基于显著图的电磁信号对抗样本生成方法

周 侠,张 剑,李宁安
(武汉数字工程研究所,湖北武汉 430205)

摘 要: 基于深度学习的电磁信号识别模型具有高效、准确和人工干预少的优点,然而其与传统神经网络模型一样容易受到对抗样本的影响. 研究对抗样本对测试和提升模型的安全性和鲁棒性有着重要意义. 为生成高质量电磁信号对抗样本,本文提出了基于雅可比显著图批量特征点攻击算法(Batch Points Jacobian-based Saliency Map Attack, BP-JSMA). 与传统雅可比显著图的攻击方法相比,BP-JSMA 通过批量选取关键特征点能够更快生成对抗样本. 此外,针对电磁信号数据的特点,增加自适应扰动限制,使得生成的对抗样本更具隐蔽性. 在公开数据集的实验结果表明,与雅可比显著图攻击方法相比,BP-JSMA 在生成速度方面提升了 11 倍,隐蔽性提升了 10%;而与传统快速梯度符号攻击算法相比,攻击成功率提升了 24%,隐蔽性提升了 20%.

关键词: 人工智能;深度学习;对抗样本;电磁信号识别;显著图;目标攻击
基金项目: 国防科技技术领域基金项目(A 类)(No.A24011)
中图分类号: TP183 **文献标识码:** A **文章编号:** 0372-2112(2023)07-1917-12
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20230125

An Electromagnetic Signal Adversarial Examples Generation Method Based on Saliency Map

ZHOU Xia, ZHANG Jian, LI Ning-an
(*Wuhan Digital Engineering Institute*, *Wuhan*, *Hubei* 430205, *China*)

Abstract: The electromagnetic signal recognition model based on deep learning has the advantages of high efficiency, high accuracy, and less manual intervention. However, it is as susceptible to adversarial examples as traditional neural network models. Studying adversarial examples is important for testing and improving the security and robustness of neural network models. In order to generate high-quality electromagnetic signal adversarial examples, this paper proposed a batch feature points Jacobian-based saliency map attack method (BP-JSMA). Compared with the traditional Jacobian saliency map attack method, BP-JSMA can generate adversarial examples faster by selecting key feature points in a batch. In addition, according to the characteristics of electromagnetic signal data, adaptive disturbance limitation is proposed to make the generated adversarial examples more covert. Experimental results on public datasets show that compared to the Jacobian saliency map attack method, BP-JSMA improves generation speed by 11 times and concealment by 10%. Compared with traditional fast gradient sign attack method, the attack success rate has been improved by 24%, and the concealment has been improved by 20%.

Key words: artificial intelligence; deep learning; adversarial examples; electromagnetic signal recognition; saliency map; target attack
Foundation Item(s): National Defense Science and Technology Fund Projects of China (Category A) (No.A24011)

1 引言

随着无线电磁技术的不断发展,电磁空间争夺已成为军事斗争的重要维度,因此对电磁信号的监测与识别具有重大的军事意义. 深度神经网络(Deep Neural Network,DNN)在图像识别[1]等领域有着巨大优势,受此启发,研究者将其应用于电磁信号调制识别领域并取得不错效果[2~5]. 其中,文献[2~4]以电磁信号的同相和正交(In-phase and Quadrature,I、Q)两路数据为输入,以调制方式为输出完成调制类别识别;文献[5]将电磁信号转换为图像,进而使用传统图像识别模型进

收稿日期:2023-02-13;修回日期:2023-05-08;责任编辑:覃怀银

行识别. 然而,研究表明深度学习模型容易受到“对抗样本”[6]的影响. 对抗样本指的是向正常样本中添加细微扰动生成使得深度模型决策错误的样本. 利用对抗样本和正常样本进行“对抗训练”能有效提升模型的鲁棒性. 因此,研究对抗样本生成对提升模型鲁棒性和安全性具有现实意义.

在图像识别领域,对抗样本生成方法主要有快速梯度符号算法(Fast Gradient Sign Method,FGSM)[7]及其变体[8~10],这类方法通过全局添加扰动生成对抗样本,在生成样本时往往只要求模型识别出错即可,攻击力较弱. 为减少改动数据点的数量,研究人员提出基于显著图的对抗样本生成方法[11,12],这类方法通过寻找对抗显著特征点生成对抗样本,同时能够将样本攻击为攻击者指定类别,具有较强攻击性. 为了扩充电磁信号识别领域对抗样本研究,研究人员以图像领域FGSM为基础,提出了不同的改进方法[13~16].

目前,针对电磁信号的对抗样本生成方法多数为全局攻击,在添加对抗扰动时缺乏特异性,导致生成的对抗样本质量低. 为生成更高质量的电磁信号对抗样本以完成强力攻击,受图像识别中雅可比显著图攻击算法(Jacobian-based Saliency Map Attack,JSMA)[11]的启发,本文提出一种利用显著图进行扰动添加生成对抗样本的技术思路:首先计算正常电磁样本的特征显著图,然后根据特征显著图批量选择关键特征点,在关键特征点上添加对抗扰动生成对抗样本. 基于该思路,提出基于雅可比显著图的批量特征点攻击算法(Batch Points-JSMA, BP-JSMA). 实验表明BP-JSMA能够将样本攻击为指定类别,同时提升了对抗样本的生成速度和隐蔽性,证明了该算法的可行性和有效性.

2 相关工作研究

2.1 基于深度神经网络的电磁信号识别

利用DNN的平移不变性,O'Shea等人先后将卷积神经网络(Convolutional Neural Networks,CNNs)和残差网络(Residual Networks,ResNets)[17]进行调整构建了新的CNN[2]和ResNet[3]模型,进而用于电磁信号识别,同时构建了开源数据集RML2016.10A和2018.01.OSC.

在这两个数据集中,每个样本均由正交同步采样的I、Q两维数据构成,区别在于数据量和包含信噪比(Signal-to-Noise Ratio, SNR)范围的不同. CNN和ResNet模型分别将数据集中的样本作为输入,以信号的调制方式作为输出,减少了专家特征提取的步骤,实现了端到端的识别.

2.2 对抗攻击的概念及分类

对于一个已训练好的深度学习模型$F(x)$:$x\in D\rightarrow y\in Y$,其对于干净样本x有着唯一正确的输出y. 在x邻域内寻找对抗样本x^{adv}并利用其攻击深度学习模型的方式称为对抗攻击. 根据攻击者的攻击目的可将对抗攻击分为“目标攻击”和“非目标攻击”. 目标攻击使得目标模型将对抗样本x^{adv}识别为攻击者指定的类别t,如式(1)所示,此时$t\neq y$;非目标攻击只需使得模型识别为非y类,如式(2)所示.

$$\begin{cases} x^{\text{adv}}=x+\eta \\ F\left(x^{\text{adv}}\right)=t\neq y \end{cases} \tag{1}$$

$$\begin{cases} x^{\text{adv}}=x+\eta \\ F\left(x^{\text{adv}}\right)\neq y \end{cases} \tag{2}$$

2.3 对抗样本生成方法

2.3.1 快速梯度符号方法

Goodfellow等人[7]从损失梯度入手提出了FGSM方法,该方法通过最大化模型的损失函数来添加扰动,仅需一步便能生成对抗样本.

$$x^{\text{adv}}=x+\varepsilon\cdot\text{sign}\left(\nabla_x L\left(x,y;\ \theta\right)\right) \tag{3}$$

式(3)中,x是原样本输入,y是样本的真实标签,θ是深度学习模型的参数,$L(\cdot)$是交叉熵损失函数,∇_x表示对损失函数求导(也即原始样本的梯度信息),sign$(\cdot)$是符号函数,用于标记梯度方向;ε为超参,用于控制扰动幅度. FGSM优点是扰动计算快,缺点是扰动大,攻击成功率不高.

2.3.2 动量迭代快速梯度符号方法

Dong等人[8]在FGSM的基础上,结合物理学中动量的思想提出了基于动量的多次迭代快速梯度符号方法(Momentum Iterative Fast Gradient Sign Method, MI-FGSM),如式(4)所示.

$$\begin{cases} x_0^{\text{adv}}=x, g_0=0 \\ g_{n+1}=\mu g_n+\dfrac{\nabla_{x_n^{\text{adv}}}L\left(x_n^{\text{adv}},y\right)}{\nabla_{x_n^{\text{adv}}}L\left(x_n^{\text{adv}},y\right)_1} \\ x_{n+1}^{\text{adv}}=\text{Clip}_{x,\varepsilon}\left\{x_n^{\text{adv}}+\alpha\cdot\text{sign}\left(g_{n+1}\right)\right\} \end{cases} \tag{4}$$

其中,g_{n+1}为前n次迭代中累加的梯度,μ为动量系数,Clip$(\cdot)$为裁剪函数,当超过阈值时进行裁剪限制.

2.3.3 投影梯度下降方法

基于迭代思想,Madry等人[9]提出了投影梯度下降方法(Projected Gradient Descent,PGD),少量多次添加扰动,一旦扰动超过预先设定的范围,则进行投影操作将对抗扰动约束到合法范围,如式(5)所示.

$$\begin{cases} x_0^{\text{adv}}=x \\ x_{i+1}^{\text{adv}}=\prod_{\varepsilon}\left\{x_i^{\text{adv}}+\alpha\cdot\text{sign}\left(\nabla_x L\left(x_i^{\text{adv}},y;\ \theta\right)\right)\right\} \end{cases} \tag{5}$$

其中,$\prod_{\varepsilon}(\cdot)$表示投影过程,保证对抗扰动始终在阈值$\varepsilon$内,$\alpha$为单步扰动大小. 相较于FGSM,PGD能够在增加损失函数的同时尽可能小的添加扰动.

2.3.4　基于雅可比显著图方法

Papernot 等人[11]首次将“显著图”的概念引入对抗样本生成领域,并基于雅可比前向导数提出了基于雅可比显著图的对抗样本生成方法 JSMA.“显著图”的作用是返回输入对输出结果的影响,显著性越强则影响越大.JSMA 首先计算攻击类别的雅可比前向导数,然后根据雅可比前向导数生成特征显著图,之后在显著图中通过式(6)迭代寻找显著性最强的特征点对(p_1,p_2),最后在(p_1,p_2)上添加对抗扰动生成对抗样本.

$$\arg\max_{(p_1,p_2)}\left(\sum_{i=p_1,p_2}\frac{\partial F_t(x)}{\partial x_i}\right)\times\left|\sum_{i=p_1,p_2}\sum_{j\neq t}\frac{\partial F_t(x)}{\partial x_i}\right| \tag{6}$$

其中,$\partial\left[F_t(x)/x_i\right]$表示样本 x 中第 i 个数据点在输出类别 t 上的显著性.

JSMA 的主要思想是通过改动较少的数据点,达到对抗攻击的目的.但是,虽然改动的数量少,但会使得在某些显著特征点上的改动很大,肉眼能够轻易察觉.

3　基于显著图的批量特征点 BP-JSMA 攻击算法

3.1　BP-JSMA 攻击算法基本思路

传统 JSMA 攻击方法利用雅可比前向导数可以反映输入对输出影响程度的特性,通过生成特征显著图寻找关键特征进行扰动添加,直至生成对抗样本.具体流程如图 1 中路径(1)所示:首先选择目标标签 t,通过目标模型 F 计算雅可比前向导数 J,然后通过雅可比前向导数生成特征显著图 S,之后通过 S 选择显著性最强的特征点对(p_1,p_2),再在(p_1,p_2)上添加扰动 ε 得到单次扰动,最后将单次扰动添加到原始样本 x 上即可生成中间对抗样本 x^{adv},如果 $F(x^{adv})=t$,则返回对抗样本,否则再判断每个点添加扰动是否大于阈值 γ,小于 γ 则在(p_1,p_2)上再次添加扰动 ε,否则判断显著图中特征点是否为空,为空则生成对抗样本失败,否则选择新的特征点对,迭代添加扰动,直至生成满足条件的对抗样本 x^{adv}.

然而,JSMA 通过计算显著特征点对选择关键特征点时存在以下限制:特征点对组合方式过多,计算存在负担.而且由于电磁信号数据并不像图像数据一样存在初始阈值(黑白图像[0,1],彩色图像[0,255]),因此该方法在用于电磁信号时会使得单点添加的扰动过大,降低了对抗隐蔽性.基于这两点,本文在 JSMA 的研究基础上提出批量特征点 BP-JSMA 攻击方法,该方法利用 FGSM 的思想,在选取特征点时根据显著图 S 批量选择.此外,为了增强对抗样本的隐蔽性,设置一个阈值对其进行限制,然而不同类别的电磁信号的峰值范围大相径庭,因此需要对阈值进行自适应处理.BP-JSMA 流程如图 1 中路径(2)所示,具体描述见 3.2 节.

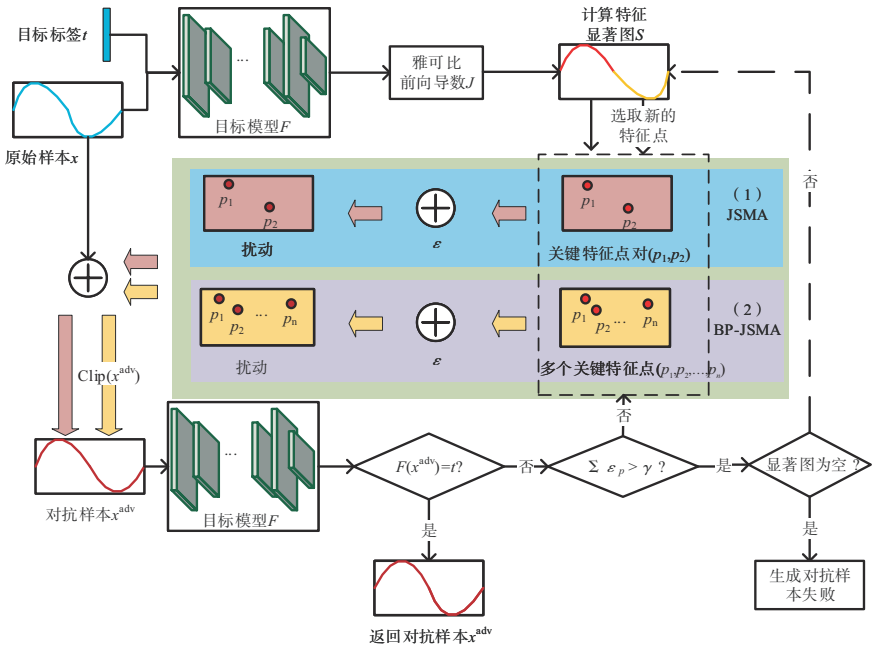


图 1　BP-JSMA 攻击流程图

3.2　BP-JSMA 攻击方法描述

通过计算函数输出关于输入的雅可比前向导数,可以得到输入对输出的影响程度,值越大,说明影响程度越高.深度神经网络属于特殊的函数映射,因此,也可计算输入 x 在标签 r 上的雅可比前向导数,如式(7)所示.

$$J(F_r(x))=\frac{\partial F_r(x)}{\partial x}=\left[\frac{\partial F_r(x)}{\partial x_i}\right]_{i\in\{1,2,\cdots,a\times b\}} \tag{7}$$

其中，$F_r(x)$表示输入x在类别r上的得分，x_i表示x中的第i个特征点，x共有$a\times b$个特征点.

由于神经网络并不是简单的函数计算，其中间隐含有许多隐藏层，因此需要根据函数的链式计算法则求得雅可比前向导数，如式(8)所示.

$$\frac{\partial F_r(x)}{\partial x}=\frac{\partial F_r(x)}{\partial H(x)}\cdot\frac{\partial H(x)}{\partial x} \tag{8}$$

其中，$H(x)$表示模型F的中间隐藏层，一般情况下，$H(x)$如式(9)所示，其中k表示第k隐藏层，w表示$k-1$层输出作为第k层输入的权重，b表示偏置，f表示激活函数.

$$H_k(x)=f_k\left(w\cdot H_{k-1}(x)+b_k\right) \tag{9}$$

所以，$\partial H_k(x)/\partial x$的计算如式(10)所示.

$$\frac{\partial H_k}{\partial x}=\frac{\partial\left[f_k\left(w\cdot H_{k-1}(x)+b_k\right)\right]}{\partial x_i} \tag{10}$$

计算得到所有类别的雅可比前向导数J，然后计算生成指定类别的显著图S. 显著图生成分为扰动添加和扰动减少两个方向. 扰动添加方向指的是特征点值的增加会导致目标类的分类得分增加，且非目标类的得分下降，此时特征点对目标类的分类结果影响为正，如式(11)所示. 扰动减少方向指的是特征值的增加会导致目标类的分类得分降低，且非目标类的得分增加，此时特征点对目标类的分类结果影响为负，如式(12)所示.

$$S(x,t)[i]=\begin{cases}J_{it}(x)\left|\sum\limits_{r\neq t}J_{ir}(x)\right|, J_{it}(x)>0\wedge\sum\limits_{r\neq t}J_{ir}(x)<0\\ 0, J_{it}(x)=0\wedge\sum\limits_{r\neq t}J_{ir}(x)=0\end{cases} \tag{11}$$

$$S(x,t)[i]=\begin{cases}0, J_{it}(x)=0\wedge\sum\limits_{r\neq t}J_{ir}(x)=0\\ J_{it}(x)\sum\limits_{r\neq t}J_{ir}(x), J_{it}(x)<0\wedge\sum\limits_{r\neq t}J_{ir}(x)>0\end{cases} \tag{12}$$

式(11)、式(12)中，x表示输入样本，t表示攻击类别，r表示其他非目标类别，i表示样本x中的第i个特征点.

得到类别t的显著图之后，传统JSMA根据式(6)选择关键特征点对(p_1,p_2)，然后根据所选择的特征点对迭代添加扰动，直至生成对抗样本.

为了加快对抗样本生成速度，BP-JSMA则通过批量选取特征点的方式进行. 为了使得选取更具客观性，本文设置一次特征点的选取数量为显著图的5%，假设S中有N个非零特征点，则一次迭代选择特征点数$n=N\times 5\%$. 具体操作为：将显著图内的各个特征点的显著值进行排序，然后选择前5%显著性最高的特征点. 因此特征点的选取如式(13)所示.

$$\begin{aligned}S'(x,t)&=\left[\mathrm{Sort}\left(S(x,t)[i]\right)\right]_{i\in\{1,2,\cdots,a\times b\}}\\ p_i&=S'(x,t)[i]_{i\in\{1,2,\cdots,n\}}\end{aligned} \tag{13}$$

之后添加单步扰动ε生成对抗样本，为了增强隐蔽性，将单点扰动限制θ设为样本x的数据点最大值或最小值绝对值的1/3(与扰动添加方向有关，该值是一个经验值，如果设置过大则约束效果不佳，扰动容易察觉，太小则会导致需要扰动更多的特征点，多次实验显示1/3比较合适)，如式(14)所示. 然后通过式(15)迭代生成对抗样本.

$$\theta=\frac{1}{3}\cdot\max\left(x_i\right) \tag{14}$$

$$x_n^{\mathrm{adv}}=\mathrm{Clip}_\theta\left[\left(x_{i,n-1}^{\mathrm{adv}}\right)_{i\in\{p_1,p_2,\cdots,p_n\}}+\varepsilon\right] \tag{15}$$

式(15)中，x_n^{adv}表示第n次迭代生成对抗样本，$\mathrm{Clip}_\theta(\cdot)$为裁剪函数，将添加的扰动限制在$\theta$以内. 如果该批数据点不能生成对抗样本，则判断显著图是否为空，若不为空则再次根据式(13)选择下一批特征数据点，然后再根据式(15)循环添加扰动，若为空则停止运算，生成对抗样本失败.

BP-JSMA具体描述如算法1所示.

算法1 基于显著图的批量特征点BP-JSMA攻击算法

输入：正常样本x，深度学习模型F，目标类别t，单步扰动大小ε，单个数据点扰动限制θ，总扰动限制γ，迭代轮数m.

输出：对抗样本x^{adv}，$F(x^{\mathrm{adv}})=t$.

```
将样本x输入模型F，并返回各个类别得分F_r(x)
根据F_r(x)和式(7)计算各个类别的雅可比前向导数J(F_r(x))
根据式(11)或式(12)生成类别t的特征显著图S(x,t)
WHILE 显著图不为空
    FOR (i=0, i<m, i++)
        根据式(13)选择特征点{p_1,p_2,…,p_n}
        根据式(15)生成对抗样本
        IF F(x^adv)=t
            RETURN 对抗样本x^adv;
        ELSE
            IF 单点扰动小于θ AND 总扰动小于γ
                CONTINUE;
            ELSE
                生成对抗样本失败;
                BREAK FOR
    END FOR
END WHILE
```

4 实验结果与分析

4.1 实验设置

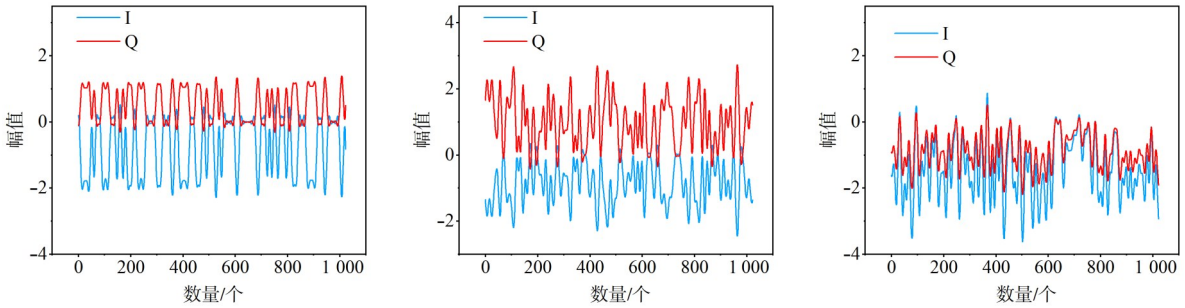
对于数据集，本文选取了O'Shea等人构建的公开数据集2018.01.OSC[3]和RML2016.10A[2]. 对于网络模型结构，针对2018.01.OSC数据集，本文选择了O'Shea

等人在构建该数据集时搭建的CNN和ResNet模型;针对RML2016.10A,我们选择了Sadeghi等人在对抗攻击时搭建的VT-CNN2[18]和Sainath等人构建的CLDNN模型[19]. 对于评估指标,本文从攻击有效性、攻击效率和对抗样本隐蔽性三个维度进行指标选择.

4.1.1　数据集

(1)　2018.01.OSC

该数据集以调制方式作为分类类别,包含24种调制方式:OOK,4ASK,8ASK,BPSK,QPSK,8PSK,16PSK,32PSK,16APSK,32APSK,64APSK,128APSK,16QAM,32QAM,64QAM,128QAM,256QAM,AM-SSB-WC,AM-SSB-SC,AM-DSB-WC,AM-DSB-SC,FM,GMSK,OQPSK. 其中,每种调制方式包含26种SNR,SNR的范围在[−20 dB,30 dB],间隔2 dB,每种SNR包含4 096条数据,每条数据包含I、Q两路数据,每路数据有1 024个数据点. 部分数据可视化如图2所示,其中图2(*a*)~(*c*)分为类别32PSK,16APSK和32QAM在SNR=30 dB的样本.

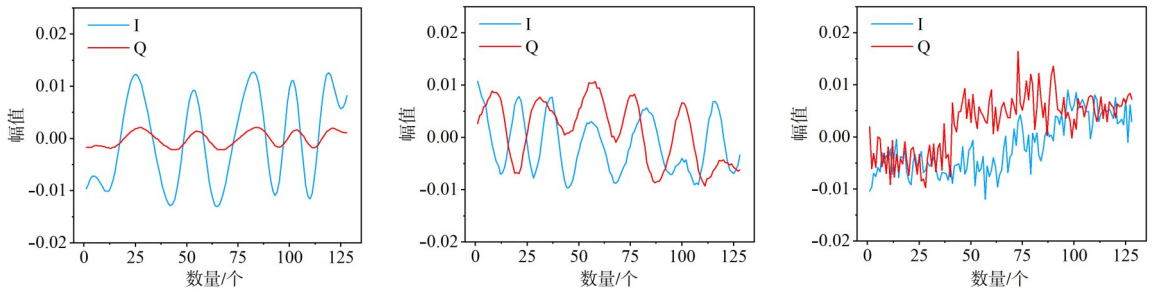


(*a*) 32PSK　(*b*) 16APSK　(*c*) 32QAM

图2　2018.01.OSC部分数据可视化

(2)　RML2016.10A

该数据集包含11类调制方式,其中包括8种数字调制方式(8PSK,BPSK,CPFSK,GFSK,PAM4,16QAM,64QAM,QPSK)和3种模拟调制方式(AM-DSB,AM-SSB,WBFM). 每种调制方式包含20种SNR(范围在[−20 dB,18 dB]之间,间隔2 dB),每种SNR下有1 000条数据,每条数据同样包含I、Q两路数据,每路数据有128个数据点. 部分数据如图3(*a*)~(*c*)所示.



(*a*) 8PSK(18 dB)　(*b*) 8PSK(10 dB)　(*c*) GFSK(0 dB)

图3　RML2016.10A部分数据可视化

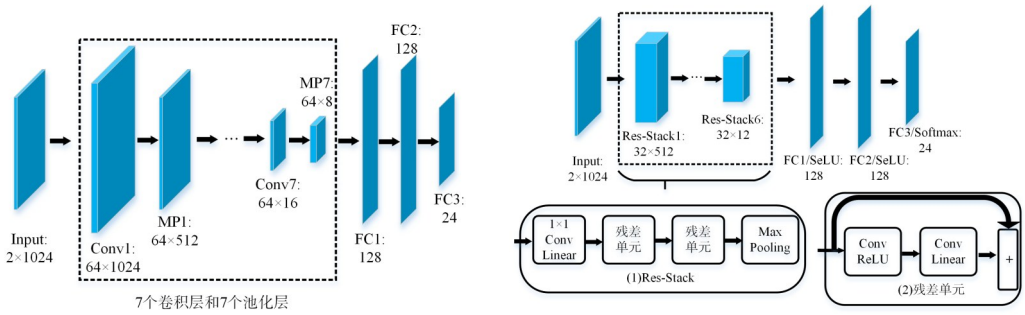
实验中将数据集按7:3随机划分为训练集和测试集,为了验证攻击的有效性,仅使用测试集中的数据进行添加对抗扰动.

4.1.2　受攻击模型

对抗攻击的目标模型的优劣是反映攻击效果的关键,如果目标模型本身识别正确率不高,那么攻击也就失去了意义.

如图4所示,O'Shea等人[3]在构建数据集2018.01.OSC时搭建了CNN和ResNet模型. 具体地,CNN模型如图4(*a*)所示,以电磁信号I、Q为输入,调制方式为输出,该模型包括1个输入层、7个卷积层、7个池化层和3个全连接层,FC1和FC2后连接SeLU激活函数,FC3使用Softmax激活函数最终得到模型输出. ResNet模型如图4(*b*)所示,包含了1个输入层、6个残差模块和3个全连接层,前两个全连接层同样采用SeLU激活函数,而FC3采用Softmax激活函数. 每个残差块包含1个1×1的线性卷积层、2个残差单元和1个最大池化层. 每个残差单元包含1个连接ReLU激活函数的卷积层和1个线性卷积层.

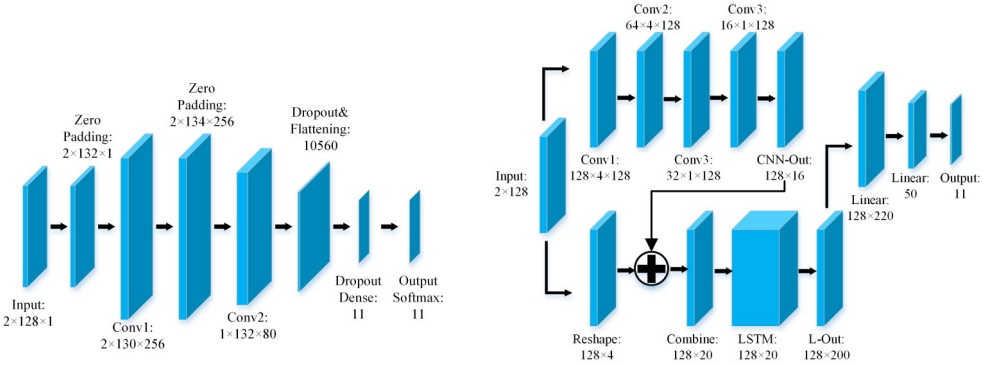
针对数据集RML2016.10A,Sadeghi等人[18]搭建的VT-CNN2模型如图5(*a*)所示,该模型包含了1个输入



(*a*) CNN模型结构 (*b*) ResNet模型结构

图4 用于2018.01.OSC数据集的网络结构

层、2个卷积层、2个全连接层和1个输出层. 此外,如图5(*b*)所示,Sainath等人[19]将CNN和长短时记忆网络LSTM[20](Long Short-Term Memory)结合搭建了CLDNN网络模型,该模型首先经过4个卷积层,然后将输入层Reshape之后与经过卷积层后的输出相加,再经过LSTM模块,最后输入2个全连接层.



(*a*) VT-CNN2模型结构 (*b*) CLDNN模型结构

图5 用于RML2016.10A数据集的网络结构

4.1.3 评估指标

为有效评估本文算法,从攻击有效性、攻击效率和对抗样本隐蔽性三个维度选取了不同的评估指标. 其中攻击有效性由模型的脆弱性体现,模型越脆弱则攻击越有效;攻击效率和对抗样本隐蔽性由对抗样本体现,样本生成速度越快则效率越高,样本与原始样本更相似则隐蔽性越强.

(1) 攻击有效性

(*a*) 攻击成功率

攻击成功率(Attack Success Rate,ASR)从对抗样本角度出发,指的是使得目标模型出错的样本数占总样本数的比率,该指标反映了攻击的有效性. 假设总测试样本数为M,成功攻击模型的样本数为N,则ASR如式(16)所示:

$$\mathrm{ASR}=\frac{N}{M}\times 100\% \tag{16}$$

(*b*) 正确类别平均置信度

正确类别平均置信度(Average Confidence of True Class,ACTC)从模型角度出发,指的是针对一批攻击成功的对抗样本,模型在其正确类别上的平均置信度,该值越小,则攻击效果越好. 假设模型对单个样本的正确类别的置信度为a,总样本数为M,则该指标计算方法如式(17)所示:

$$\mathrm{ACTC}=\left(\frac{1}{M}\sum_{1}^{M}a\right)\times 100\% \tag{17}$$

(*c*) 对抗类别平均置信度

对抗类别平均置信度(Average Confidence of Adversarial Class,ACAC)同样从模型角度出发,指的是对一批攻击成功的对抗样本,模型对这些样本在攻击者指定类别上的平均置信度,该值越大,攻击效果越好. 假设单个样本的攻击类别的置信度为b,总样本数为M,则该指标计算方法如式(18)所示:

$$\mathrm{ACAC}=\left(\frac{1}{M}\sum_{1}^{M}b\right)\times 100\% \tag{18}$$

(2) 攻击效率

攻击效率由平均生成单个对抗样本所需的消耗时长(Average Time Consumption,ATC)体现. 假设成功攻击模型的样本数为N,总耗费时长为T,则ATC如式(19)所示:

$$\mathrm{ATC}=\frac{T}{N} \tag{19}$$

(3) 对抗样本隐蔽性

(a) 对抗样本与样本间结构相似度

对抗样本与样本间结构相似度(Structural SIMilarity,SSIM)是图像相似度的衡量指标,指的是两个样本之间的结构差异. SSIM计算过程如式(20)所示,其中包括计算对比两张图片的亮度$l(x,x^{\mathrm{adv}})$、对比度$c(x,x^{\mathrm{adv}})$和结构$s(x,x^{\mathrm{adv}})$,式中$\alpha>0,\beta>0,\gamma>0$为调整亮度、对比度和结构的相对重要性参数. 具体亮度、对比度和结构的计算如式(21)所示,式中μ_{x_1}和μ_{x_2}、σ_{x_1}和σ_{x_2}分别为x_1和x_2的平均值和标准差,$\sigma_{x_1x_2}$为x_1和x_2的协方差,C_1、C_2、C_3均为常数,用以避免分母趋于0. SSIM反映了生成对抗样本的隐蔽性,取值范围为0到1,越接近1则说明二者越相似.

$$\begin{aligned}\mathrm{SSIM}\left(x,x^{\mathrm{adv}}\right)=&\left[l\left(x,x^{\mathrm{adv}}\right)\right]^{\alpha}\cdot\left[c\left(x,x^{\mathrm{adv}}\right)\right]^{\beta}\\&\cdot\left[s\left(x,x^{\mathrm{adv}}\right)\right]^{\gamma}\end{aligned} \tag{20}$$

$$\begin{cases}l\left(x_1,x_2\right)=\dfrac{2\mu_{x_1}\mu_{x_2}+C_1}{\mu_{x_1}^2+\mu_{x_2}^2+C_1}\\c\left(x_1,x_2\right)=\dfrac{2\sigma_{x_1}\sigma_{x_2}+C_2}{\sigma_{x_1}^2+\sigma_{x_2}^2+C_2}\\s\left(x_1,x_2\right)=\dfrac{\sigma_{x_1x_2}+C_3}{\sigma_{x_1}\sigma_{x_2}+C_3}\end{cases} \tag{21}$$

由于将电磁信号样本视为图像处理,则其亮度和对比度的计算结果都为1,因此实际只需计算二者结构相似指标$s(x,x^{\mathrm{adv}})$即可.

(b) 扰动比率L_0

扰动比率L_0指的是样本中改动数据点占整个样本数据点的比率,该指标值取值范围为0到1,越小则说明改动的数据点越少. 假设数据样本维度为$n\times m$,单个对抗样本平均添加扰动的数据点的数量为p,则L_0如式(22)所示:

$$L_0=\frac{p}{n\times m}\times 100\% \tag{22}$$

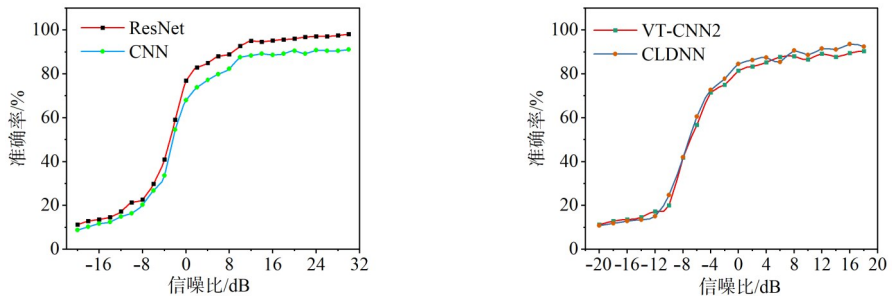
(c) 最大单点扰动L_∞

最大单点扰动L_∞指的是对抗样本中数据点添加的最大扰动值,该值反映了单个数据点的扰动大小,值越大说明扰动越大,也更容易被人眼察觉,因此对抗样本的隐蔽性越高该指标值越小.

4.2　实验结果及分析

4.2.1　目标模型训练设置及结果

本文使用的深度学习框架为Mindspore1.6.2,GPU为CUDA11.1. 在目标模型训练阶段,CNN模型、ResNet模型、VT-CNN2模型和CLDNN模型的迭代次数、学习率、损失函数等均保持一致,其中迭代次数设置为1 000,初始学习率设置为0.01,并且加入了学习率自动更新机制:每训练10轮则学习率减半. 由于训练数据过多,且没有预训练权重,因此为了加快训练速度,本次实验在训练集的每个类别的每个信噪比上又随机选取70%的数据作为最终的训练集,验证集则保持不变. 最终在各个信噪比下测试训练模型的识别准确率,结果如图6所示.



(a) 2018.01.OSC上的模型准确率　　(b) RML2016.10A上的模型准确率

图6　模型准确率

由图6可知,随着信噪比的增加,四种模型的准确率都呈现上升趋势. 对于2018.01.OSC数据集,其正确率如图6(a)所示,ResNet模型略优于CNN模型. 当信噪比较低时,模型的准确率通常不高,因为此时噪声占比很大,已经高于信号本身的功率,这就导致信号波形严重失真. 为了验证本文对抗攻击的有效性,需要选择信噪比较高的信号数据,因为模型对信噪比较低的信号识别率不高,也就失去了攻击的意义. 在ResNet模型和CNN模型中,当信噪比大于12 dB时,模型的正确率均超过了90%. 对于RML2016.10A数据集,如图6(b)所示,在VT-CNN2模型和CLDNN模型中,当信噪比大于4 dB时,模型准确率达到了88%左右,满足对抗攻击

对目标模型的需求.

4.2.2 攻击效果及分析

为了验证本文方法的有效性,引入了传统JSMA[11]、FGSM[7]、MI-FGSM[8]和PGD[9]方法,为便于与本文方法对比,将这些方法设置为目标攻击方式. 实验结果如表1~4所示.

表1 CNN模型受攻击结果

	ASR/%	ACTC/%	ACAC/%	ATC/s	SSIM/%	L_0/%	L_∞
BP-JSMA	**97.5**	0.12	75.3	28.5	**90.3**	9.3	**0.5**
JSMA	96.8	**0.09**	**77.5**	366.4	82.6	**4.3**	1.8
FGSM	65.4	9.26	48.7	**0.4**	55.8	95.8	0.9
MI-FGSM	73.3	7.43	54.6	0.7	66.9	94.6	0.8
PGD	68.5	8.58	51.8	1.3	65.4	56.2	0.8

表2 ResNet模型受攻击结果

	ASR/%	ACTC/%	ACAC/%	ATC/s	SSIM/%	L_0/%	L_∞
BP-JSMA	**98.3**	0.22	72.4	43.2	**91.5**	11.2	**0.5**
JSMA	96.4	**0.14**	**76.2**	486.7	80.7	**5.7**	2.3
FGSM	60.7	11.71	51.7	**0.5**	57.2	94.8	1.6
MI-FGSM	71.5	9.39	57.4	0.8	63.4	95.5	0.9
PGD	65.9	10.48	54.9	1.6	66.8	60.3	1.4

(1) 攻击成功率ASR

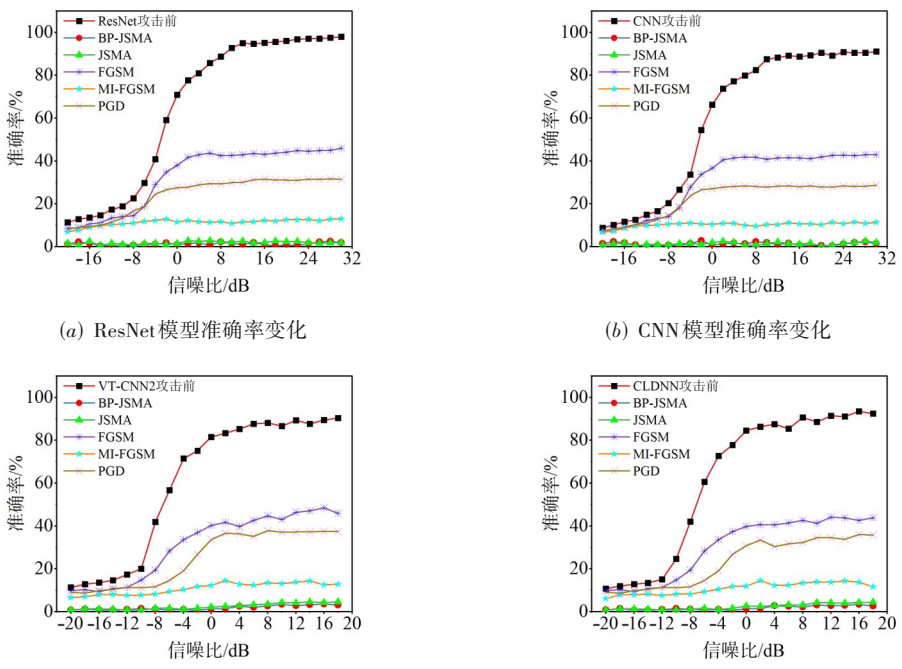
由表1~4可知,无论CNN模型、ResNet模型、VT-CNN2模型还是CLDNN模型,对抗样本攻击成功率都在60%以上,其中BP-JSMA与JSMA攻击成功率甚至高达95%以上,高于FGSM、MI-FGSM和PGD. 攻击成功率可以由模型在对抗样本上的正确率体现,攻击成功率越高,则模型的正确率越低. 如图7所示,在BP-JSMA和JSMA攻击下,四种模型的正确率均下降到10%以下,而在FGSM、MI-FGSM和PGD的攻击下也有不同程度的降低,但总体来说BP-JSMA和JSMA对模型的攻击成功率更高.

表3 VT-CNN2模型受攻击结果

	ASR/%	ACTC/%	ACAC/%	ATC/s	SSIM/%	L_0/%	L_∞
BP-JSMA	**97.6**	**0.36**	**71.5**	16.6	**89.7**	13.7	**0.42**
JSMA	97.1	0.40	67.4	201.3	76.3	**7.9**	2.6
FGSM	62.4	8.56	44.7	**0.4**	55.2	93.7	1.6
MI-FGSM	76.3	8.44	56.2	0.6	62.6	96.6	1.2
PGD	69.2	9.39	49.8	1.2	65.9	67.6	0.9

表4 CLDNN模型受攻击结果

	ASR/%	ACTC/%	ACAC/%	ATC/s	SSIM/%	L_0/%	L_∞
BP-JSMA	**97.3**	0.39	**66.9**	20.8	**92.2**	15.1	**0.38**
JSMA	96.5	**0.38**	62.1	266.3	79.8	**8.6**	2.1
FGSM	65.9	9.77	42.8	**0.7**	51.7	92.8	1.7
MI-FGSM	78.6	8.36	53.7	0.9	65.3	94.3	1.3
PGD	68.8	9.48	46.9	1.6	67.4	69.9	1.3



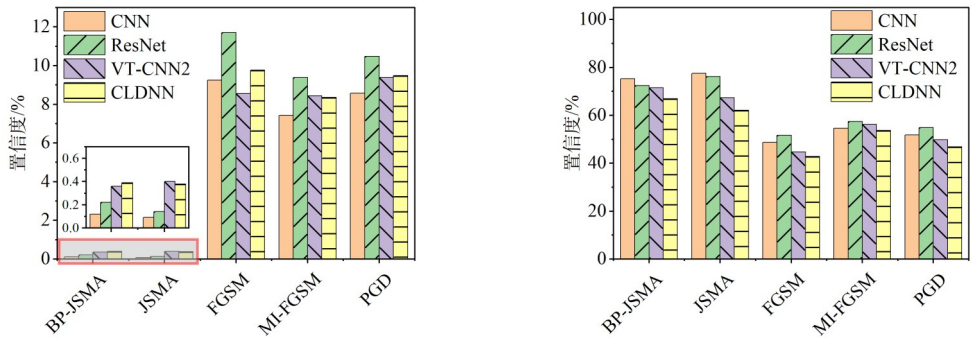
(a) ResNet模型准确率变化 (b) CNN模型准确率变化

(c) VT-CNN2模型准确率变化 (d) CLDNN模型准确率变化

图7 模型准确率变化

(2) 正确类别平均置信度 ACTC 和对抗类别平均置信度 ACAC

由表 1~4 绘制如图 8 所示的 ACTC、ACAC 对比图，其中可知，模型在对抗样本的决策上出现了较大偏差，如图 8(a)所示在正确类别的置信度 ACTC 大都小于 10%；而如图 8(b)所示在对抗类别上的置信度 ACAC 在 40% 以上. 在 BP-JSMA 和 JSMA 的攻击下，ACTC 的值甚至小于 1%，而 ACAC 大于 60%. 由此可知，BP-JSMA 能够保持与 JSMA 相同的攻击水平，且强于 FGSM、MI-FGSM 和 PGD.



(a) ACTC　　(b) ACAC

图 8　ACTC、ACAC 对比

(3) 平均耗时 ATC

由表 1~4 可知，BP-JSMA 在生成速度 ATC 方面虽然与 FGSM、MI-FGSM、和 PGD 有着不小差距，但明显优于 JSMA. 出现该现象的原因是因为 FGSM 是单步攻击方法，不需要寻找特异性扰动，只需计算梯度符号即可完成攻击，因此速度很快. 同理 MI-FGSM 作为变体也继承了 FGSM 快速的特性，而 PGD 算法也是利用投影梯度进行攻击，因此速度也较快. 而 BP-JSMA 通过批量选择特征点，这使得一次迭代比 JSMA 添加了更大的特征扰动，因此速度也得到了快速提升. 由表 1~4 数据可得，BP-JSMA 在 ATC 上比 JSMA 提升了 11 倍左右. 为了更直观感受，绘制如图 9 所示的耗时结果图.

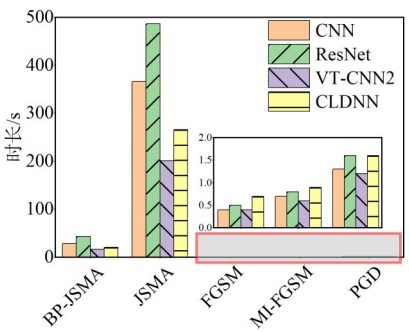


图 9　单个样本生成耗时 ATC

(4) 结构相似度 SSIM

由表 1~4 可知，BP-JSMA 在攻击三种模型时生成的对抗样本与原始样本的结构相似度 SSIM 达到了 90% 以上，相较于 JSMA 提升了大约 10%，而较于 FGSM、MI-FGSM 和 PGD 更是提升超过 20%，这表明了 BP-JSMA 能够生成与原始样本更为相似的对抗样本，其隐蔽性更好. 这是由于 BP-JSMA 通过在选取的显著特征点上添加单点限制，相较于 JSMA 降低了局部感知性，相较于 FGSM 等减少扰动特征点数量，增强了相似性. 直观结果如图 10 所示.

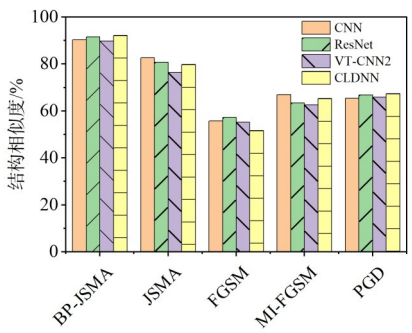


图 10　结构相似度 SSIM

(5) 扰动比率 L_0 和单点最大扰动 L_∞

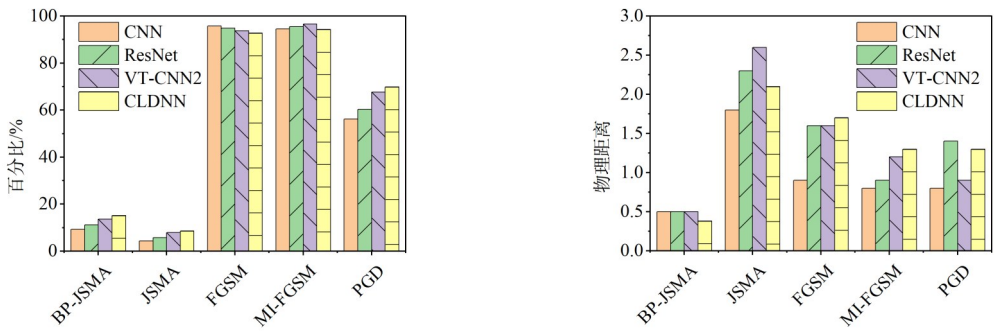
扰动比率 L_0 和单点最大扰动 L_∞ 具体如图 11 所示. 对于 L_0，如图 11(a)所示，BP-JSMA 在 10% 左右，JSMA 低于 10%，FGSM 和 MI-FGSM 均超过了 90%，PGD 在 60% 左右. 这是由于 BP-JSMA 和 JSMA 是特征攻击方法，通过寻找关键特征点进行添加扰动，因此只需改动较少的数据点即可完成攻击，而 BP-JSMA 批量选择特征点，因此该指标略高于 JSMA；而 FGSM、MI-FGSM 和 PGD 属于全局攻击方法，攻击时在全局添加扰动，因此该指标往往较高. 对于 L_∞，如图 11(b)所示，BP-

JSMA 明显优于其他方法，这是由于添加了自适应扰动限制 θ，将对抗样本限制在较小范围，增强了隐蔽性，JSMA 则在某个特征点上添加较大扰动，因此该指标值最高.

由以上分析可知，BP-JSMA 通过单次迭代更多数据点从而加快了生成速度，虽然增加了需要改动的数据点数量，但整体扰动更为细微，隐蔽性更好. 攻击效果如图 12 所示，图 12(a)是原始样本，模型将其正确识别为类别 QPSK，将攻击目标设为 AM-DSB-WC 之后，通过 JSMA 方法生成的对抗样本如图 12(b)所示，而通过本文算法 BP-JSMA 方法生成的对抗样本如图 12(c)所示. 类似的，图 12(d)是原始样本 OOK 且模型正确识别，将攻击目标设为 32APSK 后，通过 JSMA 方法生成对抗样本为图 12(e)，通过 BP-JSMA 生成的对抗样本为图 12(f). 分别对比图 12(b)和(c)、图 12(e)和(f)，可以看到 JSMA 在某些特征点上添加过大的对抗扰动，这虽然降低了改动的特征点数量，但是改动明显，隐蔽性差，而 BP-JSMA 将特征点增加后，降低了扰动的大小，增强了隐蔽性.

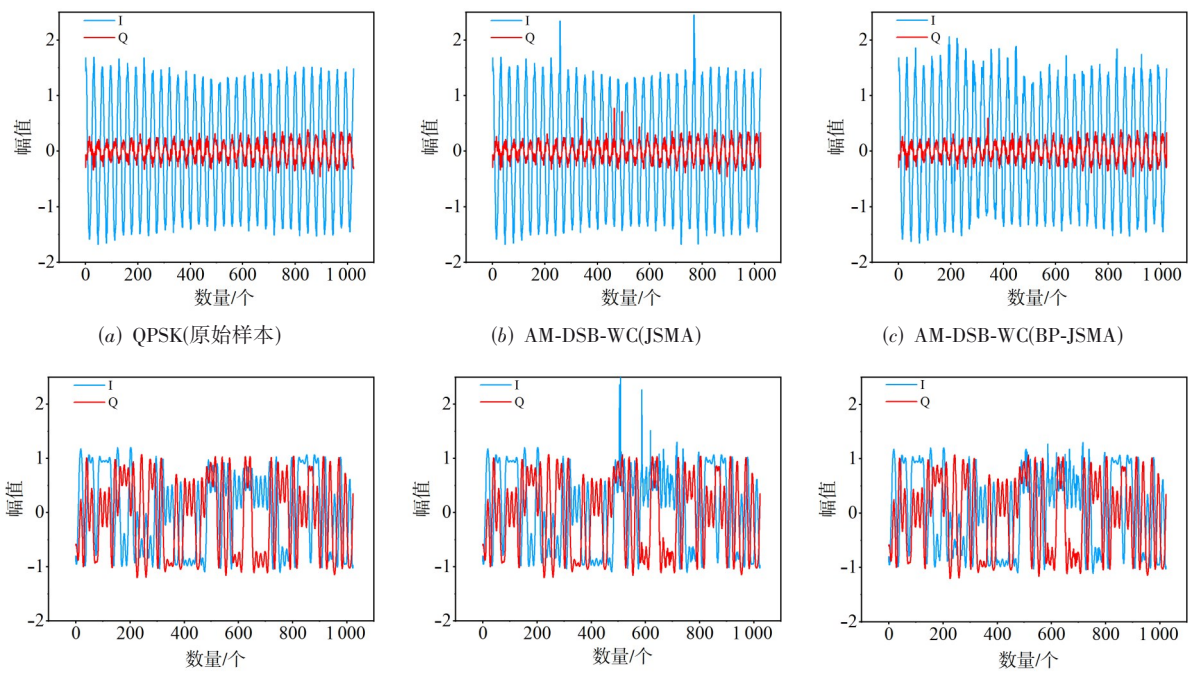
5 结论

为了扩充电磁信号目标攻击研究，本文在传统雅可



(a) L_0距离 (b) L_∞距离

图 11 样本与对抗样本间 L_0 和 L_∞ 距离



(a) QPSK(原始样本) (b) AM-DSB-WC(JSMA) (c) AM-DSB-WC(BP-JSMA)

(d) OOK(原始样本) (e) 32APSK(JSMA) (f) 32APSK(BP-JSMA)

图 12 JSMA 与 BP-JSMA 攻击效果对比

比显著图攻击方法JSMA的基础上对特征点选取方式进行改进,提出了批量显著特征点攻击方法BP-JSMA. 实验表明批量选择特征点的思想使得BP-JSMA能够显著提升对抗样本生成速度,引入的自适应L_∞限制能够有效提升对抗样本隐蔽性. 因此,BP-JSMA能够在加速生成速度的同时保证隐蔽性. 所以除了电磁信号数据外,BP-JSMA适用于其他数据量较多的数据类型,未来可将该方法应用于较大图像等领域的对抗样本研究.

参考文献

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

[2] O'SHEA T J, CORGAN J, CLANCY T C. Convolutional radio modulation recognition networks[M]//Engineering Applications of Neural Networks. Cham: Springer International Publishing, 2016: 213-226.

[3] O'SHEA T J, ROY T, CLANCY T C. Over-the-air deep learning based radio signal classification[J]. IEEE Journal of Selected Topics in Signal Processing, 2018, 12(1): 168-179.

[4] 李钦, 刘伟, 牛朝阳, 等. 低信噪比下基于分裂EfficientNet网络的雷达信号调制方式识别[J]. 电子学报, 2023, 51(3): 675-686.
LI Q, LIU W, NIU C Y, et al. Radar signal modulation recognition based on split EfficientNet under low signal-to-noise ratio[J]. Acta Electronica Sinica, 2023, 51(3): 675-686. (in Chinese)

[5] 周鑫, 何晓新, 郑昌文. 基于图像深度学习的无线电信号识别[J]. 通信学报, 2019, 40(7): 114-125.
ZHOU X, HE X X, ZHENG C W. Radio signal recognition based on image deep learning[J]. Journal on Communications, 2019, 40(7): 114-125. (in Chinese)

[6] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[C]//2014 the 2nd International Conference on Learning Representations. Banff: Conference Track Proceedings, 2014: 1-10.

[7] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[C]//2015 the 3rd International Conference on Learning Representations. San Diego: Conference Track Proceedings, 2015: 1-11.

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

[9] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[C]//2018 the 6th International Conference on Learning Representations. Vancouver: Conference Track Proceedings, 2018: 1-28.

[10] 邹军华, 段晔鑫, 任传伦, 等. 基于噪声初始化、Adam-Nesterov方法和准双曲动量方法的对抗样本生成方法[J]. 电子学报, 2022, 50(1): 207-216.
ZOU J H, DUAN Y X, REN C L, et al. Perturbation initialization, Adam-Nesterov and quasi-hyperbolic momentum for adversarial examples[J]. Acta Electronica Sinica, 2022, 50(1): 207-216. (in Chinese)

[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). Piscataway: IEEE, 2016: 372-387.

[12] COMBEY T, LOISON A, FAUCHER M, et al. Probabilistic Jacobian-based saliency maps attacks[J]. Machine Learning and Knowledge Extraction, 2020, 2(4): 558-578.

[13] LIN Y, ZHAO H J, TU Y, et al. Threats of adversarial attacks in DNN-based modulation recognition[C]//IEEE INFOCOM 2020 - IEEE Conference on Computer Communications. Piscataway: IEEE, 2020: 2469-2478.

[14] FLOWERS B, BUEHRER R M, HEADLEY W C. Evaluating adversarial evasion attacks in the context of wireless communications[J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 1102-1113.

[15] ZHAO H J, LIN Y, GAO S, et al. Evaluating and improving adversarial attacks on DNN-based modulation recognition[C]//GLOBECOM 2020 - 2020 IEEE Global Communications Conference. Piscataway: IEEE, 2021: 1-5.

[16] 王超, 魏祥麟, 田青, 等. 基于特征梯度的调制识别深度网络对抗攻击方法[J]. 计算机科学, 2021, 48(7): 25-32.
WANG C, WEI X L, TIAN Q, et al. Feature gradient-based adversarial attack on modulation recognition-oriented deep neural networks[J]. Computer Science, 2021, 48(7): 25-32. (in Chinese)

[17] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Confer-

ence on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 770-778.

[18] SADEGHI M, LARSSON E G. Adversarial attacks on deep-learning based radio signal classification[J]. IEEE Wireless Communications Letters, 2019, 8(1): 213-216.

[19] SAINATH T N, VINYALS O, SENIOR A, et al. Convolutional, long short-term memory, fully connected deep neural networks[C]//2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2015: 4580-4584.

[20] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

作者简介

周 侠 男,1996年11月出生于贵州省安顺市. 现为武汉数字工程研究所在读硕士. 研究方向为人工智能对抗攻防.
E-mail: zhou_xia1110@163.com

张 剑(通讯作者) 男,1979年1月出生于湖北省宜昌市. 博士毕业于华中科技大学电信系. 现为武汉数字工程研究所博士生导师、研究员,主要研究方向为人工智能、电子战.
E-mail: 1893664@qq.com

李宁安 男,1992年2月出生于河南省开封市. 现为武汉数字工程研究所高级工程师. 主要研究方向为人工智能安全测评.
E-mail: 1912665736@qq.com