基于雅可比显著图的电磁信号快速对抗攻击方法

张剑，周侠，张一然，王梓聪
（武汉数字工程研究所，湖北 武汉 430205）

摘 要：为了生成高质量的电磁信号对抗样本，提出了快速雅可比显著图攻击（FJSMA）方法。FJSMA 通过计算攻击目标类别的雅可比矩阵，并根据该矩阵生成特征显著图，之后迭代选取显著性最强的特征点及其邻域内连续特征点添加扰动，同时引入单点扰动限制，最后生成对抗样本。实验结果表明，与雅可比显著图攻击方法相比，FJSMA 在保持与之相同的高攻击成功率的同时，生成速度提升了约 10 倍，相似度提升了超过 11%；与其他基于梯度的方法相比，攻击成功率提升了超过 20%，相似度提升了 20%～30%。

关键词：深度神经网络；对抗样本；电磁信号调制识别；雅可比显著图；目标攻击
中图分类号：TP183
文献标志码：A
DOI: 10.11959/j.issn.1000−436x.2024021

Electromagnetic signal fast adversarial attack method based on Jacobian saliency map

ZHANG Jian, ZHOU Xia, ZHANG Yiran, WANG Zicong
Wuhan Digital Engineering Institute, Wuhan 430205, China

Abstract: In order to generate high-quality electromagnetic signal countermeasure examples, a fast Jacobian saliency map attack (FJSMA) method was proposed. The Jacobian matrix of the attack target class was calculated and feature saliency maps based on the matrix were generated, then the most salient feature points were iteratively selected and perturbations in their neighborhood were continuously added while introducing a single point perturbation constraint, finally adversarial examples were generated. Experimental results show that, compared with Jacobian saliency map attack method, FJSMA improves the generation speed by about 10 times while maintaining the same high attack success rate, and improves the similarity by more than 11%, and compared with other gradient-based methods, the attack success rate is improved by more than 20%, and the similarity is improved by 20% to 30%.

Keywords: deep neural network, adversarial sample, electromagnetic signal modulation recognition, Jacobian saliency map, target attack

0 引言

深度神经网络（DNN, deep neural network）具有高效率、高准确率等特点，可应用于电磁信号识别领域[1-6]，大幅提升了对电磁信号样本的识别能力，具有重要意义与实用价值。文献[1-3]将电磁信号的同相和正交分量作为模型输入，以调制方式作为输出；文献[4-6]将电磁信号转换为图像，进而利用图像领域识别模型进行识别，以信号名称及时频坐标作为输出。然而，Szegedy 等[7]表明对抗样本的存在会对深度学习模型造成巨大威胁。对抗样本是指向原始样本添加精心设计的扰动，使深度神经网络模型对样本识别错误。利用对抗样本对神经网络模型进行攻击的方式称为对抗攻击。文献[7]表明，生成高质量对抗样本能够用于干扰敌方识别模型，降低其模型识别准确性，使之无法正确识别目标，从而达到保护己方重要

收稿日期：2023−07−26；**修回日期：**2023−10−24
通信作者：周侠，zhou_xia1110@163.com
基金项目：国家自然科学基金资助项目（No.61873040）
Foundation Item: The National Natural Science Foundation of China (No.61873040)

目标的目的。与此同时，文献[8-9]将高质量对抗样本打上正确的类别标签，并添加到正常数据集中对己方模型进行训练，以此提升己方模型的鲁棒性和安全性，从而达到抵御敌方对抗样本攻击的目的。由此可见，研究对抗样本生成方法极具现实意义与实用价值。因此，本文将对如何生成高质量对抗样本进行详细讨论。

目前，应用最广泛的对抗样本生成方法有快速梯度符号方法（FGSM, fast gradient sign method）[10]及其变体[11-14]，这类方法通过在梯度的反方向上添加扰动，进而最大化损失函数以生成对抗样本，其优点在于速度快、易实现，缺点在于其攻击方式为全局攻击，容易被人眼识别。为了减少数据点的改动数量，雅可比显著图攻击（JSMA, Jacobian saliency map attack）方法[15]及其变体[16]应运而生，这类方法属于 L_0（扰动数据点数量占比）限制攻击[16]，即尽可能少地改动数据点。

对抗样本生成方法于 2019 年被引入电磁信号领域[17]，其研究可大致分为两类。第一类是研究人员结合真实物理场景和邻域特有知识，针对性地优化对抗样本生成方法，以增强对抗样本对真实物理场景的适应能力[18-19]。第二类是将现有的图像对抗样本方法应用到电磁信号识别领域[20-25]，证明对抗样本在电磁信号领域中的可行性。但这些方法大多在 FGSM 的基础上进行改进，其攻击目的是使模型决策错误，而使模型将样本识别为指定类别的有目标攻击方法却鲜有涉及。

为了扩充该领域的有目标攻击方法，增强对电磁信号样本的攻击能力，本文将 JSMA 方法的思路应用于电磁信号对抗样本生成，但该方法仍存在以下问题：首先，该方法在每轮选择特征点时需要通过两两组合的方式选取 2 个特征点，因此只适用于小数据样本，对于数据量较多的电磁信号样本来说速度太慢；其次，该方法没有限制样本与对抗样本之间的 L_∞（单点最大扰动）范数[16]，因此在某些特征点添加的扰动很大，隐蔽性差。为了有效解决以上问题，本文提出一种快速雅可比显著图攻击（FJSMA, fast JSMA）方法。

FJSMA 方法充分利用了电磁信号数据连续性强的特点，选择连续特征点进行扰动添加，显著提升对抗样本生成速度；此外，该方法对特征点进行了 L_∞限制，使生成的对抗样本更隐蔽。

本文主要贡献介绍如下。

1) 实验证明了 JSMA 在攻击电磁信号调制识别模型时存在不足，并根据电磁信号数据强连续性的特点，提出了一种适用于电磁信号识别领域的对抗样本生成方法 FJSMA。

2) FJSMA 采用连续选取特征点操作方式，将时间复杂度由 $O(n^2)$降低至 $O(n)$，显著提升对抗样本生成速度。

3) FJSMA 引入了 JSMA 方法中不具备的裁剪函数，在原始的 L_0 限制以外添加了 L_∞ 限制，保证样本的扰动不会超出一定范围，限制了原始样本与对抗样本间的 L_∞ 距离，使生成的对抗样本更隐蔽。

1　相关工作

1.1　对抗样本基本概念

对抗样本生成过程可形式化表示为

$$x^{\text{adv}} = x + \arg\min \| \delta \|_P \tag{1}$$

$$F(x^{\text{adv}}) \neq F(x) \tag{2}$$

其中，x^{adv} 为对抗样本，δ 为对抗扰动，P 为对扰动 δ 的范数约束，0 范数表示约束对抗扰动中扰动数量，2 范数表示约束对抗扰动的模长，∞ 范数表示约束对抗扰动的最大值，F 为深度学习模型，x 为原始图像。约束是为了生成质量更高的对抗样本。

对抗样本生成示意如图 1 所示。图 1(a)表示模型对原始样本分类为 FM，图 1(b)表示在该样本上添加的对抗扰动，图 1(c)表示模型将添加扰动后的样本分类为 GMSK，即对抗样本。

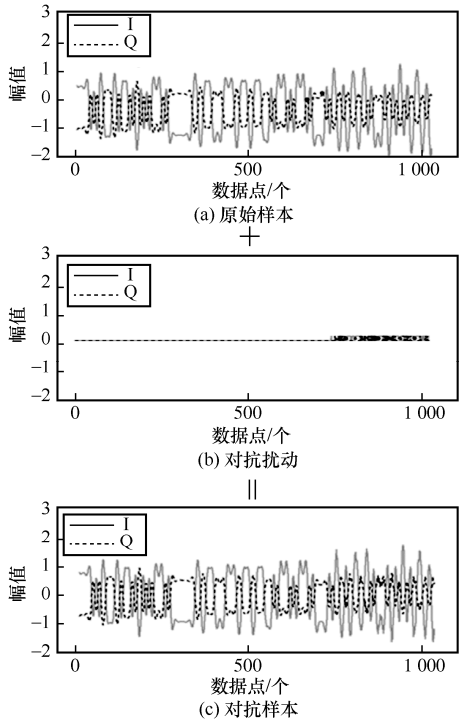


图 1　对抗样本生成示意

1.2 对抗样本生成方法

对抗样本生成方法可根据被攻击环境、攻击者目的、攻击类型、扰动类型等角度进行分类，以下根据攻击类型分别介绍本文实验所用到的几种经典对抗样本生成方法。

1.2.1 快速梯度符号方法

Goodfellow 等[10]提出的 FGSM 通过计算模型损失函数对输入的梯度得到对抗样本，如式(3)所示。

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

其中，x 是原始样本输入，y 是样本的真实标签，x^{adv} 是生成的对抗样本，θ 是深度学习模型的参数，ε 是扰动大小，$L(\cdot)$ 是损失函数，∇_x 是对损失函数求导（即原始样本的梯度信息），$\operatorname{sign}(\cdot)$ 是符号函数，用于标记梯度方向。FGSM 的优点在于速度快，一步就能生成对抗样本，然而也正是由于单步攻击的缘故，生成的扰动幅度大，易被人眼识别。

1.2.2 基本迭代方法

为了克服 FGSM 的缺点，Kurakin 等[11]提出基本迭代方法（BIM, basic iterative method）。BIM 也称为迭代快速梯度符号方法（I-FGSM, iterative fast gradient sign method），其主要思想是将 FGSM 的单步添加扰动改为少量多次，即每次迭代添加少量扰动，观察模型的决策变化，一旦出现错误就返回对抗样本，其过程如式(4)所示。

$$\begin{aligned} x_0^{\text{adv}} &= x \\ x_{i+1}^{\text{adv}} &= \operatorname{Clip}_{x,\varepsilon}\{x_i^{\text{adv}} + \alpha \operatorname{sign}(\nabla_x L(x_i^{\text{adv}}, y; \theta))\} \end{aligned} \tag{4}$$

其中，Clip{}为裁剪函数，保证对抗样本在合法范围内；x_i^{adv} 为第 i 次迭代时的生成的样本（不一定是对抗样本），α 是单步扰动大小，θ 是扰动限制。该方法能够找到更精细的对抗扰动，虽然增加了时间开销，但提高了对抗样本质量。

1.2.3 多样性输入快速梯度符号方法

Xie 等[13]基于数据扩充思想，提出了多样性迭代快速梯度符号方法（DI-FGSM, diversity iterative FGSM），该方法以概率 p 对样本进行随机多样性转换，例如调整大小、裁剪和旋转等操作。然后利用 FGSM 生成对抗样本，其转换过程如式(5)所示。

$$T(x_t^{\text{adv}}, p, s) = \begin{cases} T(x_t^{\text{adv}}, s), \text{概率} p \\ x_t^{\text{adv}}, \text{概率} 1-p \end{cases} \tag{5}$$

其中，p 为对样本进行多样性转换的概率，s 为转换后的样本维度，$T(\cdot)$ 为对样本的随机多样性转化操作。

1.2.4 投影梯度下降方法

基于迭代的思想，Madry 等[25]提出了投影梯度下降对抗（PGD, projected gradient descent）样本生成方法。该方法能够通过投影操作将对抗扰动约束到合法范围，其过程如式(6)所示。

$$\begin{aligned} x_0^{\text{adv}} &= x \\ x_{i+1}^{\text{adv}} &= \prod_{\varepsilon}\{x_i^{\text{adv}} + \alpha \operatorname{sign}(\nabla_x L(x_i^{\text{adv}}, y; \theta))\} \end{aligned} \tag{6}$$

其中，$\prod_{\varepsilon}\{\cdot\}$ 是投影操作，可将添加扰动后的对抗样本投影到合法范围。

1.2.5 显著图方法

显著图方法本质上是找出样本中的重要特征部分并对其添加扰动。显著图方法一般分为模型解释的方法和显著目标检测的方法。

模型解释的方法可以将分类结果通过反向传播算法逐层传递到输入层，此类方法包括 JSMA 方法以及 Grad-CAM 方法[26]，具体介绍如下。

FGSM 及其变体攻击方法大多为非目标攻击。为了完成目标攻击，同时减少改动数据点的数量，Papernot 等[15]提出了 JSMA 方法。该方法利用雅可比矩阵反映输入对输出的影响的特性，通过目标类别 t 计算雅可比矩阵生成特征显著图，然后从特征显著图中选取关键特征点对添加扰动生成对抗样本。其关键特征点对选取方式如式(7)所示。

$$\arg\max_{(p_1,p_2)} \left(\sum_{i=p_1,p_2} \frac{\partial F_t(x)}{\partial x_i} \right) \left| \sum_{i=p_1,p_2} \sum_{j \neq t} \frac{\partial F_t(x)}{\partial x_i} \right| \tag{7}$$

其中，$F_t(x)$ 为 x 在类别 t 上的分类得分，i 为数据点，p_1和p_2 为根据特征所选取的特征点对。

周侠等[26]将 Grad-CAM 方法应用于电磁信号领域。首先，反向计算最高层特征图的权重信息；然后，根据该权重值对特征图进行加权叠加，经过 ReLU 激活得到显著特征图；最后，设定一个参数作为基准对特征图进行二值化处理，对处理后的非 0 数据点添加扰动生成对抗样本。

显著目标检测的方法只需要输入原始图像，不需要攻击模型的梯度信息，此类方法包括 MA-NA-FGSM[27]。

李哲铭等[27]提出的 MA-NA-FGSM 通过对显著特征区域迭代添加扰动生成对抗样本。具体步骤为

在图像的底层细节与高层语义中提取不均匀的特征，并在空间域和通道域中分配特征的自适应权重，生成像素值为 0～255 的灰度图，其中图像的主体部分（灰度图中白色部分），即所需要迭代添加扰动的显著特征区域。

2　FJSMA 方法

为了解决 JSMA 方法用于电磁信号对抗样本生成存在的问题，本文提出 FJSMA 方法，其具体攻击流程描述如下。

首先，对于一个已知训练好的电磁信号深度学习模型 $F(x):x\in X\to y\in Y$，获取神经网络模型 F 的参数信息（包括层数及权重等）。其次，将原始电磁信号样本 x 输入模型 F，获取模型对 x 的各类预测值，根据获取的每一类的预测值分别对 x 求雅可比矩阵。再次，选择原始样本 x 的攻击目标类别 t，根据攻击类别 t 结合雅可比矩阵生成攻击类别 t 的特征显著图。最后，根据类别 t 的特征显著图选取显著性最强的特征点及其邻域内的连续特征点进行扰动添加得到 x^{adv}，如果该样本使模型 F 将样本 x^{adv} 识别为类别 t，则生成对抗样本成功，返回对抗样本；否则，对特征点进行迭代扰动添加，每次判断单点扰动是否达到设置的扰动上限 ε，若达到上限，则选取下一个显著特征点及其邻域内的特征点，并重复上述操作直至对抗样本生成成功。如果显著图中的可用特征点在用完之前能够使模型 F 将其识别为类别 t 的对抗样本 x^{adv}，则成功生成对抗样本，反之生成对抗样本失败。该攻击方法流程如图 2 所示。

攻击流程的主要步骤如下。1)计算雅可比矩阵，雅可比矩阵是生成显著图的基础。根据输入 x 在模型 F 上的得分，前向求导得到雅可比矩阵 $\boldsymbol{J}(F(x))$。2)生成显著图，显著图展示了每个特征点的显著性。根据 $\boldsymbol{J}(F(x))$和攻击目标 t 求解生成显著图 $S(x,t)$。3)生成对抗样本，根据 $S(x,t)$选择显著特征点添加扰动 δ，生成使 $F(x^{\mathrm{adv}})=t$ 的对抗样本 x^{adv}。下面对这 3 个步骤进行详细介绍。

2.1　计算雅可比矩阵

雅可比矩阵反映了函数计算中输入特征点对输出结果的影响，而深度神经网络可以视为一种复杂的函数计算。将函数输出值对输入进行前向求导即可得到雅可比矩阵。

设模型 F 是一个 n 分类模型，其输入 x 的维度为 $a\times b$。将 F 对 x 关于每一类别的得分值作为计算起点，根据式(8)计算 x 对于 n 个类别的雅可比矩阵。

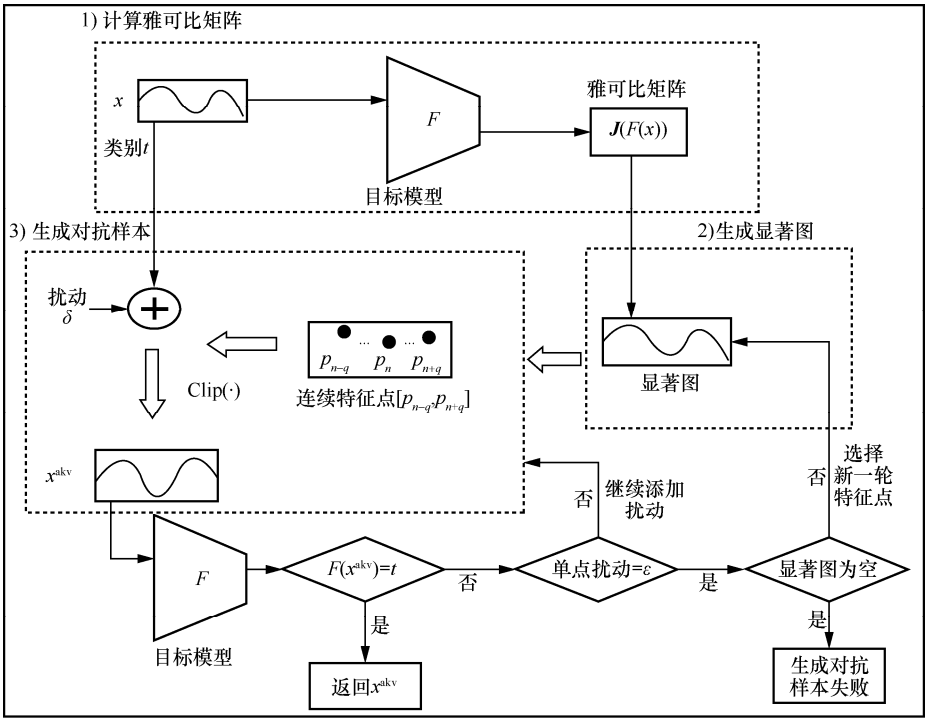


图 2　FJSMA 流程

$$\boldsymbol{J}(F_r(x))=\frac{\partial F_r(x)}{\partial x}=\left[\frac{\partial F_r(x)}{\partial x_i}\right]_{i\in[1,a\times b]} \tag{8}$$

其中，$F_r(x)$ 表示输入 x 在类别 r 的得分，x_i 表示 x 的第 i 个特征点。$\boldsymbol{J}(F_r(x))$ 的计算结果是一个与 x 同维度的矩阵，其各个位置的值表示该位置数据点对 $F_r(x)$ 的贡献度，该值越大，说明该点对于将 x 分类为类别 r 的贡献就越大，反之越小。

由于深度神经网络 F 是一个复杂的函数计算，输入与输出之间存在许多隐藏层。因此，在计算分类得分对于输入的雅可比矩阵时，需要使用式(9)和式(10)进行链式计算。

$$\frac{\partial F_r(x)}{\partial x}=\frac{\partial F_r(x)}{\partial H(x)}\frac{\partial H(x)}{\partial x} \tag{9}$$

$$\frac{\partial H_k(x)}{\partial x_i}=\left[\frac{\partial f_{k,e}(W_{k,e}H_{k-1}+b_{k,e})}{\partial x_i}\right]_{e\in[1,m_k]} \tag{10}$$

其中，$H_k(x)$ 是第 k 层隐藏层，e 是第 k 层的一个神经元，$f_{k,e}$ 是神经元 e 的激活函数，$W_{k,e}$ 是 e 的权重，$b_{k,e}$ 是 e 的偏置值，m_k 是第 k 层神经元个数。

2.2 生成显著图

FJSMA 作为一种目标攻击方法，在攻击时需要指定攻击目标 t。JSMA 生成显著图时分成了扰动添加和扰动减少 2 个方向，FJSMA 将两者进行结合。根据 2.1 节计算得到对于输入 x 的雅可比矩阵 $\boldsymbol{J}(F(x))$，然后通过式(11)计算得到使 x 分类为 t 的关键特征点，最终生成 x 对于 t 的特征显著图 $S(x,t)$。

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

其中，i 表示输入 x 的第 i 个特征点，$J_{it}(x)$ 表示 x 在类别 t 的得分对第 i 个特征点的雅可比值。$J_{it}(x)>0$ 表示特征点 i 对于将 x 分类为 t 的贡献为正，即增加 i 的值会使 x 在类别 t 的得分增加；$\sum_{r\neq t}J_{ir}(x)<0$ 表示 i 对其他类别的总贡献为负，即增加 i 的值会导致其他类别的得分减少。$J_{it}(x)>0$ 且 $\sum_{r\neq t}J_{ir}(x)<0$ 表示该点对分类为 t 的贡献为正；$J_{it}(x)<0$ 且 $\sum_{r\neq t}J_{ir}(x)>0$ 表明 i 值增加会导致 t 的得分下降。

为了促进将 x 分类为 t，对于 $J_{it}(x)>0$ 且 $\sum_{r\neq t}J_{ir}(x)<0$ 的特征点，其扰动值添加方向为正；对于 $J_{it}(x)<0$ 且 $\sum_{r\neq t}J_{ir}(x)>0$ 的特征点，其扰动值添加方向为负。通过对 x 中每个特征点进行如上计算，即可得到特征显著图 $S(x,t)$。

2.3 生成对抗样本

由 2.2 节得到类别 t 的特征显著图 $S(x,t)$，然后循环选择显著性最高的特征点添加扰动，当 $F(x^{\text{adv}})=t$ 时，返回对抗样本。

为了提升对抗样本生成速度，FJSMA 充分利用电磁信号数据连续性特点，对特征点的选取方式进行改进。

首先选择 $S(x,t)$ 中特征值最大的数据点（如式(12)所示），然后选择该点邻域内的数据点添加扰动（如式(13)所示）。假设 $S(x,t)$ 中最大显著性特征点为 q，则在每次迭代中选择点 q 邻域 $[q-p,q+p]$ 中的特征点进行扰动添加。实验中 $p=5$，$[q-p,q+p]\in[0,1\,023]$。

$$\left[x_i^{\text{adv}}=x_i+\theta\,\text{sign}(S(x,t)[i])\right]_{i=q} \tag{12}$$

$$\left[x_i^{\text{adv}}=x_i+\frac{1}{5}\theta\,\text{sign}(S(x,t)[i])\right]_{i\in[q-p)\cup(q+p]} \tag{13}$$

其中，$\text{sign}(\cdot)$ 表示对选取的一段特征点进行符号运算，i 为所选取的特征点。添加扰动过程如式(12)所示，首先对特征点 q 添加单步扰动 θ，然后对其邻域内的特征点添加如式(13)所示的扰动，除了特征点 q，其邻域中特征点添加 $\left(\frac{1}{5}\right)\theta$ 的扰动值，这样既能够快速添加扰动，同时也能降低邻域内其他特征点的显著性，保证特征点 q 的显著性。

$$x^{\text{adv}}=x+x_i^{\text{adv}} \tag{14}$$

如式(14)所示，将添加了扰动的特征点与原始样本 x 相加，得到对抗样本 x^{adv}。如果 $F(x^{\text{adv}})\neq t$，则选取下一个显著特征点及其邻域添加扰动，直到 $F(x^{\text{adv}})=t$，返回对抗样本 x^{adv}。

此外，JSMA 属于 L_0 限制，导致在特征点上添加的扰动过大，易被人眼识别。因此，FJSMA 如式(15)所示，引入 L_∞ 限制 ε，即每个显著性特征点添加的扰动最大值不超过 ε。

$$\left\| x_i^{\text{adv}} - x_i \right\|_\infty \leqslant \varepsilon \tag{15}$$

FJSMA 方法流程如算法 1 所示。

算法 1　FJSMA 方法

输入　干净信号样本 x，标签 y_{true}，n 分类的深度学习模型 F，攻击目标类别 $t\left(t \neq y_{\text{true}}\right)$，单步扰动大小 θ，单个扰动限制 ε，总扰动限制 γ

输出　使 $F(x^{\text{adv}}) = t$ 的对抗样本 x^{adv}

1) 将样本 x 输入模型 F，返回类别 r 的 $F_r(x)(r \in [1,n])$；
2) 根据式(8)计算类别 r 的雅可比矩阵 $\boldsymbol{J}(F_r(x))$；
3) 通过式(11)，根据 $\boldsymbol{J}(F_r(x))$，生成目标类别 t 对于 x 的显著图 $S(x,t)$；
4) while 显著图 $S(x,t)$ 不为空
5) for $i = 0, i < m, i{+}{+}$
6) 根据式(12)和式(13)从 $S(x,t)$ 中选择显著特征点 q 及其邻域 $[q-p, q+p]$ 中的特征点添加扰动 θ
7) if $F(x^{\text{adv}}) = t$
8) return 对抗样本 x^{adv}
9) else
10) if 单点扰动小于 ε and 总扰动小于 γ
11) continue
12) else
13) 生成对抗样本失败
14) break for
15) end if
16) end if
17) end for
18) end while

3　实验和结果分析

3.1　实验设置

3.1.1　数据集

本文实验在数据集 2018.01.OSC[2]上进行。该数据集是电磁信号领域十分完善的数据集，已被众多研究者[1-6]使用并进行实验。该数据集在多个网络模型下都能训练出较高的识别正确率，符合实验需要高正确率模型的需求，数据集本身完整公开，不存在数据无效或数据缺失等现象，不需要进行数据清洗操作，且其本身的安全性问题也能够得到保障。数据集共有 2 555 904 条数据，包含 24 种调制类型。每种调制类型包含 26 种信噪比，从−20 dB 到 30 dB，步长为 2 dB。每种信噪比下有 4 096 条数据。每条数据有 I/Q 两路信号，每路信号有 1 024 个数据点。具体 24 种调制方式分别为 32PSK，16APSK，32QAM，FM，GMSK，32APSK，OQPSK，8ASK，BPSK，8PSK，AM-SSB-SC，4ASK，16PSK，64APSK，128QAM，128APSK，AM-DSB-SC，AM-SSB-WC，64QAM，QPSK，256QAM，AM-DSB-WC，OOK，16QAM。

部分样本可视化如图 3 所示。其中，图 3(a)是调制方式为 32APSK 在信噪比为−16 dB 下的一个数据样本；图 3(b)的调制方式为 OOK，信噪比为 10 dB；图 3(c)的调制方式为 4ASK，信噪比为 30 dB。

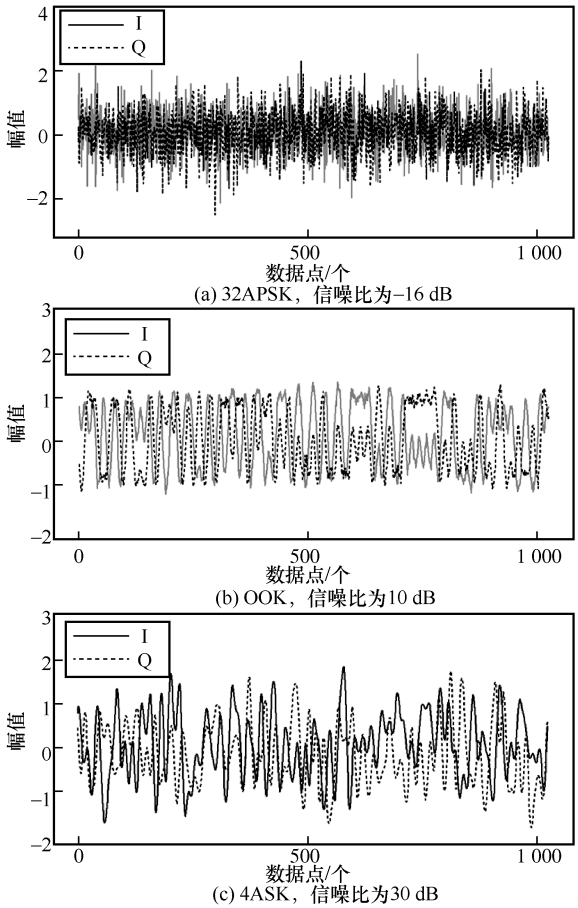


图 3　部分样本可视化

3.1.2　网络模型

O'Shea 在构建数据集 2018.01.OSC[2]时设计了针对该数据集的 ResNet 和 CNN 模型。其中，ResNet 模型由输入层、6 个残差块、2 个全连接

层和输出层组成，如图 4(a)所示。其中，输入层大小为 2×1 024，每个残差块包含一个 1×1 的线性卷积层、2 个残差单元和一个最大池化层，模型中所有卷积层的卷积核数量均为 32，卷积核大小为 3×1。CNN 模型由输入层、7 个卷积和最大池化层、2 个全连接层组成，输入层大小为 2×1 024，卷积层中卷积核大小为 3×3，步长与填充均设为 1，其余参数如图 4(b)所示。

此外，针对 O'Shea 在文献[1]中提出的数据集，Xu 等[28]提出多通道卷积长短期深度神经网络（MCLDNN, multi-channel convolutional long short-term deep neural network）框架，该框架集成了一维（1D）卷积、二维（2D）卷积和长短期记忆（LSTM）层，从时间和空间角度更有效地提取特征，本文将其改造以适配本文数据集 2018.01.OSC，改变输入层大小为 2×1 024，此外，还包含 2 个 1D 卷积层（Conv2、Conv3）和 3 个 2D 卷积层（Conv1、Conv4、Conv5）、2 个具有 128 个单元的 LSTM 层以及 2 个具有 128 个神经元的全连接层和输出层，结构如图 4(c)所示。

3.1.3　评价指标

为了对 FJSMA 进行有效评估，本文从攻击有效性、攻击效率和对抗样本隐蔽性 3 个方面引入了如下指标。

1) 攻击有效性

攻击有效性包括攻击成功率（ASR, attack success rate）、对抗类别平均置信度（ACAC, average confidence of adversarial class）以及真实类别平均置信度（ACTC, average confidence of true class）。

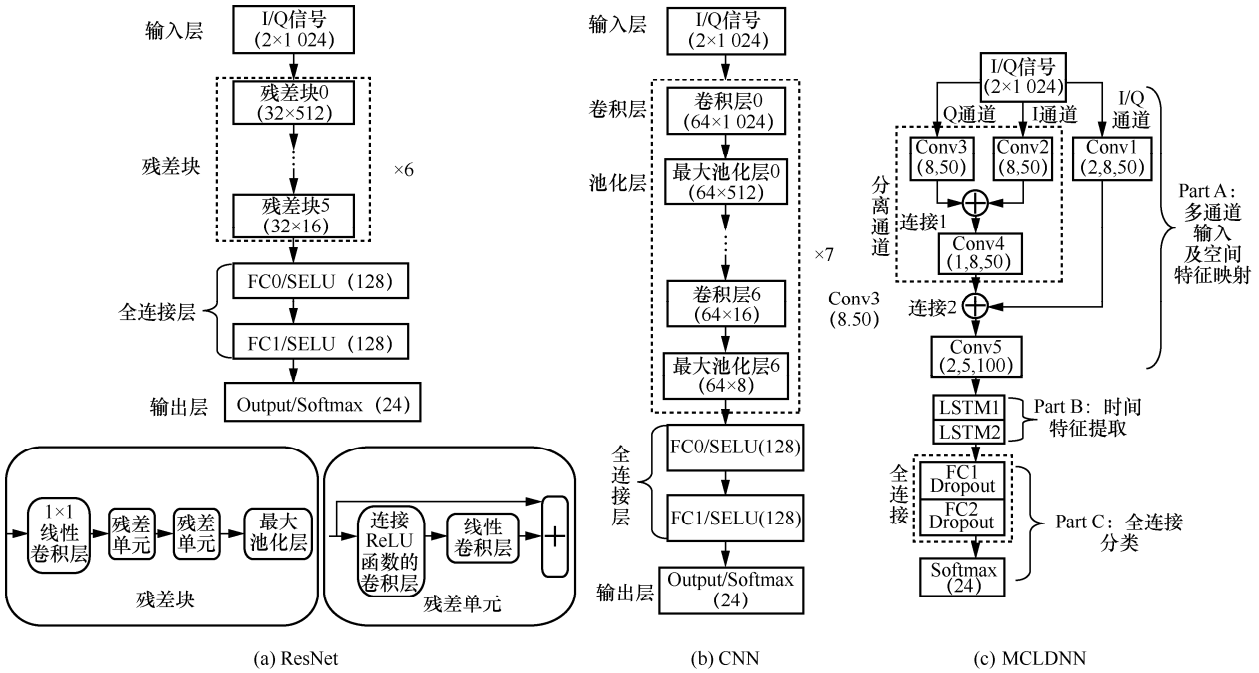
攻击成功率表示给定一批样本，其中攻击成功的样本数占总样本数的比率，该值在 0 到 1 之间，越接近 1 说明成功率越高。设总测试样本数为 M，成功攻击模型的样本数为 N，则 ASR 为

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

对抗类别平均置信度表示模型将给定样本识别为攻击目标类别的置信度，该值在 0 到 1 之间，越接近 1 说明模型将该样本识别为目标类别的置信度越高，攻击也就越有效。假设单个样本的攻击类别置信度为 a，总样本数为 M，则 ACAC 为

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

真实类别平均置信度表示模型将给定样本识别为正确类别的置信度，该值在 0 到 1 之间，越接近 1 说明模型将该样本识别为正确类别的置信度越高，攻击也就越弱。假设模型对单个样本的正确类别的置信度为 b，总样本数为 M，则 ACTC 为



(a) ResNet　　(b) CNN　　(c) MCLDNN

图 4　网络结构

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

2) 攻击效率

攻击效率由生成对抗样本平均耗时（ATS, average time spend）反映。ATS 是一种统计指标值，具体为给定一批样本，统计这批样本生成对抗样本所需的平均时间。假设成功攻击模型的样本数为 N，总耗费时长为 T，则 ATS 为

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

3) 对抗样本隐蔽性

对抗样本隐蔽性可由对抗样本与原始样本之间的差异度表示，因此引入以下 4 个指标。

结构相似度（SSIM, structural similarity），常用于刻画图像之间的相似差异，而本文实验所用的电磁信号数据可视为单通道图像。该指标在 0 到 1 之间，越接近 1 说明对抗样本与原始样本更相似。

扰动数据点数量占比（L_0），表示对抗样本中改动的数据点数量占总数据点数量的比率。该值在 0 到 1 之间，越大说明改动的数据点数量越多，越容易破坏隐蔽性。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，则扰动数据点数量占比的计算式为

$$L_0=\frac{P}{nm}\times 100\% \tag{20}$$

欧氏距离（L_2），表示对抗样本与原始样本之间的欧氏距离，该值越小说明两者越相似。设原始样本为 x，对抗样本为 x^{adv}，数据点为 i，则欧氏距离的计算式为

$$L_2=\mathrm{dist}(x,x^{\mathrm{adv}})=\sqrt{\sum_{i=1}^{n}(x_i-x_i^{\mathrm{adv}})^2} \tag{21}$$

单点最大扰动（L_∞），表示对抗样本中数据点改动的最大值，该值越大则越可能被人眼识别。

3.2　结果分析

本次实验使用的显卡为 GeForce RTX 3090，深度学习环境为华为昇思 MindSpore1.9.0，对抗样本库为 MindArmour1.9，GPU 版本为 CUDA11.1。

3.2.1　网络训练设置和结果

训练时，将数据集进行分层抽样，即每种类别中的每种信噪比的 70%作为训练样本，30%作为测试样本。对于学习率 l_{r}，3 个模型设置相同，初始 l_{r}=0.01，每训练迭代一轮 l_{r} 降为原来的一半，这使前期训练能够快速收敛，而训练到后期会更精确。3 个模型的准确率如图 5 所示。针对 ResNet 模型，为提升训练速度，除了残差块 0 之外，其余残差块的卷积核从 3×2 改为 3×1，这使模型参数量由原来的 23.6 万个降到了 14.1 万个，当信噪比大于 10 dB 时，平均准确率达到了 96.2%。对于 CNN 模型，当 SNR 大于 10 dB 时，平均正确率达到了 88.6%。对于 MCLDNN 模型，当信噪比大于 4 dB 时，正确率达到了 97.3%。由实验结果图 5 可知，信噪比太低的数据本身就难以被模型正确识别，对其进行攻击没有实际意义。信噪比大于 10 dB 的样本在各网络模型上被识别的正确率皆能达到 80%，因此本文要进行数据预处理工作，挑选出信噪比大于 10 dB 的数据进行实验，该部分数据能够更显著地反映对抗样本的实际攻击能力。这证明模型是被对抗样本干扰判断出现错误，而不是数据本身噪点过多而导致识别错误。

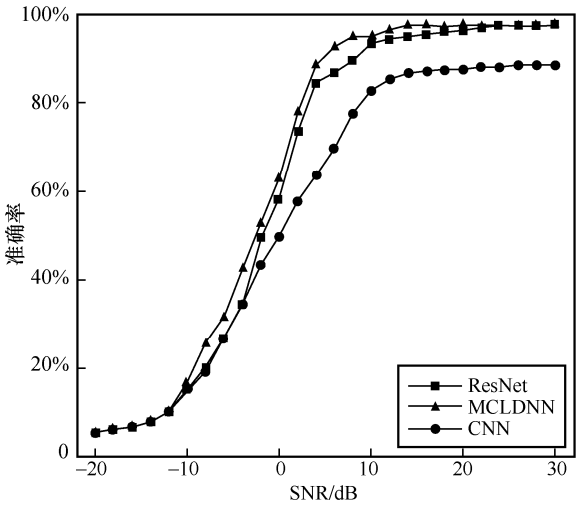


图 5　3 个模型的准确率

3.2.2　FJSMA 效果

FJSMA 与 JSMA 效果对比如图 6 所示，其中图 6(a)为原始样本（被识别为正确标签 OOK），此时攻击目标标签为 32APSK，图 6(b)为 JSMA 生成的对抗样本，图 6(c)为 FJSMA 在扰动限制 $\varepsilon=0.2$ 时生成的对抗样本。对比可以看出，FJSMA 方法生成的对抗样本更加隐蔽。

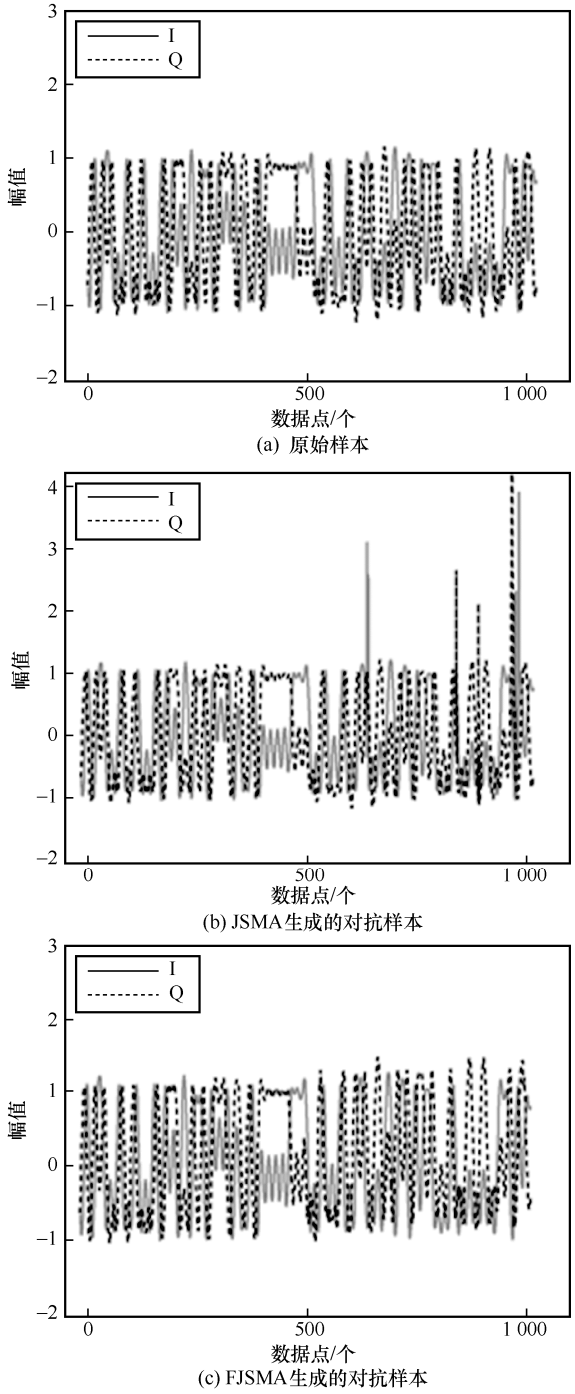


图 6 JSMA 与 FJSMA 效果对比

为了更好地对 FJSMA 进行评价，选取 JSMA[15]、DI-FGSM[13]、PGD[22]、Grad-CAM[26]和 MA-NA-FGSM[27] 攻击方法作为对比，其中 MA-NA-FGSM 由图像领域的对抗样本方法改动而来。由于 FJSMA 为目标攻击，因此 DI-FGSM 和 PGD 也设置为目标攻击模式。此外，本文还研究了 ε 的影响，分别将其设置为 0.2、0.4 和 0.6，并观察攻击效果。ResNet 模型、CNN 模型、MCLDNN 模型受攻击结果分别如表 1～表 3 所示。

1) 攻击有效性

由表 1～表 3 可知，FJSMA 保持与 JSMA、MA-NA-FGSM 同水平的攻击成功率，均大于 95%，强于 DI-FGSM、PGD 和 Grad-CAM。为了直观了解对抗攻击对模型的影响，绘制如图 7 所示的模型正确率。模型正确率从侧面反映了攻击成功率，模型正确率越高则攻击成功率越低，反之亦然。由图 7 可知，攻击之前模型具有较好的识别准确率，但是在对抗样本下其决策效果大打折扣，由此可知深度学习模型容易受到对抗样本的影响，同时，本文提出的 FJSMA 继承了目标攻击强攻击性的优点。

由表 1～表 3 可知，对抗样本能够使模型以较大置信度识别为指定的类别，其中，FJSMA 达到了 70%左右，略低于 JSMA 和 MA-NA-FGSM，但高于 DI-FGSM、PGD 和 Grad-CAM。此外，模型在对抗样本正确类别的置信度急剧下降，在 FJSMA 与 JSMA 的攻击下仅有不到 0.2%的正确类别置信度，而在 DI-FGSM、PGD、MA-NA-FGSM 和 Grad-CAM 上也不到 10%，由此可知对抗样本具有很强的诱导性使模型决策错误。

由以上分析可知，FJSMA 能够保持与 JSMA 同水平的攻击有效性，且强于其他 4 种方法。

2) 攻击效率

单个对抗样本生成耗时对比如图 8 所示。由图 8 可知，JSMA 具有很高的时间复杂度，其生成对抗样本的平均耗时在 ResNet 模型上为 611.22 s，在 CNN 模型上为 432.08 s，在 MCLDNN 模型上为 488.36 s。而 FJSMA 将该指标大幅降低，在 ResNet 模型上为 63.94 s，在 CNN 模型上为 40.51 s，在 MCLDNN 模型上为 47.29 s，生成速度提升了约 10 倍。这是因为 JSMA 在选择特征点时采用的是两点选择策略，如式(7)所示选择最显著的特征点对，然而特征点对的组合方式过多，由式(7)可知，JSMA 在选择特征点对的时间复杂度为 $O(n^2)$，而 FJSMA 直接选择最显著的单个特征点，再选择其前后邻域内的特征点，该方式避免了式(7)的计算，时间复杂度降为 $O(n)$，所以 FJSMA 能够将速度提升一个数量级左右。其余方法都不需要进行单点迭代选择扰动，因此整体速度较快，但代价是改动了很多影响较小的数据点，导致隐蔽性升高。

表 1　　ResNet 模型受攻击结果

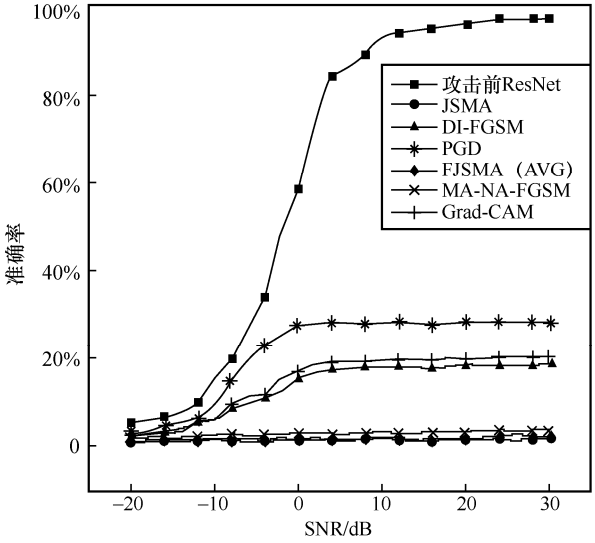
攻击方法	ASR	ACAC	ACTC	ATS/s	SSIM	L_0	L_2	L_∞
JSMA	97.76%	76.33%	0.003%	611.22	80.79%	3.61%	0.11	3.08
DI-FGSM	84.75%	65.71%	7.331%	0.58	53.66%	75.00%	0.68	2.26
PGD	72.58%	53.82%	4.884%	0.63	67.39%	93.84%	0.45	2.18
FJSMA($\varepsilon = 0.2$)	96.91%	66.96%	0.097%	78.44	93.93%	26.81%	0.09	0.19
FJSMA($\varepsilon = 0.4$)	97.33%	69.27%	0.051%	63.18	92.92%	18.94%	0.12	0.39
FJSMA($\varepsilon = 0.6$)	97.64%	74.85%	0.013%	50.21	90.25%	15.64%	0.18	0.56
FJSMA(AVG)	97.29%	70.51%	0.054%	63.94	92.37%	20.46%	0.13	0.38
MA-NA-FGSM	95.20%	72.33%	1.94%	1.76	76.8%	87.21%	0.49	1.04
Grad-CAM	80.91%	65.26%	4.791%	11.46	85.41%	24.33%	0.31	1.74

表 2　　CNN 模型受攻击结果

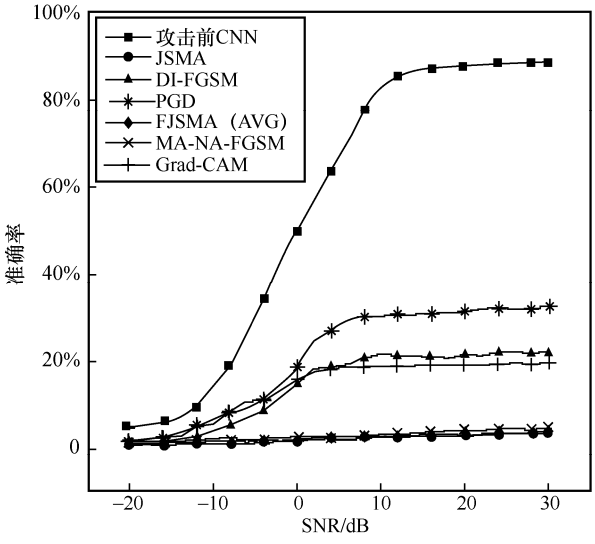
攻击方法	ASR	ACAC	ACTC	ATS/s	SSIM	L_0	L_2	L_∞
JSMA	97.12%	73.03%	0.128%	432.08	76.48%	8.07%	0.13	3.11
DI-FGSM	83.88%	66.81%	2.146%	0.42	56.71%	88.96%	0.73	2.69
PGD	74.12%	49.85%	17.251%	0.36	65.46%	100.00%	0.65	2.13
FJSMA($\varepsilon = 0.2$)	94.55%	64.91%	0.107%	46.38	92.38%	28.26%	0.11	0.18
FJSMA($\varepsilon = 0.4$)	95.48%	70.33%	0.081%	41.25	90.31%	18.93%	0.16	0.37
FJSMA($\varepsilon = 0.6$)	96.35%	75.82%	0.034%	33.89	87.90%	10.58%	0.21	0.58
FJSMA(AVG)	95.46%	70.35%	0.074%	40.51	90.20%	19.26%	0.16	0.38
MA-NA-FGSM	94.36%	73.21%	1.88%	1.55	77.31%	88.90%	0.53	1.13
Grad-CAM	81.42%	66.34%	4.63%	10.98	86.37%	25.61%	0.29	1.58

表 3　　MCLDNN 模型受攻击结果

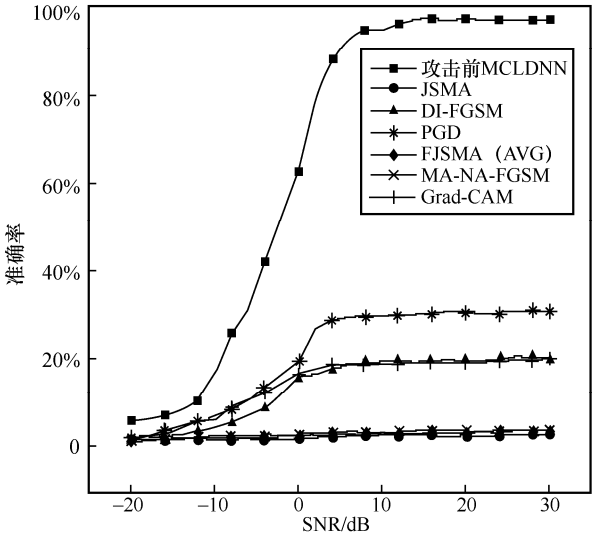
攻击方法	ASR	ACAC	ACTC	ATS/s	SSIM	L_0	L_2	L_∞
JSMA	97.33%	74.83%	0.038%	488.36	78.23%	6.59%	0.13	2.79
DI-FGSM	83.97%	62.18%	6.012%	0.56	54.62%	87.89%	0.77	2.12
PGD	71.44%	52.91%	7.297%	0.73	64.93%	92.47%	0.61	2.36
FJSMA($\varepsilon = 0.2$)	96.35%	67.48%	0.307%	53.93	91.87%	27.33%	0.13	0.19
FJSMA($\varepsilon = 0.4$)	96.12%	70.35%	0.196%	47.29	88.69%	19.26%	0.19	0.38
FJSMA($\varepsilon = 0.6$)	97.24%	73.67%	0.103%	40.66	86.31%	13.95%	0.25	0.53
FJSMA(AVG)	96.57%	69.55%	0.202%	47.29	88.96%	20.18%	0.19	0.37
MA-NA-FGSM	94.97%	72.95%	1.76%	1.58	74.29%	87.13%	0.47	0.98
Grad-CAM	80.57%	64.47%	5.81%	11.23	85.92%	25.82%	0.34	2.01



(a) ResNet



(b) CNN



(c) MCLDNN

图 7　攻击前后模型的准确率

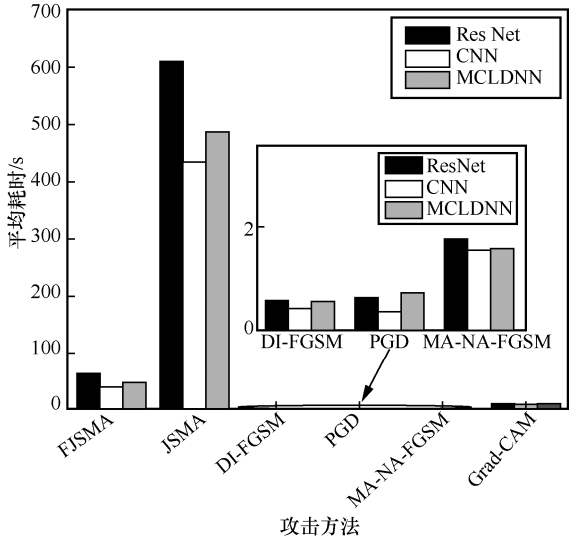


图 8　单个对抗样本生成耗时对比

3) 对抗样本隐蔽性

由表 1～表 3 可知，FJSMA 生成的对抗样本与原始样本的结构相似度达到了 90%左右，和最新的对抗样本生成方法 MA-NA-FGSM 以及 Grad-CAM 处于同一水平，相较于 JSMA 提升了 11%，相较于 DI-FGSM 和 PGD 更是提升了超过 20%。为直观感受相似度变化，绘制如图 9 所示的相似度柱状图。

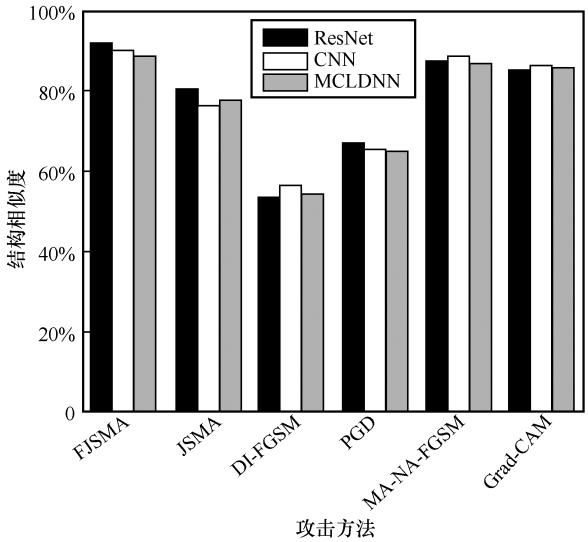


图 9　相似度柱状图

由表 1～表 3 可知，扰动比率 L_0 在 JSMA 上最小，平均不超过 10%；FJSMA 与 Grad-CAM 略高，达到了 20%左右；DI-FGSM、PGD 和 MA-NA- FGSM 最高，平均达到了 90%。出现该现象的原因是 JSMA 是特征扰动方法，其目的是改动较少的数据点完成攻击，因此扰动比率最

小；FJSMA 由于添加了单点扰动限制，不得不增加选择特征点的数量，因此较 JSMA 有所增加；而 DI-FGSM、PGD 和 MA-NA-FGSM 属于全局攻击方法，在添加扰动时是整体更新，因此改动的特征点数量最多。

由于全局攻击往往较特征攻击添加更多的扰动值，因此 DI-FGSM、PGD 和 MA-NA-FGSM 在 L_2 上的值高于 JSMA 和 FJSMA；而 JSMA 由于改动的特征点最少，因此平均至所有数据点上的欧氏距离也相应最小，但 FJSMA 也保持了与之相同的水平。

由于添加了单点扰动限制，因此 FJSMA 在 L_∞ 上最小，而 JSMA 由于改动较少的数据点，在某些显著特征点上添加了较大扰动值，因此 JSMA 在该指标上最大；DI-FGSM 和 PGD 则通过计算特征点的梯度信息添加扰动，也会使较大梯度的特征点添加较大的扰动值。

由以上分析可知，FJSMA 添加了单点扰动限制，虽然增加了改动特征点的数量，但是有效提升了结构相似度，降低了欧氏距离和单点扰动幅度，增强了对抗样本的隐蔽性。

由表 1～表 3 可知，当单点扰动限制 ε 的值由 0.2 增加至 0.6 之后，SSIM 会有所降低，在 ResNet 上由 93.93%降低至 90.25%，但平均耗时也由原来的 78.44 s 降低至 50.21 s，另外 2 个模型的表现与之类似。由此可知，随着扰动限制的增加，会降低对抗样本与原始样本的结构相似度，但速度会有所提升。因此，当侧重点在速度时，ε 的值可以适当增大；当侧重点在隐蔽性时，ε 的值可以适当减小。不同限制 ε 下生成的对抗样本如图 10 所示，其中，图 10(a)是标签为 8PSK 的原始样本，图 10(b)～图 10(d)分别是目标标签为 4ASK 时，ε 取值为 0.2、0.4 和 0.6 时生成的对抗样本。从图 10 可以看到，随着 ε 值的增加，添加的对抗扰动越明显。由以上分析可知，FJSMA 在各方面均有不同的效果提升。从攻击有效性和攻击效率来看，该方法在保证攻击成功率的基础上，比 JSMA 提升了约 10 倍的对抗样本生成速度，比 DI-FGSM 等提升了超过 20%的攻击成功率。从对抗样本隐蔽性来看，FJSMA 的 SSIM 比同类型的 JSMA 提升了 11%，比 DI-FGSM 等提升了 20%。此外，FJSMA 对抗样本的扰动比率 L_0 也能维持在 10%～30%，且因设置了最大扰动限制，FJSMA 的 L_∞ 指标也最小。与最新的对抗样本生成方法 MA-NA-FGSM 与 Grad-CAM 相比，FJSMA 部分指标也存在优势。总体来说，FJSMA 能在保证隐蔽性的前提下提升攻击效率与攻击成功率，具有现实意义。

4　结束语

本文提出的基于雅可比显著图的快速目标攻击方法 FJSMA 有效解决了 JSMA 在电磁信号识别时存在的问题。实验结果表明了 FJSMA 能够有效

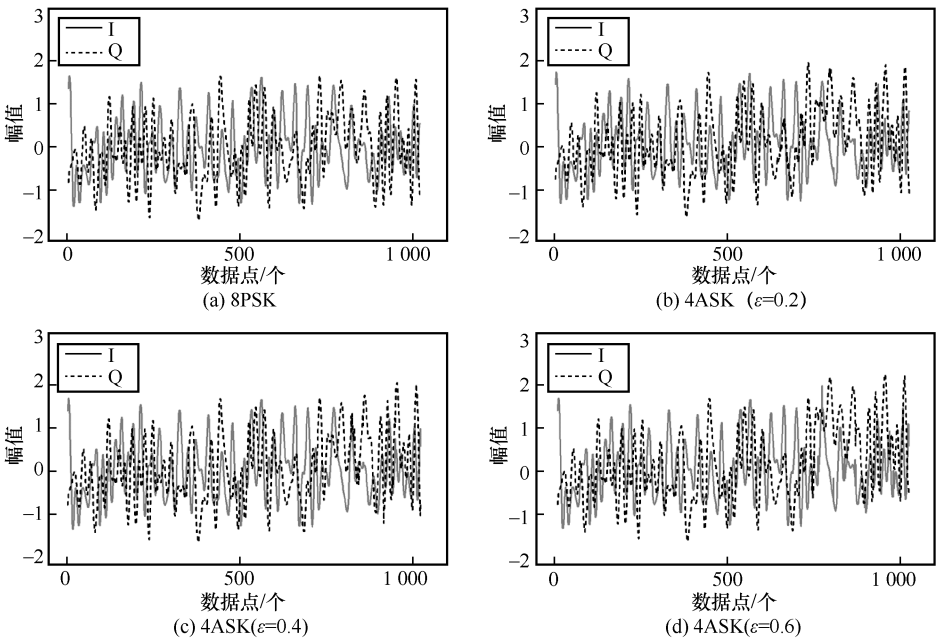


图 10　不同限制 ε 下生成的对抗样本

提升对抗样本生成速度和隐蔽性，证明了连续特征点添加对抗扰动的策略适用于强连续性的数据类型，因此未来可将该思路应用于语音等具有连续特征的数据类型对抗样本生成领域。

除了对抗攻击以外，FJSMA 所生成的对抗样本也能用于对抗防御中的对抗性训练。将 FJSMA 所生成的对抗样本赋予正确标签作为数据集与正常样本一起进行训练，能够提升模型鲁棒性，增强模型的防御能力。对抗防御还有很多方法，例如网络蒸馏、对抗性检测、网络验证等。鉴于本文篇幅有限，无法对其一一阐述。笔者所在团队已经对模型防御方法开展了相关研究，后续将根据研究进展，结合实验分析，对成果进行提炼总结。

参考文献：

[1] O'SHEA T J, CORGAN J, CLANCY T C. Convolutional radio modulation recognition networks[C]//Proceedings of International Conference on Engineering Applications of Neural Networks. Berlin: Springer, 2016: 213-226.

[2] 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.

[3] LIU K, XIANG X, LIANG Y, et al. Automatic modulation recognition through wireless sensor networks in aeronautical wireless channel[J]. IEEE Sensors Journal, 2021, 21(20): 23125-23132.

[4] 林心桐, 张琳, 吴志强, 等. 基于卷积神经网络与循环谱图的调制识别方法[J]. 太赫兹科学与电子信息学报, 2021, 19(4): 617-622.
LIN X T, ZHANG L, WU Z Q, et al. Modulation recognition method based on convolutional neural network and cyclic spectrum images[J]. Journal of Terahertz Science and Electronic Information Technology, 2021, 19(4): 617-622.

[5] JDID B, HASSAN K, DAYOUB I, et al. Machine learning based automatic modulation recognition for wireless communications: a comprehensive survey[J]. IEEE Access, 2021, 9: 57851-57873.

[6] LUAN S Y, GAO Y R, ZHOU J C, et al. Automatic modulation classification based on cauchy-score constellation and lightweight network under impulsive noise[J]. IEEE Wireless Communications Letters, 2021, 10(11): 2509-2513.

[7] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[C]//Proceedings of 2nd International Conference on Learning Representations. Piscataway: IEEE Press, 2014: 1-10.

[8] TRAMÈR F, KURAKIN A, PAPERNOT N, et al. Ensemble adversarial training: attacks and defenses[J]. arXiv Preprint, arXiv: 1705.07204, 2017.

[9] 钱亚冠, 张锡敏, 王滨, 等. 基于二阶对抗样本的对抗训练防御[J]. 电子与信息学报, 2021, 43(11): 3367-3373.
QIAN Y G, ZHANG X M, WANG B, et al. Adversarial training defense based on second-order adversarial examples[J]. Journal of Electronics & Information Technology, 2021, 43(11): 3367-3373.

[10] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[C]//Proceedings of 3rd International Conference on Learning Representations. Piscataway: IEEE Press, 2015: 1-11.

[11] KURAKIN A, GOODFELLOW I, BENGIO S. Adversarial examples in the physical world[C]//Proceedings of 4th International Conference on Learning Representations(ICLR) . Piscataway: IEEE Press, 2016: 1-14.

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

[13] XIE C H, ZHANG Z S, ZHOU Y Y, et al. Improving transferability of adversarial examples with input diversity[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 2725-2734.

[14] DING Y, ZHU G Q, CHEN D J, et al. Adversarial sample attack and defense method for encrypted traffic data[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(10): 18024-18039.

[15] PAPERNOT N, MCDANIEL P, JHA S, et al. The limitations of deep learning in adversarial settings[C]//Proceedings of the 2016 IEEE European Symposium on Security and Privacy (EuroS&P). Piscataway: IEEE Press, 2016: 372-387.

[16] 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.

[17] 黄知涛, 柯达, 王翔. 电磁信号对抗样本攻击与防御发展研究[J]. 信息对抗技术, 2023, 2(S1): 37-52.
HUANG Z T, KE D, WANG X. Research on the development of electromagnetic signal against sample attack and defense[J]. Information Countermeasure Technology, 2023, 2(S1): 37-52.

[18] KIM B, SAGDUYU Y, ERPEK T, et al. Adversarial attacks on deep learning based mmWave beam prediction in 5G and beyond[C]//Proceedings of the 2021 IEEE Statistical Signal Processing Workshop (SSP). Piscataway: IEEE Press, 2021: 590-594.

[19] KIM B, SAGDUYU Y E, ERPEK T, et al. Channel effects on surrogate models of adversarial attacks against wireless signal classifiers[C]//Proceedings of the IEEE International Conference on Communications. Piscataway: IEEE Press, 2021: 1-6.

[20] 王满喜, 史明佳, 陆科宇, 等. 电磁信号调制识别中的对抗性攻击技术研究[J]. 无线电通信技术, 2022, 48(6): 1098-1104.
WANG M X, SHI M J, LU K Y, et al. Research on adversarial attacks technology in modulation recognition[J]. Radio Communications Technology, 2022, 48(6): 1098-1104.

[21] SADEGHI M, LARSSON E G. Adversarial attacks on deep-learning

based radio signal classification[J]. IEEE Wireless Communications Letters, 2019, 8(1): 213-216.

[22] 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, 2019, 15: 1102-1113.

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

[24] 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.

[25] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[C]//Proceedings of 6th International Conference on Learning Representations. Piscataway: IEEE Press, 2018: 1-28.

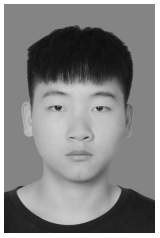
[26] 周侠, 张一然, 张剑. 基于 Grad-CAM 的电磁信号对抗攻击方法[J]. 舰船电子工程, 2023, 43(6): 204-208.
ZHOU X, ZHANG Y R, ZHANG J. Adversarial attack algorithm for electromagnetic signal based on Grad-CAM[J]. Ship Electronic Engineering, 2023, 43(6): 204-208.

[27] 李哲铭, 王晋东, 侯建中, 等. 基于显著区域优化的对抗样本攻击方法[J]. 计算机工程, 2023, 49(9): 246-255, 264.
LI Z M, WANG J D, HOU J Z, et al. Adversarial example attack method based on salient region optimization[J]. Computer Engineering, 2023, 49(9): 246-255, 264.

[28] XU J L, LUO C B, PARR G, et al. A spatiotemporal multi-channel learning framework for automatic modulation recognition[J]. IEEE Wireless Communications Letters, 2020, 9(10): 1629-1632.

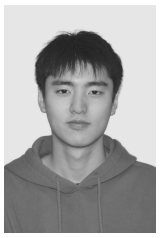
[作者简介]

张剑（1979− ），男，湖北宜昌人，博士，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、作战指挥。



周侠（1996− ），男，贵州安顺人，武汉数字工程研究所硕士生，主要研究方向为人工智能对抗样本攻防。

张一然（1999− ），女，辽宁阜新人，武汉数字工程研究所硕士生，主要研究方向为人工智能可解释性、对抗样本生成。



王梓聪（2000− ），男，湖南长沙人，武汉数字工程研究所硕士生，主要研究方向为人工智能对抗攻防。