https://doi.org/10.1051/jnwpu/20234161170

DE-JSMA:面向 SAR-ATR 模型的稀疏对抗攻击算法

金夏颖[1], 李扬[2], 潘泉[2]

(1.西北工业大学 网络空间安全学院, 陕西 西安 710072; 2.西北工业大学 自动化学院, 陕西 西安 710072)

摘 要:DNN 易受攻击的特点使得以智能算法为识别手段的 SAR-ATR 系统也存在一定脆弱性。为验证其脆弱性,结合 SAR 图像特征稀疏的特点,在显著图对抗攻击算法和差分进化算法基础上提出了 DE-JSMA 稀疏攻击算法,精确筛选出对模型推理结果影响较大的显著特征后,为显著特征优化出合适的特征值。为了更全面地验证攻击的有效性,构建了一种结合攻击成功率和对抗样本平均置信度的新指标 F_c 值。实验结果表明,在没有增加过多耗时,且保证高攻击成功率情况下,DE-JSMA 将只能定向攻击的 JSMA 扩展到了非定向攻击场景,且在 2 种攻击场景下均实现了可靠性更高、稀疏性更优的稀疏对抗攻击,仅扰动 0.31%与 0.85%的像素即可达到 100%与 78.79%以上的非定向与定向攻击成功率。

关 键 词:合成孔径雷达;自动目标识别;深度学习;对抗攻击;稀疏攻击

中图分类号:TP391.4 **文献标志码**:A **文章编号**:1000-2758(2023)06-1170-09

合成孔径雷达(synthetic aperture radar,SAR)广泛应用于地理勘测、军事信息处理等领域。由于 SAR 图像人工解译十分困难,且 SAR 系统获取的数据量不断增加,近年来 SAR 图像解译已转向智能化,基于深度神经网络模型(deep neural network,DNN)的合成孔径雷达自动目标识别系统(synthetic aperture radar automatic target recognition,SAR-ATR)可自动提取特征,大幅提高了识别准确率和效率[1-3]。但研究表明[4-9] DNN 模型易受到对抗样本的攻击,且对抗样本能够以高置信度被 DNN 模型误分类,但人眼几乎无法辨别。

2014 年,Szegedy 等[10]首次发现 DNN 模型的脆弱性。而后,Goodfellow 等[4]解释了这种“对抗攻击”现象存在的原因,即 DNN 模型高维空间中的线性性质导致无限小的对抗扰动在 DNN 模型前向传播过程中不断积累,最终引起输出的巨大变化。根据对抗扰动的疏密程度,可将对抗攻击分为密集攻击和稀疏攻击。密集攻击通常以全尺寸分布的方式向输入图像添加扰动,例如:Goodfellow 等[4]提出了利用损失函数梯度的快速梯度符号方法(fast gradient sign method,FGSM);Moosavi-Dezfooli 等[5]提出了基于超平面分类的攻击方法 DeepFool;Carlini 和 Wagner[6]提出了 C&W 算法,将距离度量引入到目标函数中,迭代寻找其最优解。与密集攻击不同,稀疏攻击主要需要考虑 2 个问题,即扰动位置和扰动强度的选择,例如:Papernot 等[7]提出了基于雅各比矩阵的显著图攻击算法(Jacobian saliency map attack,JSMA),构建显著图挑选显著特征,并对其添加扰动;Su 等[8]提出了 OnePixel 算法,结合差分进化(differential evolution,DE)[11]算法确定扰动像素坐标和扰动强度,利用 DNN 模型的结果引导进化方向;Modas 等[9]提出了 SparseFool 算法,是 DeepFool 算法在稀疏攻击上的应用,但只能进行非定向攻击。

上述攻击算法均针对光学图像,而近年来,针对 SAR-ATR 系统的对抗攻击研究也开始涌现。2020

收稿日期:2022-12-27 **基金项目**:国家自然科学基金(62103330,62233014)资助

作者简介:金夏颖(2000—),西北工业大学硕士研究生,主要从事面向 SAR 图像的对抗攻击研究。

通信作者:李扬(1990—),西北工业大学副教授,主要从事人工智能安全研究。e-mail:liyangnpu@nwpu.edu.cn

年,Huang 等[12]使用 3 种主流对抗攻击算法攻击基于 DNN 的 SAR-ATR 模型,识别准确率降低 90%以上,首次证明了 SAR-ATR 模型的脆弱性。2021 年,Du 等[13]提出了 Fast C&W 算法,引入编码器网络实现了原始样本到对抗样本的映射,提高了算法实时性。最近,Peng 等[14]提出了斑点变异攻击(speckle-variant attack,SVA)算法,通过操纵斑点噪声模式增强对抗样本的迁移性。然而现有针对 SAR 图像的对抗攻击算法均为密集型,扰动范围大,物理可行性差。而且 SAR 图像由离散散射中心亮斑组成,目标区域占比较小,导致其特征冗余程度高[15],因而采用稀疏攻击添加针对性扰动的方式可解释性更强,稀疏对抗样本也更贴合物理世界中监测区域被部分遮挡或附着异物的情况。

为了解决上述问题,针对 SAR 图像特征稀疏性,本文在 JSMA 算法的基础上提出了一种新的稀疏攻击方法——基于差分进化的雅各比显著图攻击算法(differential evolution-Jacobian saliency map attack,DE-JSMA),在精确筛选出对模型推理结果影响较大的显著特征的同时,动态选择合适的特征值,并延伸设计,使其既可以定向攻击,又可以非定向攻击。为了更全面地验证攻击有效性,本文构建了一种结合攻击成功率和对抗样本平均置信度的新指标 F_c 值。实验结果表明,在没有增加过多耗时,且保证高攻击成功率的情况下,DE-JSMA 实现了可靠性更高、稀疏性更优的稀疏攻击,并且表现出了优越的重定向能力。

1　SAR-ATR 系统攻击分析

SAR-ATR 系统通常包括图像采集、数据预处理、深度神经网络训练和验证、深度识别模型测试和部署等几个重要的信息处理步骤,这些步骤都具有潜在的安全问题。①图像采集阶段:攻击者使用特定方式干扰 SAR 传感器的数据采集链路。在 SAR 成像的过程中,攻击者可以设置不同的干扰模式来攻击 SAR 抗欺骗干扰技术。②图像预处理阶段:在提取图像信息前,需对原始图像数据进行一系列校正和预处理操作,在这个过程中也会存在新的安全风险,例如重采样攻击,攻击图片被图像识别模型缩放到特定的空间分辨率后才会得以显现。③模型训练阶段:典型的攻击方式包括投毒攻击和后门攻击。投毒攻击是指攻击者设法对训练数据进行某种恶意操作进而污染训练数据,使原有数据的概率分布发生改变,模型可用性被破坏的一种攻击方式。而后门攻击是指模型在训练阶段使用某些精心制作的恶意样本进行训练,此模型对于正常样本的预测结果表现正常,但对于某些具有特定属性的测试样本,便可以触发后门使模型的预测结果受攻击者操控。④模型测试阶段:典型的攻击方式是规避攻击,规避攻击的一个代表性工作就是针对 DNN 模型的对抗攻击。⑤模型部署阶段:典型的攻击方式包括面向设备的硬件攻击、操作系统攻击等。许多支持神经网络加速的开源软硬件平台的底层安全问题尚未得到充分验证,这给模型部署带来了很大的安全风险。

本文的工作是针对 SAR-ATR 系统的模型测试阶段实施对抗攻击,在图像分类任务中,对抗攻击的目的是针对输入图像(即原始样本) $X \in \boldsymbol{R}^{H\times W}$ 及其真实标签 $Y \in \boldsymbol{R}^{J}$,找到一个人眼难以察觉的对抗扰动 δ,从而误导目标识别模型 $F:\boldsymbol{R}^{H\times W}\rightarrow\boldsymbol{R}^{J}$ 输出错误预测结果。衡量对抗扰动 δ 的大小通常使用 L_p 范数。对抗攻击的形式化表达如(1)式所示

$$\begin{aligned}&\min \|\delta\|_p\\&\text{s.t. } F(X)=Y\\&\quad F(X')=Y'\\&\quad Y'\neq Y\\&\quad X'\neq X+\delta\in D\end{aligned}\tag{1}$$

式中:F 为深度神经网络模型;X 为原始样本;X' 为对抗样本;Y 为原始样本的标签;Y' 为对抗样本的标签;δ 为对抗扰动;D 为样本取值范围,图像中通常为像素值范围[0,255];$\|\cdot\|_p$ 为度量对抗扰动大小的范数,$p=0$ 时,范数值表示相较于原始样本,对抗样本改变的像素个数大小;$p=2$ 时,范数值表示对抗样本和原始样本之间的欧几里得距离。

由(1)式可知,对抗攻击可看作一个在规定条件内最小化的优化问题,但 DNN 模型高度非线性的性质导致此优化问题非线性且非凸,因此找到这个问题的最优解并非易事。

2　DE-JSMA 算法

本文将(1)式中的求解转化为最佳扰动位置与大小选择的问题,因此,设计的 DE-JSMA 算法主要分为 3 步。第一步是利用雅克比显著性计算的方法实现显著图的计算;第二步是利用算得的显著图确定要攻击的位置;第三步则是利用差分进化算法计

算显著性图所在点的攻击强度,最终得到对抗扰动 δ,将其添加到干净样本 X 上获得对抗样本 X'从而实现有效的稀疏攻击,整体框架图如图 1 所示。在 JSMA 的基础上,本文进行了延伸设计,使得提出的 DE-JSMA 既可以实现非定向攻击,也可以实现定向攻击,具体的算法设计将在下文进行详细介绍。

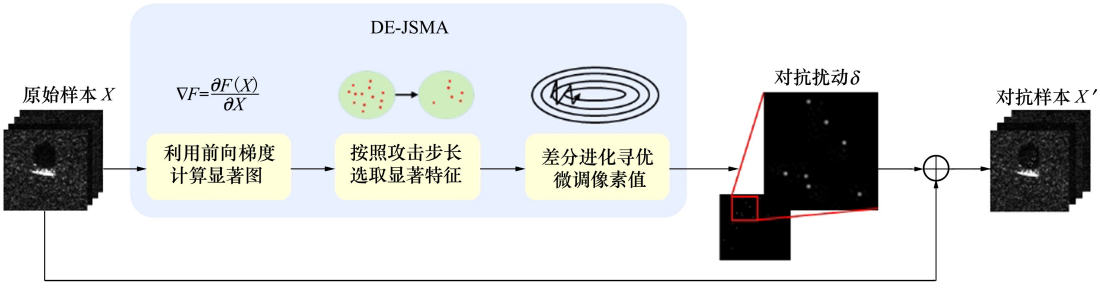

图 1 DE-JSMA 算法框架图

2.1 显著性检测方法

显著性检测方法主要分为以下 2 步:

1) 求前向梯度

前向梯度的定义是目标模型输入与隐藏层输出映射函数的雅各比矩阵,可表示单个隐藏层输出关于单个输入特征的变化关系,其公式为

$$\nabla F(X)=\frac{\partial F(X)}{\partial X}=\left[\frac{\partial F_j(X)}{\partial x_i}\right]_{i=1,2,\cdots,M,j=1,2,\cdots,N} \tag{2}$$

式中:X 为输入图像;$F(\cdot)$ 为目标模型输入到隐藏层输出的映射函数;F_j 为第 j 个隐藏层输出;x_i 为输入图像的第 i 个特征;M 为输入图像的特征总数,即像素总数;N 为目标模型隐藏层的输出个数。

2) 计算显著图

JSMA 算法提出了基于雅各比矩阵的显著图计算方法,此方法只适用于定向攻击,其公式为

$$S(X,t)[i]=\begin{cases}0, & \dfrac{\partial F_t(X)}{\partial x_i}<0 \text{ 或} \sum\limits_{j\neq t}\dfrac{\partial F_j(X)}{\partial x_i}>0\\ \left(\dfrac{\partial F_t(X)}{\partial x_i}\right)\left|\sum\limits_{j\neq t}\dfrac{\partial F_j(X)}{\partial x_i}\right|, & \text{其他}\end{cases} \tag{3}$$

式中,t 为目标类别标签。

在定向攻击场景中选择显著特征时,应当选取 $\dfrac{\partial F_t(X)}{\partial x_i}$ 为正且 $\sum\limits_{j\neq t}\dfrac{\partial F_j(X)}{\partial x_i}$ 为负的特征。$\dfrac{\partial F_t(X)}{\partial x_i}$ 为正,则特征值 x_i 增加时,$F_t(X)$ 会随之增加。同理,$\sum\limits_{j\neq t}\dfrac{\partial F_j(X)}{\partial x_i}$ 为负,则特征值 x_i 增加时,$\sum\limits_{j\neq t}F_j(X)$ 会随之减小或者保持不变。而可选特征的显著值 $S(X,t)[i]$ 定义为 $\dfrac{\partial F_t(X)}{\partial x_i}$ 与 $\left|\sum\limits_{j\neq t}\dfrac{\partial F_j(X)}{\partial x_i}\right|$ 的乘积,可综合考虑所有前向梯度分量的影响,特征显著值越高,扰动该特征时更容易增大目标类别的输出结果或减小非目标类别的输出结果。

JSMA 算法只实现了定向攻击,本文将此算法推广到非定向攻击场景,其公式为

$$S(X,l)[i]=\begin{cases}0, & \dfrac{\partial F_l(X)}{\partial x_i}>0 \text{ 或} \sum\limits_{j\neq l}\dfrac{\partial F_j(X)}{\partial x_i}<0\\ \left(\dfrac{\partial F_l(X)}{\partial x_i}\right)\left|\sum\limits_{j\neq l}\dfrac{\partial F_j(X)}{\partial x_i}\right|, & \text{其他}\end{cases} \tag{4}$$

式中,l 为输入图像的真实标签。

在非定向攻击场景中选择显著特征时,应当选取 $\dfrac{\partial F_l(X)}{\partial x_i}$ 为负且 $\sum\limits_{j\neq l}\dfrac{\partial F_j(X)}{\partial x_i}$ 为正的特征。$\dfrac{\partial F_l(X)}{\partial x_i}$ 为负,则特征值 x_i 增加时,$F_l(X)$ 会随之减小。同理,$\sum\limits_{j\neq l}\dfrac{\partial F_j(X)}{\partial x_i}$ 为正,则特征值 x_i 增加时,$\sum\limits_{j\neq l}F_j(X)$ 会随之增大或者保持不变。而可选特征的显著值 $S(X,l)[i]$ 定义为 $\dfrac{\partial F_l(X)}{\partial x_i}$ 与 $\left|\sum\limits_{j\neq l}\dfrac{\partial F_j(X)}{\partial x_i}\right|$ 的乘积,特征显著值越高,扰动该特征时更容易减小标签类别的输出结果或增大其他类别的输出结果。

2.2 差分进化算法

Storn 和 Price[11] 提出差分进化算法,采用基于群体的启发式搜索解决连续空间全局优化问题。DE 算法通过有针对性地选择系统参数优化系统的某些属性,DE-JSMA 中选取的待优化系统参数为挑选出的显著特征特征值。差分进化算法优化问题的

标准流程是,首先设计一个目标函数对问题目标进行建模,该函数可结合任意约束。目标函数大多会将优化问题定义为最小化问题,这种目标函数可被更准确地描述为成本函数。DE-JSMA 中差分进化的目标函数 f_{DE} 设为目标网络模型输出的目标类别置信度,如(5)式所示,以此来确定特征值进化的方向。

$$f_{\mathrm{DE}}=\begin{cases}c(X'(t)), & \text{定向攻击}\\ c(X'(l)), & \text{非定向攻击}\end{cases} \tag{5}$$

式中:$c(X'(t))$ 为对抗样本目标类别 t 的置信度;$c(X'(l))$ 为对抗样本标签类别 l 的置信度。

而后算法初始化种群即待优化系统参数的候选值,循环执行变异、交叉和选择操作,完成种群的进化,直到算法迭代次数达到上限,或种群最优解达到预设误差精度时,算法结束。

算法 1　DE-JSMA

输入　干净样本 X; 标签类别 l; 定向攻击的目标类别 t;
神经网络模型输入到隐藏层输出的映射函数 F;
攻击上限 γ; 攻击步长 n;
DE 算法最大迭代次数 $I_{\max}$

输出　对抗样本 X^*

1) $X^* \leftarrow X$
2) while $N \leqslant \gamma$ and 攻击未成功 do
3)　利用(2) 式计算前向梯度 $\nabla F(X^*)$
4)　if 非定向攻击 then
5)　　利用(4) 式计算显著图
　　$S=\mathrm{saliency\_map}(\nabla F(X^*), l)$;
6)　else if 定向攻击 then
7)　　利用(3) 式计算显著图
　　$S=\mathrm{saliency\_map}(\nabla F(X^*), t)$;
8)　end if
9)　按照攻击步长 n 选取显著特征;
10)　while $j < I_{\max}$ do
11)　　利用 DE 算法微调显著特征的特征值;
12)　　$N \leftarrow N+n$;
13)　end while
14) end while

DE-JSMA 算法执行过程如算法 1 所示。开始攻击前,需设置攻击方式(非定向攻击/定向攻击)、攻击步长 n、DE 算法最大迭代次数 $I_{\max}$ 和攻击上限 γ。若攻击方式设为定向攻击,则还要设置攻击的目标类别。攻击步长是指单次扰动的特征数量。攻击上限指扰动特征数量的上限,γ 最大可设为 16 384(即 128×128)。由于显著图的计算需要较大计算成本,因此为减少计算显著图的次数,每轮需要选取多个显著特征,即攻击步长 n,但 n 也不宜过大,否则会造成后续差分进化算法寻优过程中计算成本增加。n 默认为 5。DE 算法最大迭代次数 $I_{\max}$ 也不宜过大,否则同样会造成耗时过长,而且在单步攻击中迭代次数过多,易使对抗扰动陷入局部最优解。$I_{\max}$ 默认为 3。

设置参数后,DE-JSMA 将迭代攻击直至攻击成功或扰动特征数量超出攻击限制,每轮攻击包含 3 个步骤:首先计算前馈神经网络模型的“前向梯度”,即模型输入到隐藏层输出映射函数的梯度(第 3)步),并利用前向梯度计算显著图(第 4)~8)步);然后基于显著图选取显著特征,特征越显著表示对模型输出影响越大(第 9)步);最后利用差分进化算法寻优,为选取的显著特征确定像素值(第 10)~13)步)。每轮攻击结束判断是否攻击成功,若成功则攻击结束,否则进入下一轮攻击,如此迭代直至攻击成功。

3　实验及分析

为了验证 DE-JSMA 的有效性,本文在公开的 SAR 图像数据集 MSTAR 上分别对经典的 DNN 模型如 AlexNet、VGG16 以及 ResNet18 等进行攻击,并以攻击成功率、F_c 值以及平均扰动特征个数作为评价指标,分别在定向攻击和非定向攻击场景下展开对比实验。

3.1　实验环境与实验设置

硬件环境为英特尔 Silver 4210R 处理器,128 GB内存。软件环境为 64 位 Ubuntu 20.04.1 操作系统,显卡为 GeForce RTX 3090,开发环境为 Python 3.6.13,PyTorch 1.9.0。

实验采用包含 10 类目标的 SAR 图像数据集 MSTAR,攻击的 3 个目标模型在 MSTAR 训练集上的训练准确率分别为 92.62%,97.11%,91.39%,如表 1 所示。本文从 MSTAR 测试集的每个分类分别随机选取 10 张图像,选取的 100 张图像构成验证数据集,用于测试对抗攻击算法,3 个模型攻击数据集上的分类准确率分别为 91.00%,97.00%,91.00%,如表 1 所示。

表 1 3 个目标模型的准确率 %

目标模型	训练准确率	攻击数据集上的准确率
AlexNet	92.62	91.00
VGG16	97.11	97.00
ResNet18	91.39	91.00

本文将与 FGSM、C&W、DeepFool、SparseFool、JSMA、OnePixel 5 个对抗攻击算法进行对比实验,算法特性如表 2 所示。

表 2 不同算法的特性表

对抗攻击算法	攻击模式	扰动疏密度
FGSM[4]	非定向、定向	密集
C&W[6]	非定向、定向	密集
DeepFool[5]	非定向	密集
SparseFool[9]	非定向	稀疏
JSMA[7]	定向	稀疏
OnePixel[8]	非定向、定向	稀疏
DE-JSMA	非定向、定向	稀疏

3.2 评价指标

1) 攻击成功率

攻击成功率 R_{AS}是评价对抗攻击算法最常用的评价指标,按定义可分为定向攻击和非定向攻击。定向攻击的攻击成功率表示在全部样本中,可被模型分类为目标类别的比例,其公式为

$$R_{\mathrm{AS}_{\mathrm{targeted}}}=\frac{\sum_{n=1}^{N} I(Y(X'_n)=t^{(n)})}{N} \tag{6}$$

式中, $I(\cdot)$ 为指示函数;$Y(X'_n)$ 为对抗样本的预测类别;$t^{(n)}$ 为第 n 个对抗样本的目标类别;N 为可被正确分类的样本总数。

非定向攻击的攻击成功率表示在全部样本中,可被模型错误分类的对抗样本的比例,其公式为

$$R_{\mathrm{AS}_{\mathrm{non\text{-}targeted}}}=\frac{\sum_{n=1}^{N} I(Y(X'_n)\neq y^{(n)})}{N} \tag{7}$$

式中, $y^{(n)}$ 为第 n 个干净样本的标签类别。

2) F_c 值

对抗样本的可靠性可通过其对抗类别的置信度来衡量,置信度越高,表示对抗样本具备越多对抗类别的特征,进而表明对抗样本攻击性能就越强。当目标模型发生微调时,置信度高的对抗样本仍能保持攻击效果的概率更高。R_{AS} 是评价对抗攻击算法最重要的指标,可靠性相对次要,因此本文将攻击成功率和与对抗样本置信度相关的参数 C 结合起来,将二者的调和平均数定义为一个新指标 F_c,有助于综合评价对抗样本的质量并提供更准确的判断依据,F_c 值越大,表示对抗样本在保证攻击有效性的前提下更可靠,其公式为

$$F_c=2\cdot\frac{R_{\mathrm{AS}}\cdot C}{R_{\mathrm{AS}}+C}$$

$$C=\begin{cases}\dfrac{\sum_{i=1}^{S} c(X'_i(t))}{S}, & \text{定向攻击}\\ 1-\dfrac{\sum_{i=1}^{S} c(X'_i(l))}{S}, & \text{非定向攻击}\end{cases} \tag{8}$$

式中: $c(X'_i(t))$ 为第 i 个对抗样本目标类别 t 的置信度;S 为对抗样本总数;l 为对抗样本的标签类别。

3) 平均扰动特征个数 N_{avg}

DE-JSMA 算法对扰动进行 L_0 范数约束的稀疏对抗攻击。L_0 范数可表示相较于原始样本,对抗样本改变的像素个数,因此本文将 L_0 范数的平均值 N_{avg} 作为衡量对抗扰动稀疏性和对抗样本隐蔽性的指标,其公式为

$$N_{\mathrm{avg}}=\frac{\sum_{i=1}^{S}\|X'_i-X_i\|_0}{S}=\frac{\sum_{i=1}^{S}\sum_{j=0}^{n} I(x'_{ij}\neq x_{ij})}{S} \tag{9}$$

式中: S 为对抗样本总数;$I(\cdot)$ 为指示函数。

4) 平均耗时 T_{avg}

平均耗时是评价攻击算法的一项重要评价指标,用于衡量攻击算法的实时性,其公式为

$$T_{\mathrm{avg}}=\frac{\sum_{i=1}^{S} T_i}{S} \tag{10}$$

式中, T_i 为第 i 个对抗样本的攻击耗时,单位为 s。

3.3 实验结果与分析

本节实验中,攻击上限 γ 设为 16 384,即可攻击任意多个特征,攻击步长 n 和最大迭代次数 $I_{\max}$ 分别采用默认值 5 和 3。攻击的可视化结果如图 2 所示。

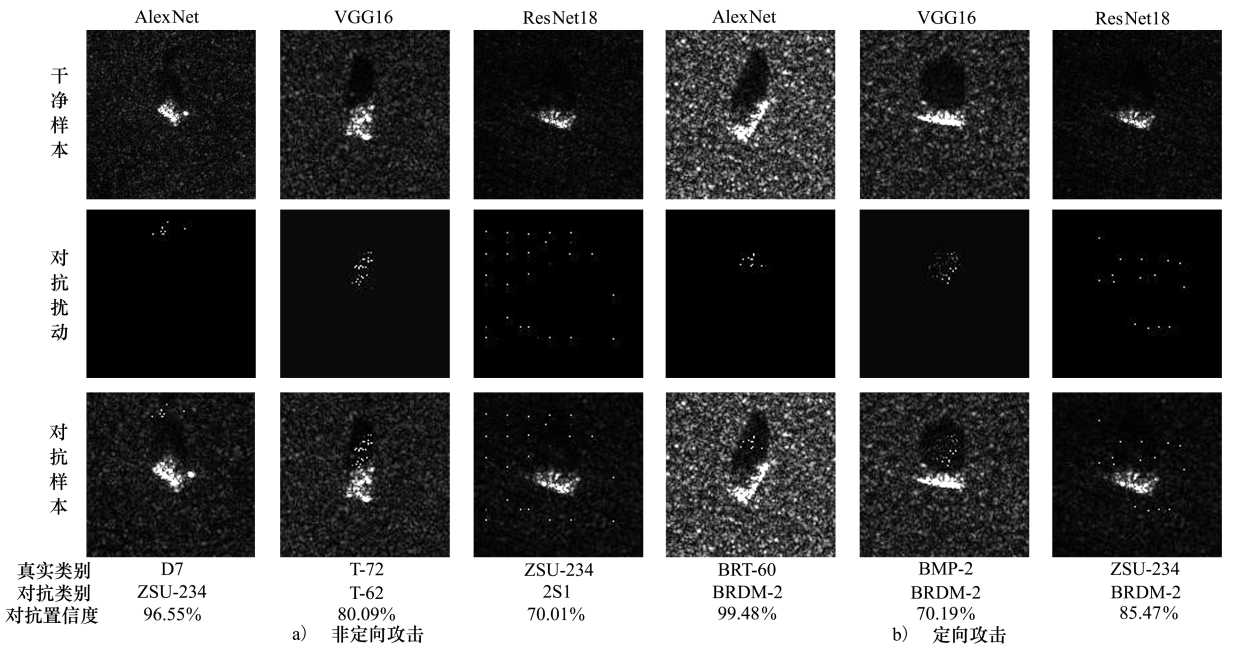

图 2　DE-JSMA 对抗攻击可视化展示

非定向攻击实验对比结果如表 3 所示。DE-JSMA 攻击成功率均达到 100%,攻击效果非常优越。然而 F_c 值结果并不突出,因为 DE-JSMA 非定向攻击时,扰动优化的目标是样本偏离原有决策边界,而非让某一对抗类别置信度增加,这就导致非定向攻击中对抗样本置信度往往较低,进而导致 F_c 值较低,但并不影响 DE-JSMA 的有效性。从 N_{avg} 来看,DE-JSMA 仅扰动 0.09%~0.31%的像素就达到了 100%的攻击成功率,且扰动特征数量比第二稀疏的 OnePixel 少 71.67%~91.62%,攻击稀疏性最好,这正是 DE-JSMA 的优势所在,将扰动集中到极少数的像素上,大大增加了攻击的物理可实现性。

针对定向攻击,本文同样在 MSTAR 数据集上展开,并对其 10 个类别全部进行了定向攻击。为了综合对比攻击效果,实验结果为各项指标全类别结果的均值,如表 4 所示。从攻击成功率 R_{AS} 来分析,DE-JSMA 和 JSMA 有着相似的表现性能,均远超其他算法,但在 F_c 值这一评价指标下,DE-JSMA 有着鲜明的优势,生成的对抗样本在保证有效性的同时表现出最优的可靠性。与非定向攻击不同,DE-JSMA 在挑选出显著像素后,差分进化算法能有效利用目标类别对目标模型进行有针对性的优化,进而优化得到具有高置信度的对抗样本,即可靠的对抗样本。由此可得,DE-JSMA 更适用于定向攻击。除此之外,DE-JSMA 同样表现出最佳的稀疏性,在扰动 0.60%~0.85%的极少数像素的情况下实现定向攻击,大大提升了攻击的实用价值。

表 3　非定向攻击对比结果

目标模型		攻击算法	R_{AS}	F_c	N_{avg}
AlexNet	密集	FGSM	1.000 0	0.949 0	14 992
		C&W	1.000 0	0.658 5	16 031
		DeepFool	1.000 0	0.754 4	14 992
	稀疏	SparseFool	1.000 0	0.927 5	1 525
		OnePixel	0.725 3	0.723 3	179
		DE-JSMA	1.000 0	0.842 0	15
VGG16	密集	FGSM	0.989 7	0.935 2	16 236
		C&W	1.000 0	0.660 0	16 091
		DeepFool	0.948 5	0.753 1	16 236
	稀疏	SparseFool	0.927 8	0.902 2	8 515
		OnePixel	0.762 9	0.753 0	191
		DE-JSMA	1.000 0	0.748 7	43
ResNet18	密集	FGSM	1.000 0	0.985 7	16 250
		C&W	1.000 0	0.669 3	16 088
		DeepFool	0.945 1	0.634 9	16 256
	稀疏	SparseFool	0.967 0	0.906 1	2 066
		OnePixel	0.912 1	0.944 9	180
		DE-JSMA	1.000 0	0.742 4	51

表 4 定向攻击对比结果

目标模型		攻击算法	R_{AS}	F_c	N_{avg}
AlexNet	密集	FGSM	0.233 0	0.307 0	12 000
		C&W	0.286 8	0.345 2	16 019
	稀疏	JSMA	0.809 9	0.613 7	226
		OnePixel	0.180 3	0.263 7	164
		DE-JSMA	0.787 9	0.678 4	98
VGG16	密集	FGSM	0.252 6	0.352 0	16 258
		C&W	0.448 5	0.428 6	16 093
	稀疏	JSMA	0.900 0	0.635 5	422
		OnePixel	0.201 3	0.274 6	174
		DE-JSMA	0.868 0	0.661 6	139
ResNet18	密集	FGSM	0.439 6	0.541 9	16 268
		C&W	0.609 9	0.504 9	16 075
	稀疏	JSMA	0.900 0	0.653 8	518
		OnePixel	0.283 5	0.321 7	163
		DE-JSMA	0.900 0	0.679 5	107

针对 SAR-ATR 系统高实时性的特点，所提出的攻击算法要有一定的时效性，因此将平均耗时 T_{avg} 设置成另一重要指标。DE-JSMA 自身具有稀疏特性，因此只和同类型的稀疏算法进行对比。稀疏攻击算法在非定向攻击和定向攻击时的平均耗时对比结果如表 5 所示，其中定向攻击的平均耗时为全类别平均耗时的均值。

由表 5 可知，非定向攻击时，DE-JSMA 耗时最短，均远远低于 OnePixel，与之相差 1～2 个数量级。然而在定向攻击时，尽管计算一次显著图选取多个显著特征，这种做法可起到省时的作用，但差分进化算法收敛缓慢，优化确定显著特征的像素值的过程会耗费大量时间，因此 DE-JSMA 比 JSMA 耗时略高，但若将差分进化算法替换成更先进的优化算法，算法的时间性能将会有进一步的提升。

表 5 稀疏攻击平均耗时对比结果 s

目标模型	攻击算法	非定向 T_{avg}	定向 T_{avg}
AlexNet	JSMA		82.693 8
	OnePixel	725.507 5	268.523 3
	DE-JSMA	17.052 0	197.179 2
VGG16	JSMA		63.954 9
	OnePixel	386.838 6	1 094.605 4
	DE-JSMA	7.783 4	143.182 1
ResNet18	JSMA		69.365 9
	OnePixel	81.987 9	763.264 0
	DE-JSMA	12.321 5	30.016 4

为了更全面地展示 DE-JSMA 算法定向攻击的性能，图 3 以热力图的方式展示了针对 3 种目标模型全类别的定向攻击结果。热力图纵轴的 0～10 表示干净样本的源类别，每个源类别从测试数据集中选取 10 个干净样本，共选取出 100 个干样本，横轴的 0～10 表示对抗样本的目标类别，图中数字表示源类别 10 个干净样本中，成功攻击成目标类别的对抗样本个数，热力图数值越大，颜色越深，表示该数值对应的源类别被攻击成目标类别的样本数越多，即该源类别更易被攻击成目标类别。单张热力图中定向攻击成功总次数除以定向攻击发起总次数可以计算得到，DE-JSMA 算法定向攻击 AlexNet、VGG16 和 ResNet18 时，分别能以 79.67%，93.56%，91.00% 的成功率将干净样本重定向为任意目标类别，这进一步说明了 DE-JSMA 在定向攻击中的有效性。

a) AlexNet b) VGG16 c) ResNet18

图 3 定向攻击源类别与目标类别相关性热力图

4　结　论

基于 DNN 的 SAR-ATR 系统存在严重的安全问题。考虑到 SAR 图像的特征稀疏性,本文提出了 DE-JSMA 稀疏攻击算法,在精确筛选出对模型推理结果影响较大的显著特征的同时,动态选择合适的特征值。本文还构建了一种结合攻击成功率和对抗样本平均置信度的新指标 F_c 值,来全面评价攻击算法的有效性和可靠性。实验结果表明,在没有增加过多耗时,且保证高攻击成功率的情况下,DE-JSMA 在定向、非定向 2 种攻击场景中均实现了可靠性更高、稀疏性更优的稀疏攻击,并且表现出了优越的重定向能力。后续工作中需进一步提升算法执行效率,以满足 SAR-ATR 系统的高实时性。

参考文献:

[1] CHEN S, WANG H, XU F, et al. Target classification using the deep convolutional networks for SAR images[J]. IEEE Trans on Geoscience and Remote Sensing, 2016, 54(8): 4806-4817

[2] SHARIFZADEH F, AKBARIZADEH G, SEIFI K Y. Ship classification in SAR images using a new hybrid CNN-MLP classifier [J]. Journal of the Indian Society of Remote Sensing, 2019, 47(4): 551-562

[3] VINT D, ANDERSON M, YANG Y, et al. Automatic target recognition for low resolution foliage penetrating SAR images using CNNs and GANs[J]. Remote Sensing, 2021, 13(4): 596

[4] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[C]//International Conference on Learning Representations, 2015

[5] MOOSAVI-DEZFOOLI S M, FAWZI A, FROSSARD P. DeepFool: a simple and accurate method to fool deep neural networks [C]//Computer Vision and Pattern Recognition, 2016: 2574-2582

[6] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//2017 IEEE Symposium on Security and Privacy, 2017: 39-57

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

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

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

[10] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[C]//International Conference on Learning Representations, 2014

[11] STORN R, PRICE K. Differential Evolution-a simple and efficient heuristic for global optimization over continuous spaces[J]. Journal of Global Optimization, 1997, 11(4): 341-359

[12] HUANG T, ZHANG Q, LIU J, et al. Adversarial attacks on deep-learning-based SAR image target recognition[J]. Journal of Network and Computer Applications, 2020, 162: 102632

[13] DU C, HUO C, ZHANG L, et al. Fast C&W: a fast adversarial attack algorithm to fool SAR target recognition with deep convolutional neural networks[J]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 4010005

[14] PENG B, PENG B, ZHOU J, et al. Speckle-variant attack: toward transferable adversarial attack to SAR target recognition[J]. IEEE Geoscience and Remote Sensing Letters, 2022, 19: 4509805

[15] 周隽凡, 孙浩, 雷琳, 等. SAR 图像稀疏对抗攻击[J]. 信号处理, 2021, 37(9): 1633-1643
ZHOU Juanfan, SUN Hao, LEI Lin, et al. Sparse adversarial attack of SAR image[J]. Journal of Signal Processing, 2021, 37(9): 1633-1643 (in Chinese)

DE-JSMA: a sparse adversarial attack algorithm for SAR-ATR models

JIN Xiaying[1], LI Yang[2], PAN Quan[2]

(1.School of Cyberspace Security, Northwestern Polytechnical University, Xi'an 710072, China;
2.School of Automation, Northwestern Polytechnical University, Xi'an 710072, China)

Abstract: The vulnerability of DNN makes the SAR-ATR system that uses an intelligent algorithm for recognition also somewhat vulnerable. In order to verify the vulnerability, this paper proposes DE-JSMA, a novel sparse adversarial attack algorithm based on a salient map's adversarial attack algorithm and differential evolution algorithm, with the synthetic aperture radar (SAR) image feature sparsity considered. After accurately screening out the salient features that have a great impact on the model inference results, the DE-JSMA algorithm optimizes the appropriate feature values for the salient features. In order to verify its effectiveness more comprehensively, a new metric that combines the attack success rate with the average confidence interval of adversarial examples is proposed. The experimental results show that DE-JSMA extends JSMA, which can be used only for targeted attack scenario, to untargeted attack scenario without increasing too much time consumption but ensuring a high attack success rate, thus achieving sparse adversarial attack with higher reliability and better sparsity in both attack scenarios. The pixel perturbations of only 0.31% and 0.85% can achieve the untargeted and targeted attack success rates up to 100% and 78.79% respectively.

Keywords: synthetic aperture radar; automatic target recognition; deep learning; adversarial attack; sparse attack

引用格式: *金夏颖, 李扬, 潘泉. DE-JSMA: 面向 SAR-ATR 模型的稀疏对抗攻击算法[J]. 西北工业大学学报, 2023, 41(6): 1170-1178*

JIN Xiaying, LI Yang, PAN Quan. DE-JSMA: a sparse adversarial attack algorithm for SAR-ATR models[J]. Journal of Northwestern Polytechnical University, 2023, 41(6): 1170-1178 (in Chinese)