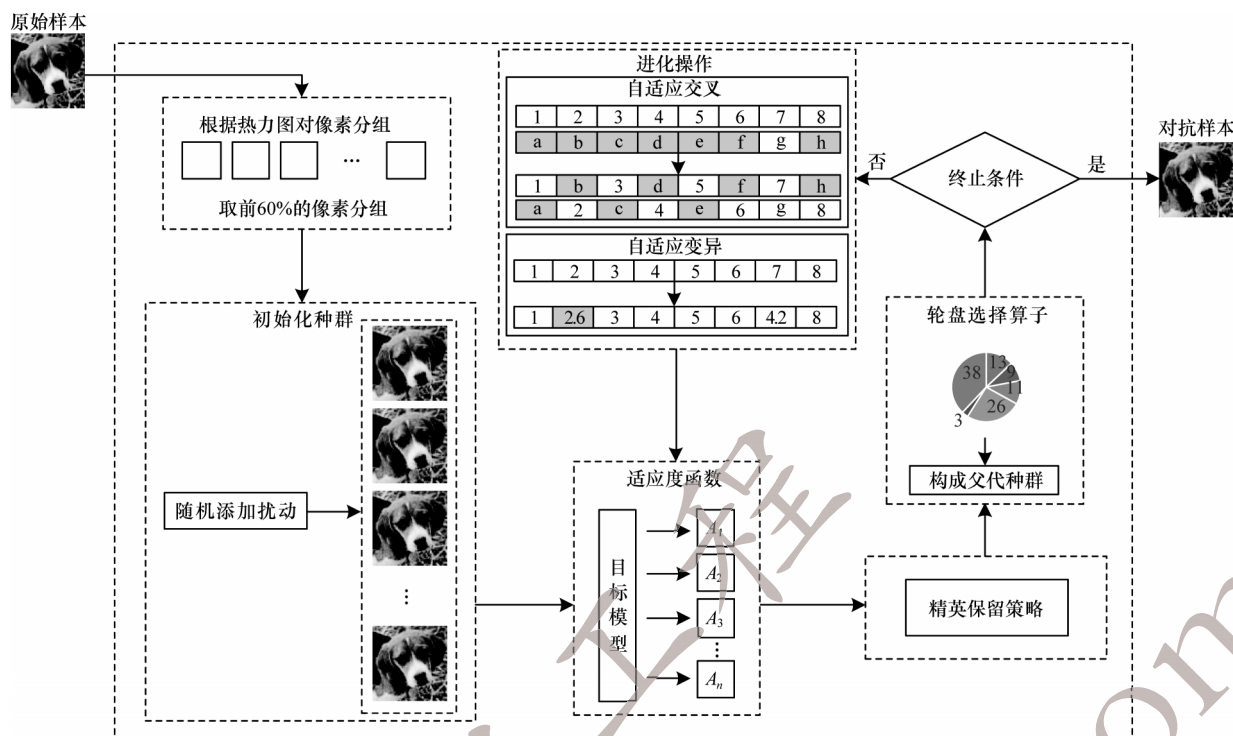


图2 AEGM-IGA算法的流程

Fig.2 Procedure of AEGM-IGA algorithm

2.1 种群初始化

初始化的目的是在开始时生成初始解,初始解的质量将直接影响算法迭代。初始解与近似最优解越接近,算法收敛越快,否则需要多次迭代才能搜索到全局最优解。将 Grad-CAM 解释方法引入到种群初始化中,在图像关键区域添加随机扰动生成初始解,并将扰动进化区域限制在图像关键区域,以提升样本的隐蔽性,加速算法收敛。

2.1.1 热力值分组方法

由于图像中的关键区域对分类影响最大,针对该区域进行对抗攻击,既能提高攻击的成功率,又能有效减少图像中添加的扰动,因此本文选择模型重点关注区域作为扰动进化区域。为确保不同模型之间拥有相似的关注区域,本文以能占比 (Energy Ratio, ER) 作为标准,衡量两张热力图的相似程度,即同一张图片在不同模型下热力图差值小于 10% 的区域占整张图的比例,表达式如式 (3) 所示:

$$E_{ER} = \frac{\sum_{i,j \in T} (H_{GradCAM}^c(x))_{i,j}}{\sum_{i,j \in T} (H_{GradCAM}^c(x))_{i,j} + \sum_{i,j \notin T} (H_{GradCAM}^c(x))_{i,j}} \times 100\% \quad (3)$$

其中: T 表示相似关注区域像素的集合。同一张图片在不同模型中热力图的相似程度如图 3 所示。通过图 3 可以明显看出同一张图片在不同模型中生成的热力图有超过 65% 的区域关注程度十分相似。

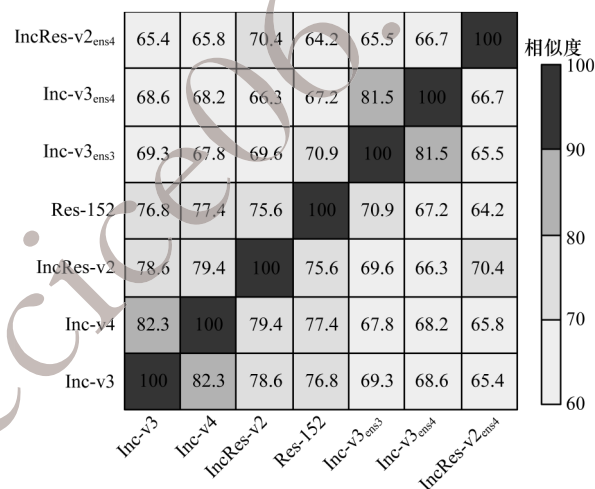


图3 不同模型中热力图的相似程度

Fig.3 Degree of similarity of heat maps in different models

本文首先使用替代模型 DM 计算图像 X 的热力图 h_{DM} 并做归一化处理,将其值域平均分为 M 个区间,并升序排列。随后将图像中的像素点按对应关系分配到这 M 个组中,如式 (4) 所示:

$$G[h] = \left\{ g \mid g \in \left(\frac{h-1}{M}, \frac{h}{M} \right] \right\}, h = 1, 2, \dots, M \quad (4)$$

其中: $G[h]$ 表示第 h 个热力值区间像素点的集合; g 表示分组 h 中的像素点,下标 h 越大,表示模型对该区域越关注。

2.1.2 扰动初始化

对于图像 X ,本文的目的是进化出一个近似最优的对抗样本 X' 。在随机扰动初始化中,使用不同

方法、不同噪声点个数和不同噪声大小,随机产生不同类型的初始扰动,以增加初始扰动的多样性,随后将生成的扰动添加至热力值分组,如式(5)所示:

$$\begin{cases} X^{t=0} = G_h + \delta'_h, \frac{h}{M} \geq w \\ \delta \sim \mathcal{N}(\mu, \sigma^2) \end{cases} \quad (5)$$

其中: $X^{t=0}$ 表示初始化对抗样本; t 表示迭代轮数; G_h 表示被选中的热力值分组; δ'_h 表示第 h 组热力值分组中添加的随机扰动; μ 和 σ^2 分别表示期望和方差; w 为控制参数,用来控制进化区域的大小, $w \in (0, 1]$,当 w 趋于0时,表示选中整张图,随机添加扰动,当 w 趋向1时,表示仅在模型关注度最高的分组中添加扰动。通过改变 w 的取值,调节图像中添加扰动的范围。

本文以热力值分组作为基因单元,完成种群初始化及后续进化中的交叉、变异操作,每轮进化中的对抗样本由原始图像及本轮得到的扰动,按热力值分组叠加形成, $X'_i = X + \delta'_i$ 。

2.2 适应度函数

为进化出近似最优的对抗样本,适应度函数应能筛选出攻击成功率高且隐蔽性好的样本。

2.2.1 扰动评价函数与攻击性评价函数

本文使用文献[21]中提出的改进扰动评估方法,该方法考虑了扰动数量与幅度之间的关系,能够准确地衡量扰动大小,且比 L_2 方法更接近肉眼评估,该方法的表达式如式(6)所示:

$$D(\delta'_i) = \sum_{a=1}^{m_a} \sum_{b=1}^{m_b} \left(\frac{1}{1 + e^{-|X'_{i(ab)}| \times p_{m1} + p_{m2}}} - \frac{1}{1 + e^{p_{m3}}} \right) \quad (6)$$

其中: $D(\delta'_i)$ 表示扰动评估函数,用来衡量扰动大小; X'_i 表示第 t 轮迭代中的第 i 个对抗样本; δ'_i 表示 X'_i 中的扰动; m_a 和 m_b 表示图像的长和宽; $X'_{i(ab)}$ 表示图像中第 a 行、 b 列的扰动像素点; p_{m1} 与 p_{m2} 表示调整扰动像素映射规则的参数。肉眼评估时 $p_{m1}=10, p_{m2}=5.8$,机器评估时 $p_{m1}=15, p_{m2}=3$ 。

使用式(7)进行样本攻击性能的衡量:

$$A(X'_i) = \begin{cases} p(y_1 | X'_i) - p(y_0 | X'_i), y_1 \neq y_0 \\ p(y_2 | X'_i) - p(y_0 | X'_i), y_1 = y_0 \end{cases} \quad (7)$$

其中: $A(X'_i)$ 表示样本的攻击性能; y_0 表示原始样本的真实标签; y_1 与 y_2 分别表示对抗样本在目标模型分类结果中排名第1与第2的标签; $p(y | X'_i)$ 表示对抗样本被目标模型标记为标签 y 的置信度。标签 y_1 与标签 y_0 置信度的差值越大,表示攻击能力越强。

2.2.2 本文适应度函数

为了将攻击成功率高且隐蔽性好的对抗样本筛选出来,本文构造适应度函数,其表达式如(8)所示:

$$\phi(X'_i) = A(X'_i) - \frac{\alpha}{\max D(\delta^0)} D(\delta'_i) \quad (8)$$

其中: $\phi(X'_i)$ 表示适应度函数; $A(X'_i)$ 表示样本 X'_i 的攻击性能; $D(\delta'_i)$ 表示样本中添加扰动的大小; α 为平

衡系数,用于调整攻击性能与扰动大小的比例。当 α 较小时,适应度函数更关心样本的攻击性能,当 α 较大时,表示适应度函数更加关心样本的隐蔽性。

2.3 遗传操作的改进

本文使用自适应交叉、变异概率,根据种群适应度函数值,自动调整交叉、变异概率,以加速算法收敛速度。此外,引入精英保留机制,将每一代进化中适应度函数值高的精英样本直接遗传至下一代。

2.3.1 选择算子

使用轮盘赌算法作为选择算子,实现对优秀样本的选择。在轮盘赌算法中,选中一个样本的概率与该样本的适应度函数值成正比,能够在确保适应度高的样本被保留下来的同时,避免适应度低的样本被淘汰,造成种群多样性的减少。

轮盘赌算法的流程如下:

- 1) 确定种群大小 G ,并计算个体适应度 $\phi(X'_i)$ 。
- 2) 计算第 t 轮中样本 i 的概率 $P(X'_i)$,如式(9)所示:

$$P(X'_i) = \frac{\phi(X'_i)}{\sum_{i=1}^n \phi(X'_i)} \quad (9)$$

- 3) 计算样本 i 的累积概率,计算式如式(10)所示:

$$Q(X'_i) = \sum_{j=1}^i P(X'_j) \quad (10)$$

- 4) 产生随机数 $r, r \in [0, 1]$ 。

- 5) 判断 $r < Q_1$,是则选择样本1,否则选择样本 k ,使得 $Q_{k-1} < r < Q_k$ 成立。

- 6) 跳至步骤4),直至完成样本选择。

2.3.2 交叉概率与变异概率的自适应调整

为使进化中的交叉、变异概率能够根据种群适应度的变化动态调整,计算交叉、变异概率,计算式如下所示:

$$P_c = \begin{cases} \frac{k_1(\phi(X)_{\max} - \phi(X))}{\phi(X)_{\max} - \phi(X)_{\text{avg}}}, \phi(X) \geq \phi(X)_{\text{avg}} \\ k_2, \phi(X) \leq \phi(X)_{\text{avg}} \end{cases} \quad (11)$$

$$P_m = \begin{cases} \frac{k_3(\phi(X)_{\max} - \phi(X)')}{\phi(X)_{\max} - \phi(X)_{\text{avg}}}, \phi(X)' \geq \phi(X)_{\text{avg}} \\ k_4, \phi(X)' < \phi(X)_{\text{avg}} \end{cases} \quad (12)$$

其中: P_c 表示交叉概率; P_m 表示变异概率; $\phi(X)_{\max}$ 表示种群中最高的适应度; $\phi(X)_{\text{avg}}$ 表示种群适应度函数的均值; $\phi(X)$ 表示参与交叉的两个样本中适应度较大者; $\phi(X)'$ 表示变异个体的适应度;其中 k_1, k_2, k_3 与 k_4 为常数,根据文献[25],分别设 $k_1=0.6, k_2=0.9, k_3=0.03, k_4=0.005$,从而根据个体的适应度动态调整交叉概率与变异概率,加速种群的前进进化与后期的收敛。

自适应交叉、变异概率的作用主要有2个:一是当种群内样本适应度差异较小时,保持较大的交叉、变异概率能够扩大搜索空间、加速种群进化;二是当

种群中样本适应度普遍较高时,降低交叉、变异概率,能够加速种群收敛。

2.3.3 均匀交叉算子

本文使用均匀交叉算子,从选择的父代样本中生成子代样本,即父代每个位点的基因以相同的概率交换,如图4所示。

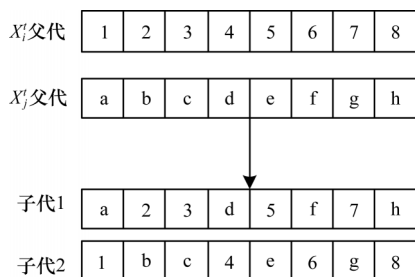


图4 交叉算子示意图

Fig.4 Schematic diagram of crossover operator

本文以扰动分组作为基因单元进行交叉运算,计算式如下所示:

$$\delta'_{\text{cross}_i} = \begin{cases} \delta'_i \times B + \delta'_j \times (1 - B), \text{rand}(0, 1) < P_c \\ \delta'_i, \text{其他} \end{cases} \quad (13)$$

$$\delta'_{\text{cross}_j} = \begin{cases} \delta'_i \times (1 - B) + \delta'_j \times B, \text{rand}(0, 1) < P_c \\ \delta'_j, \text{其他} \end{cases} \quad (14)$$

其中: δ'_i 表示父代第*t*轮迭代中样本*i*中的扰动; δ'_{cross_i} 表示新产生的子代;*B*表示与基因序列等长的选择序列,由随机的0与1组成,将基因序列与选择序列逐位相与,分割父代基因; $(1-B)$ 表示对*B*取反。

2.3.4 变异算子

本小节采用多点突变,变异算子示意图如图5所示。

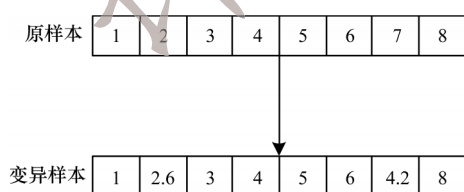


图5 变异算子示意图

Fig.5 Schematic diagram of variational operator

在子代扰动中随机选取几个扰动分组基因,依据自适应变异概率 P_m 进行变异操作,如式(15)所示:

$$\delta'_{\text{mutation}_i} = \begin{cases} \delta'_{\text{cross}_i} \times C, \text{rand}(0, 1) < P_m \\ \delta'_{\text{cross}_i}, \text{其他} \end{cases} \quad (15)$$

其中: δ'_{cross_i} 表示第*t*轮中交叉生成的子代扰动; $\delta'_{\text{mutation}_i}$ 表示本轮变异产生的扰动;*C*表示与基因序列等长的变异序列,除了极少数位置上为0到2之间的数外,其余位上全部为1。

2.3.5 精英保留策略

为避免适应度高的样本在交叉和变异过程中丢失,在遗传算法中引入精英保留策略。将第*t*轮迭代中适应度最大的样本 X'_{max} 作为第*t+1*轮迭代中的阈值,若第*t+1*轮迭代中样本适应度均小于该值,则将 X'_{max} 保留,直接进入第*t+1*轮,并将第*t+1*轮中适应度最小的样本

淘汰,以保证种群总数不变,如式(16)所示:

$$K(X) = \begin{cases} 1, X'_i < X'_{\text{max}} \\ 0, \text{其他} \end{cases} \quad (16)$$

其中: X'_i 表示第*t*轮迭代中第*i*个样本。 $K(X)$ 表示保留函数,当 $K(X)=1$ 时,对 X'_{max} 进行保留操作,当 $K(X)=0$ 时,则不保留,正常进入下一轮进化。精英样本作为截止本轮进化中适应度最高的个体,将其保留对全局收敛性作用巨大。AEGM-IGA算法的具体如下。

算法1 AEGM-IGA算法

输入 原始样本*X*,目标模型TM,替代模型DM,参数*M*,*w*, α ,最大迭代次数*T*,种群大小*G*

输出 近似最优的对抗样本 X'

1. $h_{\text{DM}} \leftarrow \text{Normalize}(\text{CAM}(\text{DM}, X))$

2. $G[h] = \left\{ g \mid g \in \left(\frac{h-1}{M}, \frac{h}{M} \right] \right\}, h = 1, 2, \dots, M$

3. $X^{t=0} = G_h + \delta_h^t, \frac{h}{N} \geq w$, for $i=1$ to N do

4. $X_{i+1}^0 \leftarrow$ 随机初始化

5. end for

6. for $t=1$ to T do

7. $X_i^t = X + \delta_i^t$

8. for $i=1$ to G do

9. if $y_1 \neq y_0$ then

10. $\phi(X_i^t) \leftarrow p(y_1 | X_i^t) - p(y_0 | X_i^t) - \frac{\alpha}{\max D(\delta^0)} D(\delta_i^t)$

11. else

12. $\phi(X_i^t) \leftarrow p(y_2 | X_i^t) - p(y_0 | X_i^t)$

13. end if

14. end for

15. for $i=1$ to G do

16. $P(X_i^t) \leftarrow \frac{\phi(X_i^t)}{\sum_{i=1}^n \phi(X_i^t)}$

17. $Q(X_i^t) \leftarrow \sum_{j=1}^i P(X_j^t)$

18. end for

19. if $\max \{ \phi(X_i^t) \} \geq X'$ then

20. Break

21. end if

22. for $n=1$ to $G/2$ do

23. if $\phi(X) \geq \phi(X)_{\text{avg}}$ then

24. $P_c \leftarrow \frac{k_1 (\phi(X)_{\text{max}} - \phi(X))}{\phi(X)_{\text{max}} - \phi(X)_{\text{avg}}}$

25. else

26. end if

27. $P_c \leftarrow k_2$

28. if $\phi(X)' \geq \phi(X)_{\text{avg}}$ then

29. $P_m \leftarrow \frac{k_3 (\phi(X)_{\text{max}} - \phi(X)')}{\phi(X)_{\text{max}} - \phi(X)_{\text{avg}}}$

30. else $P_m \leftarrow k_4$

31. end if

32. $\delta_i^t \leftarrow \text{Selection}(\delta^t)$

33. $\delta_j^t \leftarrow \text{Selection}(\delta^t)$

34. $\delta_i^{t+1}, \delta_j^{t+1} \leftarrow \text{Crossover}(P_c, \delta_i^t, \delta_j^t)$

35. $\delta_i^{t+1} \leftarrow \text{Mutation}(P_m, \delta_i^{t+1})$

```

36.if  $X_i^{t+1} < X_{\max}^t$  then
37. $X_{\min}^{t+1} = X_{\max}^t$ 
38.end if
39.end for
40.end for
41.返回适应度函数值最高的 X

```

3 实验结果与分析

本节分别对 AEGM-IGA 方法的性能和参数进行实验验证与分析。

3.1 实验设置

本文选用 MNIST、CIFAR-10 和 ImageNet 3 种公开图像数据集,其中:MNIST 为手写数字识别数据集,包含 70 000 张 28×28 像素大小的灰度图像,共 10 个类别;CIFAR-10 为小型彩色图像数据集,包含 60 000 张 32×32 像素大小的彩色图像,共 10 个类别;ImageNet 为大型图像数据集,包含约 1 500 万张图片,共 20 000 个类别。从 3 种数据集中各随机选取 1 000 张能够被所选模型正确识别的图片。

为证明本文所提方法的有效性,对于 MNIST 与 CIFAR10 两个数据集,本文使用 C&W^[10]方法中攻击的目标模型,目标模型分别记为 MNIST(M) 与 CIFAR10(M)。对于 ImageNet 数据集,本文使用 Keras 中预训练好的模型 VGG19、Resnet50、Inception-v3 (Inc-v3)、Inception-v4 (Inc-v4) 与 Inception-Resnet-v2 (IncRes-v2)^[26-28] 作为攻击模型。

本文使用 AEGM-IGA 方法对模型进行攻击,以攻击模型的成功率、扰动大小(以初始成功攻击计算)及模型访问次数作为对抗样本性能的评价标准,并与几种典型攻击方法进行比较,以验证 AEGM-IGA 方法的有效性。将 AEGM-IGA 方法与 FGSM、MI-FGSM、C&W3 种白盒攻击方法,以及 ZOO、AutoZOOM、Boundary 和 POBA-GA 4 种黑盒

攻击方法进行比较,以上方法均是具有代表性的经典对抗攻击方法。

为保证实验结果的客观性,FGSM 与 MI-FGSM 方法中涉及的参数分别依照文献[8]和文献[9]进行设置。AEGM-IGA 方法中的参数依照文献[10]与文献[21]中的设置,即在 MNIST 与 CIFAR10 数据集实验中将迭代轮次设置为 100,每轮生成 20 个样本,方差取[5, 10, 15, 20, 25],噪声点个数取[50, 100, 150, 200, 250]。在 ImageNet 数据集中,将迭代轮次设置为 400、每轮生成 50 个样本、初始扰动的方差为[5, 10, 15, 20, 25],噪声点个数取[5 000, 7 500, 10 000, 12 500, 15 000],平衡系数 $\alpha=3$ 。令本文方法中的热力图分组 $M=100$,控制参数 $w=0.6$ 。

实验部分对抗样本性能的评价标准包括攻击成功率(Attack Success Rate, ASR)、扰动大小以及模型访问次数。其中扰动的衡量以 L_2 范数量化对抗样本与原样本之间的差异;模型访问次数指攻击方法达到终止条件前,目标模型进行预测的次数,该标准只针对需要通过模型访问的对抗攻击方法。

3.2 参数研究

本节通过实验对影响 AEGM-IGA 方法性能的两个参数,即热力值分组大小 M 与控制参数 w 进行讨论,分别以攻击成功率、扰动大小以及模型访问次数 3 个指标作为评价标准,其中扰动大小与模型访问次数以初始攻击成功时计算。数据集选择 ImageNet(MNIST 和 CIFAR10 数据集中图片较小,计算出的攻击成功率较高、模型访问次数较少,但扰动较大,不具有普遍性)。热力值分组实验中 $M \in (40, 140)$;控制参数实验中 $w \in (0, 1)$,当 w 取 0 时表示对整张图片添加扰动。

热力值分组 M 对 AEGM-IGA 方法性能的影响结果如图 6 所示。

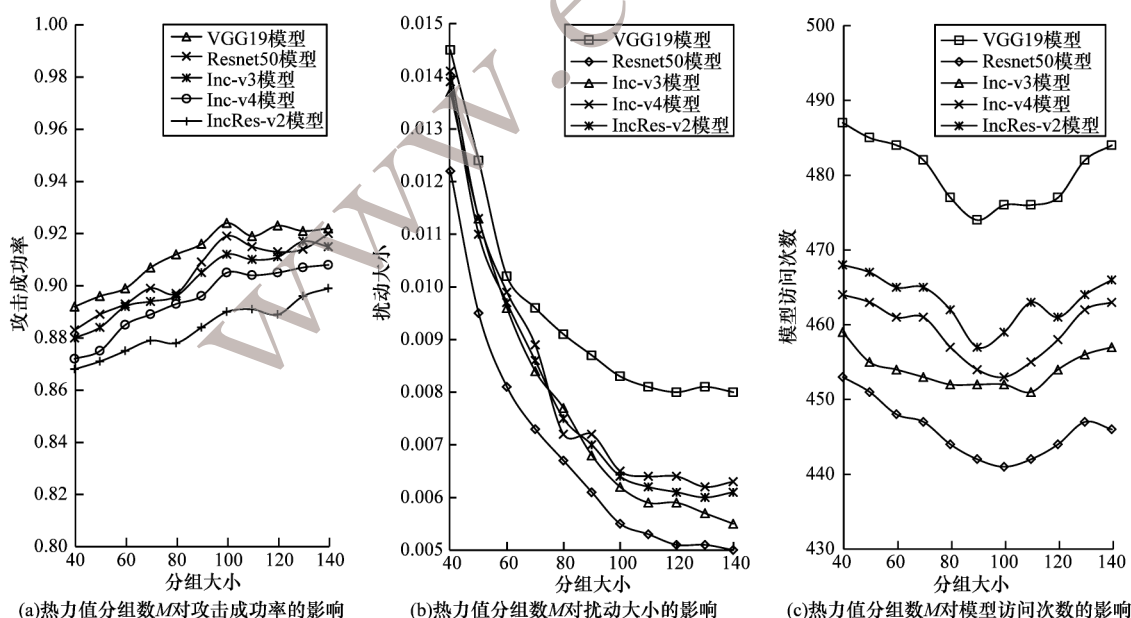


图6 热力值分组数 M 对 AEGM-IGA 方法性能的影响

Fig.6 Influence of thermal value grouping number M on the performance of AEGM-IGA method

由图6(a)可以发现,在相同控制参数下的攻击成功率实验中,随着分组数的增加,对抗样本的攻击成功率有小幅升高,但效果并不明显。当分组数由40增加至110时,攻击成功率小幅上升;当分组数由110再次增加时,对抗样本的攻击成功率几乎不再发生变化。由图6(b)可知,随着分组数的增加,添加至样本中的扰动明显减少,在分组数由40增加至100的过程中,添加至样本中的扰动明显减少,且下降趋势较为快速;当分组数大于100时,样本中扰动大小变化较为缓慢直至趋于稳定。由图6(c)可知,随着分组数的增加或减少,算法的模型访问次数并无明显变化,在分组数由40增加至90过程中,随着分组数的增加,模型访问次数小幅减少,但效果并不明显;当分组数由90继续增加时,本文方法的模型访

问次数小幅增加,但同样并不明显。本文猜测,实验结果呈现出这一变化趋势的原因是随着分组数量的增加,模型关键区域被划分得更精细,进而使每组中添加的冗余扰动更少且更容易产生攻击性。通过实验可以发现,参数 M 可以依据图像中不同位置元素的重要程度对像素分组进行筛选,进而选取其中对神经网络更为重要的像素区域,在该区域进行扰动进化,能在更快速地诱导神经网络出错的同时,有效减少对图像中所添加的扰动,从而提升对抗样本的隐蔽性。结合参数 M 对所生成对抗样本性能的综合考虑,本文取分组数 $M=100$,因为当分组为100时,对抗样本的攻击成功率虽然没有达到最高,但扰动与模型访问次数都接近最低。

控制参数 w 对对抗样本性能的影响如图7所示。

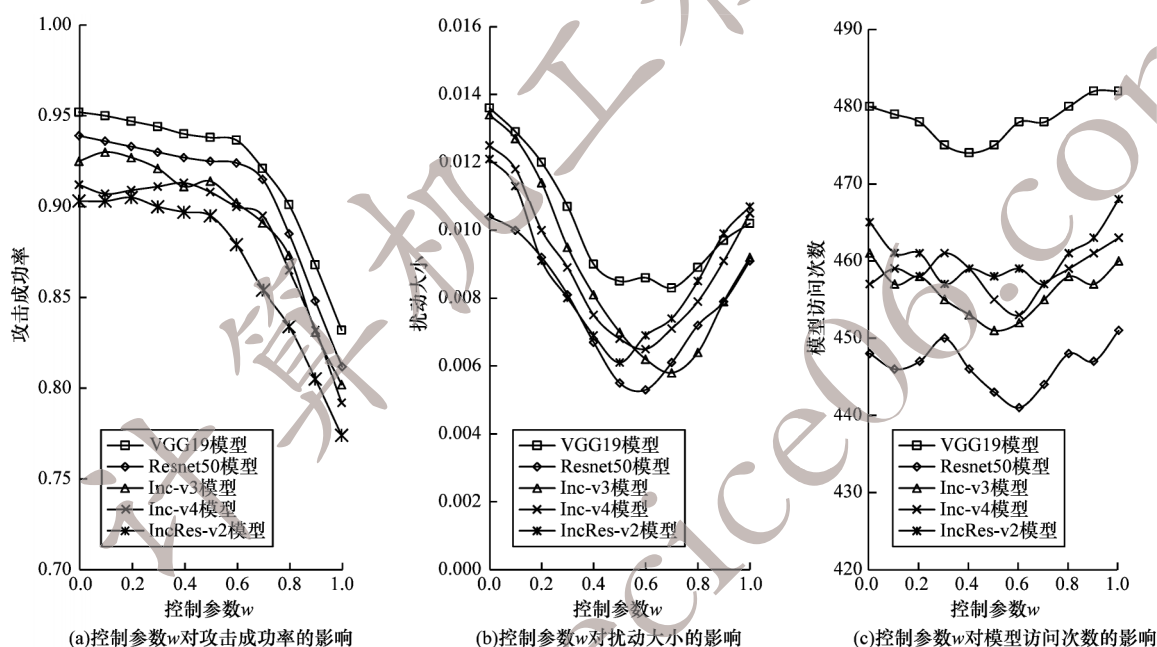


图7 控制参数 w 对AEGM-IGA方法性能的影响

Fig.7 Influence of control parameter w on the performance of AEGM-IGA method

通过图7可以发现,在攻击成功率实验中,在控制参数由0增大到0.6的过程中,随着控制参数的增大,攻击成功率小幅下降,但下降趋势较为缓慢;在控制参数由0.6增大到1的过程中,随着控制参数的增大,攻击成功率明显下降,且下降趋势较快速。在扰动实验中,在控制参数由0增大到0.5的过程中,随着控制参数的增大,添加至样本中的扰动明显减少,且下降趋势较为快速;在控制参数由0.5增大到1的过程中,随着控制参数的增大,添加至样本中的扰动又逐渐增多,且增多速度较为快速。在模型访问次数实验中,控制参数的增大或减少只能引起模型访问次数的小幅波动,对实验结果影响并不明显。本文猜测实验结果呈现出这一变化趋势的原因是控制参数决定进化区域的大小(即添加扰动区域的大小),前期,随着控制参数逐渐变大,添加扰动区域更加集中在图像的关键区域,因此攻

击成功率几乎不变,但所添加的扰动逐渐减少;后期,随着控制参数的继续变大,图像中的一些关键区域也被抛弃,导致攻击成功率下降,且为了能够完成攻击,需要增加扰动的幅度。结合实验结果可以发现, w 值越大,所筛选出的进化区域就越小,由此所生成的对抗样本的攻击成功率也相对下降,但其隐蔽性会得到一定程度的提升,选择一个合适的 w 能够在保证较高攻击成功率的同时,有效提升所生成对抗样本的隐蔽性。综合考虑对抗样本的性能,本文取控制参数 $w=0.6$,因为当控制参数为0.6时,本文方法选取的进化区域与图像关键区域重合度较高,能够在保证对抗样本拥有较高攻击成功率的同时,有效控制样本中的扰动大小。

3.3 攻击性能对比

为验证AEGM-IGA方法的有效性,将本文方法与经典的白盒与黑盒攻击方法进行对比。本文方法

虽然属于黑盒攻击方法,但其攻击成功率与隐蔽性甚至超过一些白盒攻击,能够在保证拥有较高攻击成功率的同时,有效增强样本的隐蔽性,且可以通过

少量的模型访问,完成初步攻击。

3.3.1 攻击成功率的对比

不同方法的攻击成功率如表 2 所示。

表 2 不同方法的攻击成功率对比

Table 2 Comparison of attacks success rates of different methods

方法	MNIST(M)模型	CIFAR-10(M)模型	VGG19 模型	Resnet50 模型	Inc-v3 模型	Inc-v4 模型	IncRes-v2 模型
FGSM 方法	86.3	88.4	77.2	79.8	82.5	32.7	32.6
MI-FGSM 方法	99.7	99.7	99.2	99.3	97.3	95.6	94.2
C&W 方法	100.0	100.0	99.3	100.0	99.5	99.3	99.1
ZOO 方法	100.0	99.3	87.2	88.5	87.4	85.9	81.3
AutoZOOM 方法	100.0	100.0	99.9	100.0	100.0	100.0	100.0
Boundary 方法	76.4	73.2	71.8	70.6	70.1	67.4	61.5
POBA-GA 方法	100.0	100.0	95.7	98.2	94.4	92.9	90.3
AEGM-IGA 方法	100.0	100.0	94.4	97.1	93.5	90.7	89.5

由表 2 可知,由 AEGM-IGA 方法生成的对抗样本拥有相当高的黑盒攻击成功率,其表现甚至优于一些白盒攻击方法。在表 2 中,黑盒攻击中表现最好的是 AutoZOOM 方法,随后是同样基于遗传算法的 POBA-GA 方法,本文所提出的 AEGM-IGA 方法排名第 3。在相同黑盒攻击下,与同是遗传算法的 POBA-GA 方法相比,AEGM-IGA 方法的攻击成功率略低。在攻击成功率下降最多的 Inc-v4 模型上,AEGM-IGA 方法的黑盒攻击成功率仍然达到 90.7%,仅比 POBA-GA 方法下降 2.2 个百分点,仍然在可接受范围内。出现这一结果的主要原因是 AEGM-IGA 方法为了更好地控制对抗样本的隐蔽性,将控制参数 w 选得略大,导致算法

在选择进化区域时,将一小部分关键区域排除在外,最终影响对抗样本的攻击成功率。攻击者可以根据自身的实际需求,通过牺牲样本一定的隐蔽性来增大攻击成功率。正常来说,掌握模型全部信息的白盒攻击能达到更高的攻击成功率,但从实验结果来看,使用 AEGM-IGA 方法生成的对抗样本的黑盒攻击成功率,与 MI-FGSM 方法生成的对抗样本的白盒攻击成功率相差无几,甚至远超 FGSM 方法。

3.3.2 扰动大小的对比

本文使用扰动大小衡量对抗样本的隐蔽性,扰动越小,说明该方法的隐蔽性越高。不同方法生成的对抗样本中扰动大小的结果如表 3 所示。

表 3 不同方法生成的对抗样本扰动大小对比

Table 3 Comparison of disturbance size of adversarial example generated by different methods

方法	MNIST(M)模型	CIFAR-10(M)模型	VGG19 模型	Resnet50 模型	Inc-v3 模型	Inc-v4 模型	IncRes-v2 模型
FGSM 方法	0.821 3	0.063 3	0.009 6	0.011 4	0.006 7	0.007 3	0.006 4
MI-FGSM 方法	0.449 0	0.037 9	0.007 5	0.005 8	0.004 2	0.003 9	0.003 6
C&W 方法	0.157 3	0.017 5	0.005 3	0.002 1	0.002 8	0.002 5	0.002 4
ZOO 方法	0.214 0	0.098 0	0.026 9	0.035 4	0.019 5	0.021 7	0.020 4
AutoZOOM 方法	0.323 6	0.137 3	0.041 9	0.034 9	0.036 3	0.037 2	0.035 8
Boundary 方法	0.204 1	0.016 1	0.003 2	0.002 0	0.003 8	0.004 1	0.004 0
POBA-GA 方法	0.197 3	0.045 8	0.010 1	0.009 4	0.011 4	0.010 7	0.011 7
AEGM-IGA 方法	0.178 2	0.037 5	0.008 3	0.005 3	0.005 8	0.006 5	0.006 1

由表 3 可知,在相同攻击成功率的情况下,本文方法 AEGM-IGA 生成的对抗样本中扰动较小。黑盒攻击中隐蔽性最好的是 Boundary 方法,排名第 2 的是本文方法 AEGM-IGA,其隐蔽性远超 AutoZOOM 与 ZOO 方法。与同样基于遗传算法的 POBA-GA 方法相比,本文方法能够更好地控制扰动大小,对抗样本的隐蔽性更好。如在 ImageNet 数据集上,攻击 Inc-v3、Inc-v4 与 IncRes-v2 模型生成的对抗样本中扰动大小平均降低了 47%,提升效果明显。

正常来说,白盒攻击能够掌握模型全部信息,生成的对抗样本能够更好地拟合目标模型,添加的扰动也更小,但 AEGM-IGA 方法对扰动大小的控制,与白盒攻击较为接近,甚至要优于 FGSM 方法。主要是因为本文方法在种群初始化时引入 Grad-CAM 解释方法,能够很好地将扰动进化范围控制在模型重点关注区域,减少了图像中的冗余扰动,在相同攻击效果的情况下,本文方法生成的对抗样本扰动更小,隐蔽性更强。

3.3.3 模型访问次数的对比

由于白盒攻击主要基于模型梯度完成,进行模型访问次数的对比并无实际意义,因此模型访问次数的对比

实验对象仅限于黑盒攻击,黑盒攻击模型访问次数实验的结果如表4所示,其中括号中的数据为模型询问次数,该值表示该方法初始攻击成功的模型访问次数。

表4 对抗攻击模型访问次数对比

Table 4 Comparison of the number of visits against the attack model

方法	MINIST(M) 模型	CIFAR-10(M) 模型	VGG19 模型	Resnet50 模型	Inc-v3 模型	Inc-v4 模型	IncRes-v2 模型
ZOO方法	9 317	4457	235 437	230 384	265 357	258 312	263 716
AutoZOOM方法	447(100)	309(86)	472(1 687)	4 268(1 694)	4 107(1 713)	4 214(1 784)	4 135(1 716)
Boundary方法	12 500	12 500	125 000	125 000	125 000	125 000	125 000
POBA-GA方法	427(95)	384(79)	3 786(536)	3 707(495)	3 684(473)	3 712(482)	3 751(485)
AEGM-IGA方法	374(87)	291(64)	3 154(474)	3 062(441)	3 173(451)	3 246(453)	3 257(457)

由表4可知,本文方法AEGM-IGA生成对抗样本的效率很高。在相同黑盒攻击情况下,本文方法的模型访问次数在所有对比方法中最少,相较攻击效率最低的ZOO方法(ZOO方法由于需要反复对模型进行访问来估计模型梯度),本文方法的模型访问次数不足ZOO的1.5%。与AutoZOOM方法(在ZOO基础上增加了自动编码器及自适应梯度估计策略,有效平衡了失真与访问次数)相比,本文方法具有更高的攻击效率,在ImageNet数据集上,本文方法较AutoZOOM方法的模型访问次数平均下降了25.4%。最后,与同是遗传算法的POBA-GA方法相比,本文方法的模型访问次数更少,尤其是在ImageNet数据集上,模型访问次数平均下降17%。本文方法AEGM-IGA能大幅降低模型访问次数的原因主要有3个:

- 1)在初始化时,通过在更容易诱导模型出错的区域添加扰动,能提高产生的样本质量。
- 2)前期引入较高的交叉、变异概率,加快了算法的寻优速度。
- 3)采用精英保留策略,能够避免优秀个体在进化过程中消失。

3.4 结果分析

本文分别以攻击成功率、样本的隐蔽性(扰动大小)以及攻击效率(模型访问次数)作为标准,对不同方法生成的对抗样本进行对比。由实验结果可知,使用本文方法AEGM-IGA生成的对抗样本,与同样基于遗传算法的POBA-GA方法相比,虽然在攻击成功率上略有降低,但能够有效增加所生成对抗样本的隐蔽性,且攻击效率大幅提升。例如在ImageNet数据集上,POBA-GA方法需要约500次的模型访问次数才能将攻击成功率提高至95%,但AEGM-IGA方法仅需要约450次模型访问次数即可将攻击成功率提高到95%,且此时AEGM-IGA方法中的扰动大小仅为POBA-GA方法中的1/2左右。在现实世界中,白盒攻击并没有得到广泛应用,且攻击者大多无法掌握目标模型的网络结构及相关参数,本文提出的AEGM-IGA方法更贴近实际,在仅需知道模型输

出标签及其置信度的情况下,通过较少次数的模型访问,即可生成具有较高攻击性且隐蔽性好的对抗样本。

4 结束语

现有黑盒攻击方法普遍存在攻击效率低、隐蔽性差等问题,本文提出一种基于改进遗传算法的对抗攻击方法,能够在仅知道模型输出标签及其置信度的情况下,通过少量模型访问,实现黑盒攻击。考虑到图像中的关键区域对分类结果影响更大,本文使用Grad-CAM解释方法,将扰动进化限制在图像关键区域,以减少冗余扰动,并在算法中引入自适应函数,加速算法收敛,提高攻击效率。实验结果表明,与ZOO、AutoZOOM、Boundary等方法相比,在保证攻击成功率近乎相同的情况下,本文方法拥有更高的攻击效率与更好的隐蔽性。下一步将通过Grad-CAM解释法,充分利用对抗样本,提升模型可解释性。

参考文献

[1] SZEGEDY C, LIU W, JIA Y Q, et al. Going deeper with convolutions [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2015: 1-9.

[2] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[3] REDMON J, FARHADI A. YOLOv3: an incremental improvement[EB/OL]. [2022-07-22]. <https://arxiv.org/abs/1804.02767>.

[4] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. [2022-07-22]. <https://scholar.cnki.net/zn/Detail/index/GARJ2014/DBLP6D88B9E289B257B44613EA4BE1000161>.

[5] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems. New York, USA: ACM Press, 2014: 3104-3112.

- [6] XIONG W, DROPPA J, HUANG X, et al. Achieving human parity in conversational speech recognition [EB/OL]. [2022-07-22]. <https://arxiv.org/abs/1610.05256>.
- [7] ZHANG Z X, GEIGER J, POHJALAINEN J, et al. Deep learning for environmentally robust speech recognition: an overview of recent developments[J]. *ACM Transactions on Intelligent Systems and Technology*, 2018, 9(5): 1-28.
- [8] 马玉琨,毋立芳,简萌,等. 一种面向人脸活体检测的对抗样本生成算法[J]. *软件学报*, 2019, 30(2): 469-480. MA Y K, WU L F, JIAN M, et al. Algorithm to generate adversarial examples for face-spoofing detection [J]. *Journal of Software*, 2019, 30(2): 469-480. (in Chinese)
- [9] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks [EB/OL]. [2022-07-22]. <https://arxiv.org/abs/1706.06083>.
- [10] GUO C, RANA M, CISSE M, et al. Countering adversarial images using input transformations[EB/OL]. [2022-07-22]. <https://arxiv.org/abs/1711.00117>.
- [11] SAMANGOU EI P, KABKAB M, CHELLAPPA R. Defense-GAN: protecting classifiers against adversarial attacks using generative models[EB/OL]. [2022-07-22]. <https://arxiv.org/abs/1805.06605>.
- [12] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[EB/OL]. [2022-07-22]. <https://arxiv.org/abs/1312.6199>.
- [13] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[EB/OL]. [2022-07-22]. <https://arxiv.org/abs/1412.6572>.
- [14] DONG Y P, LIAO F Z, PANG T Y, et al. Boosting adversarial attacks with momentum[C]//*Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Washington D. C. , USA: IEEE Press, 2018: 9185-9193.
- [15] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//*Proceedings of IEEE Symposium on Security and Privacy*. Washington D. C. , USA: IEEE Press, 2017: 39-57.
- [16] BRENDL W, RAUBER J, BETHGE M. Decision-based adversarial attacks: reliable attacks against black-box machine learning models[EB/OL]. [2022-07-22]. <https://arxiv.org/abs/1712.04248>.
- [17] CHEN P Y, ZHANG H, SHARMA Y, et al. ZOO: zeroth order optimization based black-box attacks to deep neural networks without training substitute models [C]//*Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*. New York, USA: ACM Press, 2017: 15-26.
- [18] TU C C, TING P S, CHEN P Y, et al. AutoZOOM: autoencoder-based zeroth order optimization method for attacking black-box neural networks[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, 33(1): 742-749.
- [19] DONG Y P, SU H, WU B Y, et al. Efficient decision-based black-box adversarial attacks on face recognition [C]//*Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Washington D. C. , USA: IEEE Press, 2020: 7706-7714.
- [20] XIE C H, ZHANG Z S, ZHOU Y Y, et al. Improving transferability of adversarial examples with input diversity [C]//*Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Washington D. C. , USA: IEEE Press, 2020: 2725-2734.
- [21] CHEN J Y, SU M M, SHEN S J, et al. POBA-GA: perturbation optimized black-box adversarial attacks via genetic algorithm[J]. *Computers & Security*, 2019, 85: 89-106.
- [22] KOZA J R. Automatic discovery of reusable programs[M]. Cambridge, USA: MIT Press, 1994.
- [23] ZHOU B L, KHOSLA A, LAPEDRIZA A, et al. Learning deep features for discriminative localization [C]//*Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. Washington D. C. , USA: IEEE Press, 2016: 2921-2929.
- [24] SELVARAJU R R, COGSWELL M, DAS A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization[C]//*Proceedings of IEEE International Conference on Computer Vision*. Washington D. C. , USA: IEEE Press, 2017: 618-626.
- [25] SRINIVAS M, PATNAIK L M. Adaptive probabilities of crossover and mutation in genetic algorithms [J]. *IEEE Transactions on Systems, Man, and Cybernetics*, 1994, 24(4): 656-667.
- [26] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. [2022-07-22]. <https://arxiv.org/abs/1409.1556>.
- [27] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//*Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. Washington D. C. , USA: IEEE Press, 2016: 770-778.
- [28] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision [C]//*Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. Washington D. C. , USA: IEEE Press, 2016: 2818-2826.