信号人工智能对抗攻击综合分析平台

宣 琦[1], 周 晴[1], 崔 慧[1], 顾淳涛[1], 徐东伟[1], 朱佳伟[2], 王 巍[2], 杨小牛[1,2]

[1]浙江工业大学网络空间安全研究院 杭州 中国 310012
[2]通信信息控制和安全技术重点实验室 嘉兴 中国 314033

摘要 为了解决信号领域针对人工智能对抗攻击缺少全面评估的平台、针对图像人工智能对抗攻击的分析指标无法完全适用于信号领域的问题, 提出了一个信号人工智能对抗攻击综合分析平台。考虑信号与图像之间的区别, 从误分类、不可感知性、信号特性、计算代价 4 个方面着手, 提出了 10 种攻击评价指标对当下常用的 8 种攻击方法进行全面的评估。研究结果表明个别攻击方法在信号上的攻击性能表现有别于图像, 攻击方法的误分类与不可感知性、信号特性以及计算代价之间也存在相互限制的关系, 这可以为我们更好地理解及防御此类对抗攻击提供见解。

关键词 深度学习; 对抗攻击; 攻击指标; 信号处理
中图法分类号 TN92 DOI 号 10.19363/J.cnki.cn10-1380/tn.2021.07.10

A Comprehensive Evaluation Platform of Adversarial Attacks on Artificial Intelligence for Signal

XUAN Qi[1], ZHOU Qing[1], CUI Hui[1], GU Chuntao[1], XU Dongwei[1],
ZHU Jiawei[2], WANG Wei[2], YANG Xiaoniu[1,2]

[1] Institute of Cyberspace Security, Zhejiang University of Technology, Hangzhou 310012, China
[2] The Science and Technology on Communication Information Security Control Laboratory, Jiaxing 314033, China

Abstract In order to cope with the lack of a platform for comprehensive evaluation of adversarial attacks on artificial intelligence (AI) methods in the area of signal, and also the evaluation indicators on images cannot be fully applicable to signals, a comprehensive evaluation platform is proposed to test the adversarial attacks on AI methods for signals. Considering the essential difference between signal and image, 10 indicators in 4 aspects (misclassification, imperceptibility, signal characteristics, and calculating cost) were proposed to comprehensively evaluate the 8 attack methods commonly used today. The results show that the performance of individual attack on signals seems to be different from that on images, and the misclassification and imperceptibility of the attack method, the signal characteristics and the calculation cost also have a mutual limitation. All of these can provide a deep insight to better understand and further defense against such adversarial attacks in the area of signal.

Key words deep learning; adversarial attacks; attack indicators; signal processing

1 引言

20 世纪以来, 电子信息技术飞速发展, 使得信号种类及数量呈指数级增长, 面对新时期各种多元化的信号, 各种不同的分类方法也开始涌现[1]。此外, 深度学习[2]逐渐广泛应用于各个领域(生物医学[3-4]、视觉场景[5]、语音识别[6]、自然语言处理[7])。它在通信领域的应用也取得长足进展, O'Shea T J[8]等人对基于深度学习的无线电信号分类方法的性能进行了深入研究, 提出了使用更高阶矩和强大的增强梯度树分类的严格基准方法; Mendis G J 等人[9]出了一种基于深度学习的智能方法来检测和识别无线电信号, 该方法可应用于识别微型无人机系统的认知雷达和航空通信系统; Rajendran S 等人[10]提出了一种基于长短期记忆(Long Short-Term Memory, LSTM[11])的数据驱动模型, 并发现其对于分类具有不同符号率的调制信号能够发挥出强大的作用。

然而即便深度神经网络已成功应用于处理复杂问题, 却不断有研究表明[12-16]它们在强对抗环境下是非常脆弱的。最近, Han X 等人[17]开发了一种方法

通讯作者: 宣琦, 博士, 教授, Email: xuanqi@zjut.edu.cn。
本课题得到国家自然科学基金(No.61973273)资助。
收稿日期: 2020-10-24; 修改日期: 2020-12-23; 定稿日期: 2021-06-24

构造平滑的心电图追踪对抗样本, 该样本无法被人类专家察觉, 却能让应用于医学影像的深度学习模型检测异常。此外, 在信号领域, 对抗攻击研究工作也有了不少成果。Sadeghi M 等人[18]通过在信道上发送利用无线信道的开放性设计的扰动信号, 将通信系统的误块率提高了几个数量级; Sagduyu Y[19]提出了一种通过发射机对其频谱感测结果进行深度学习, 以预测空闲时隙来进行数据传输的空中频谱中毒攻击; Davaslioglu K 等人[20]针对无线通信中的深度学习应用提出了木马攻击, 取得了不错的效果; Kim B 等人[21-22]研究发现, 攻击方天线数量的增加会使攻击成功率显著提高, 并证实了调制分类器对空中对抗攻击的脆弱性。Sadeghi M 等人[23]提出了一种使用通用对抗网络生成和传输无法与预期信号可靠区分的合成信号来欺骗无线信号, 成功使信号分类精度显著降低的新颖方法。随着信号攻击领域研究的不断推进, 哪些攻击更难以捉摸或者哪种攻击得到的攻击样本隐匿性更高的关键性问题也随之出现, 这意味着当下急需一个能够对信号攻击策略进行综合评估的平台。

纵观当前研究现状, 深度学习在图像领域的应用已经较为成熟, 而在信号领域的研究成果却少之又少。Ling X 等人[24]搭建了一个较为完备的深度学习模型安全分析平台, 整合了 16 个最先进的攻击方法和 10 个攻击效用指标。Sadeghi M 等人[23]在信号攻击研究过程中提出扰动信号功率比(Perturbation Signal Rate, PSR)这一指标, 该项指标能够计算对抗样本中添加的扰动与原信号的功率之比, 与扰动噪声功率比(Perturbation Noisel Rate, PNR)综合可体现信号样本攻击前后的信噪比变化。不过仅凭借这一项指标评价攻击策略的优劣尚不够全面。

为了进一步推进信号攻击的研究, 提供一个分析平台来支持对信号领域的对抗性攻击进行全面、翔实的评估至关重要, 现有的工作尚不能满足需求, 因此本文提出了第一个信号人工智能对抗攻击综合分析平台。

本文主要研究工作如下。

1) 提出信号人工智能对抗攻击综合分析平台, 该平台中特有的信号特性计算指标(VC, ACR, APD, PSR, PNR, SNRD 具体介绍见第 2 章)让本平台有别于其他已有技术, 如 Ling X 等人[24]搭建的深度学习模型安全分析平台均适用于图像处理, 平台中提到的误分类、不可感知性、计算代价等方面的指标尚可用于信号攻击评价, 但忽略了信号在处理过程中幅度、相位等特性的变化, 本平台解决了该问题。

2) 采用了 8 种常见的攻击方法对信号样本进行攻击, 并利用信号攻击综合分析平台中的 10 项攻击效果指标对其进行全面的分析, 其中的 4 项(VC, ACR, APD, SNRD)均为本文首次提出。

3) 研究发现攻击方法的误分类与不可感知性、信号特性以及计算代价之间存在相互限制的关系, 个别攻击方法在信号上的攻击性能有别于图像。

2 综合分析平台

2.1 评价指标

研究者们通常使用简单的度量标准进行攻击方法评估, 比如误分类率便是评估攻击方法的主要指标。然而各项研究均显示, 误分类率不足以全面地描述一种攻击方法。

另一方面, 区别于图像的可视性, 信号的特性更加抽象化, 信号具有相位、幅度、功率等特有的属性。此外, 在信号处理过程中一般采用复信号[25]概念来表示实信号, 规避实信号具有共轭对称的频谱会造成的处理困难。对复数信号进行采样需要同时进行 I、Q 两路采样, 这两路数据是相关的, 无法当成独立的数据点对待。在攻击评价过程中需要上述重要参考因素, 因此一个全新的信号攻击评价平台必不可少。

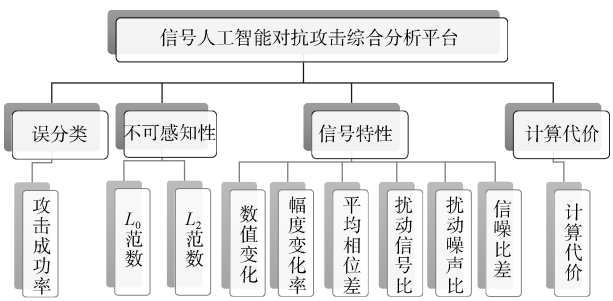


图 1 平台指标结构

Figure 1 Platform indicator structure

图像领域中已有的攻击分析平台从误分类、不可感知性、计算代价三方面对攻击策略进行分析, 分析了错误分类的百分比, 无法从表面察觉的物理量的变化, 攻击者生成对抗样本的平均速度。本文为了补充在信号攻击中对信号特性的描述, 在此基础上增加信号特性分析, 从误分类、不可感知性、信号特性、计算代价四个方面提供了如下指标(本处介绍以 Radio-ML2016.10a 数据集为例, 具体数据集介绍见第 3 章)。

1) 误分类:

(a)攻击成功率(Attack Success Rate, ASR):

$$ASR = \frac{ACC_{ori} - ACC_{adv}}{ACC_{ori}} \tag{1}$$

ASR 计算的是由攻击方造成的误分类百分比。ACC_{ori} 为原信号样本的分类精度, 而 ACC_{adv} 是对抗样本通过相同分类模型所得分类精度。攻击成功率能体现一个攻击方法导致错误分类的能力。

2) 不可感知性

(a)L_0 范数:

$$L_0 = \frac{C_c}{N} \tag{2}$$

L_0 计算的是一个信号样本在攻击后改变的点的数量占总数的比例。C_c 为一个样本(128×2 个点)中修改的点的个数, N 为一个信号样本中数据点的个数, 本文采用的数据集 N 值为 256(128×2)。

(b) L_2 范数:

$$L2 = \sqrt{\sum_{i=1}^{N} \left| V_{oi} - V_{ai} \right|^2} \tag{3}$$

L_2 计算的是一个原信号样本和对抗样本之间在数值上的欧氏距离。V_{oi} 为原始样本第 i 个数据点的数值, V_{ai} 为对抗样本第 i 个数据点的数值, N 为 256(128×2)。

3) 信号特性:

(a) 数值变化(Value Change, VC):

$$VC = \frac{1}{N} \sum_{i=1}^{N} \left| V_{oi} - V_{ai} \right| \tag{4}$$

VC 计算的是一个信号样本中每个数据点攻击前后数值变化量的平均值。VC 指标虽与 L_2 范数存在一定线性关系, 但 VC 更侧重于对幅度这一物理量的变化的描述。

(b) 幅度变化率(Amplitude Change Rate, ACR):

$$A = \sqrt{I^2 + Q^2} \tag{5}$$

$$ACR = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{A_{oi} - A_{ai}}{A_{oi}} \right| \tag{6}$$

ACR 计算的是攻击前后信号的幅度变化率。在信号处理过程中, 区别于图像中互相独立的像素点, 信号数据集中的 I/Q 两路存在一一对应的关系, 分别为复信号实部与虚部的采样值, 倘若参照与图像相同的计算方法, 则忽略了 I/Q 两路的相关性。

A 为信号有效幅值, I 和 Q 分别为信号实部和虚部的系数。A_{oi} 为原始信号第 i 个采样点的有效幅度, A_{ai} 为攻击后信号第 i 个采样点的有效幅度, n 为一个信号样本中采样点的个数, 本文所用数据集 n 值为 128。与图像中每一个独立的像素点不同, 信号中 128×2 尺寸的样本, 将其 I 路和 Q 路对应起来描述一个信号采样点会比把其当成 256 个独立点更加确切。

(c)平均相位差(Average Phase Difference, APD):

$$APD = \frac{1}{n} \sum_{i=1}^{n} \left| \arctan \frac{Q_{oi}}{I_{oi}} - \arctan \frac{Q_{ai}}{I_{ai}} \right| \tag{7}$$

APD 计算的是一个信号样本中每个采样点的相位差平均值。相位作为描述信号波形变化的重要度量, 也是评价信号攻击的重要因素, 相位的延迟可以完全改变一个信号, 从而导致无法提取出真实消息。I_{oi} 是原始信号第 i 个采样点的实部系数, Q_{oi} 是原始信号第 i 个采样点的虚部系数。I_{ai} 是攻击后信号第 i 个采样点的实部系数, Q_{ai} 是原始信号第 i 个采样点的虚部系数。

(d)扰动信号比(Perturbation Signal Rate, PSR):

$$P = \frac{\sum_{i=1}^{n} A_i^2}{n} \tag{8}$$

$$PSR = \frac{P_p}{P_s} \tag{9}$$

PSR 计算的是对抗样本添加的扰动与信号的功率比。P 是信号功率, A_i 为信号第 i 个采样点的有效幅度。

(e)扰动噪声比(Perturbation Noise Rate, PNR):

$$PNR = \frac{P_p}{P_n} \tag{10}$$

PNR 计算的是对抗样本与原样本的噪声功率比, 与上文的 PSR 综合可体现信号样本攻击前后的信噪比变化。P_p 为扰动功率, P_n 为噪声功率。信噪比是度量通信系统通信质量可靠性的一个主要技术指标, 为信号平均功率与噪声平均功率的比值, 是评价信号攻击方法优劣必不可少的一项指标。

(f) 信噪比差 (Signal Noise Rate Difference, SNRD):

$$SNR = \frac{P_s}{P_n} \tag{11}$$

$$SNRD = SNR_{adv} - SNR_{ori} \tag{12}$$

SNRD 计算的是原样本与对抗样本的信噪比差值, SNR_{ori}, SNR_{adv} 分别表示攻击前与攻击后信号样本

的信噪比。公式(11)给出的是信噪比的计算方式。

4)计算代价(Calculating Cost, CC)

计算代价定义为攻击策略生成对抗样本所用的平均时间, 本文中度量单位为秒(s)。

2.2 平台架构

图 2 给出了本平台的整体架构。通过 8 种攻击方法对信号样本进行攻击得到对样的对抗样本, 然后通过信号攻击综合分析平台提供的 10 项指标对攻击结果进行分析评价。8 种攻击方法包括现下常用的 6 种白盒攻击方法(FGSM[13], BIM[26], PGD[27], DF[16], JSMA[28], CW[29])和 2 种黑盒攻击(OPA[30], UAP[31]); 10 项评价指标包括误分类(ASR)、不可感知性(L_0 范数、L_2 范数)、信号特性(VC, APD, ACR, PSR, PNR, SNRD)、计算代价(CC)4 个模块。

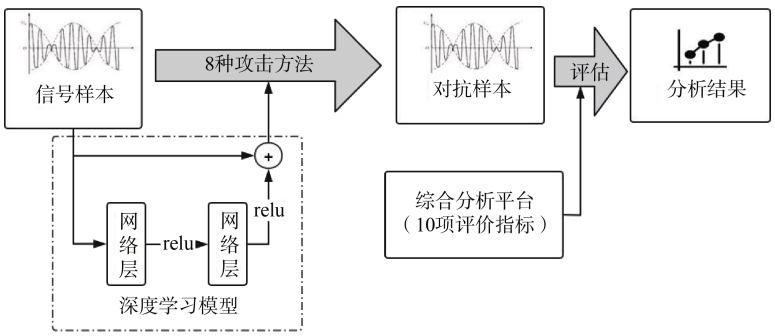


图 2 平台整体架构

Figure 2 Overall platform architecture

3 分类模型与攻击方法

3.1 分类模型

本文采用由 O'Shea T J 等人[8]提出的用于调制识别的残差神经网络 ResNet 模型。该分类方法具有容易优化、复杂度低优点的同时分类性能强大, 是当下具有强大优势的深度学习分类模型之一, 因此本文采用该模型作为深度学习分类网络。

该模型结构在残差块的基础上, 构建了残差堆叠层, 让堆叠层去拟合残差映射, 每个残差堆叠层包括一个卷积核大小为 1×1 的卷积层, 两个残差块, 一个最大池化层。在整个 ResNet 模型中,包含 6 个残差堆叠层, 数据每经过一个残差堆叠层维度减半, 且模型中所有卷积层的卷积核数量均为 32, 激活函数均采用缩放指数线性单位(Selu)。

3.2 攻击方法

由于当下已有的信号攻击工作成果较少, 本文采用的 8 种攻击方法均从图像领域迁移而来。将攻击方法从两个角度分类:根据攻击需要的条件可分为黑盒攻击和白盒攻击, 白盒攻击需要对模型结构深入了解, 知道模型的结构和各层的参数, 可以计算梯度, 黑盒攻击完全把模型当做一个黑盒, 根据一定的算法, 不断根据输出的反馈调整输入数据; 根据攻击频率, 可分为迭代攻击和非迭代攻击, 非迭代攻击只需要一步即可生成对抗样本, 而迭代攻击需要进行多次迭代更新。这两个分类并非互相独立, 两者之间存在着紧密联系, 但是为了更加清晰地分析其攻击效果, 后续将分开表述。

非迭代型白盒攻击: FGSM 在白盒环境下, 通过求出模型对输入的导数之后用符号函数得到其具体的梯度方向, 接着乘以一个步长, 得到的扰动添加在原来的输入上就得到了在 FGSM 方法的攻击样本, 表达式为:

$$X_{adv} = X + \varepsilon \cdot \mathrm{sign}(\nabla_X J(X, Y)) \tag{13}$$

X 为原始样本, X_{adv} 为对抗样本, Y 为原始样本标签, sign*()*为符号函数, ε 为可调节超参数。

迭代型白盒攻击: 在 FGSM 方法的基础上, Goodfellow I J 等人[26]提出其变体 BIM, 采取多个小步骤进行多次迭代, 并在每一步后调整方向。随后, Madry A 等人[27]提出了 BIM 的变体 PGD, 在 BIM 的基础上增加迭代轮数并添加一层随机化处理达到了更佳的分类效果。Moosavi-Dezfooli S M 等人[16]提出的 DeepFool 通过寻找原样本到目标模型决策边界的最近距离来生成对抗样本, 解决了 FGSM 未对添加的扰动范围进行界定的问题。Papernot N 等人[28]提出了 JSMA, 此方法首先计算给定样本的雅可比矩阵, 通过寻找对输出结果影响最显著的样本的输入特征来添加扰动。Carlini N 和 Wagner D 等人[29]提出的 CW, 在对目标函数的定义上进行了创新, 目标函数使用交叉熵, 迭代优化的过程就是不断减少目标函数, 因此 CW 能够调节置信度, 生成扰动也较小。

非迭代型黑盒攻击: Su J 等人[30]提出一种基于差

分进化生成单像素的对抗性扰动 OPA, 可以在最小攻击信息的条件下对网络进行欺骗。

迭代型黑盒攻击: Moosavi-Dezfooli S M 等人[31]提出的 UAP 是一种通用扰动, 具有很强迁移性, 跨数据且跨模型。

表 1 攻击方法

Table 1 Attack methods

类别	缩写	全称
白盒攻击	FGSM	快速梯度符号法 Fast Gradient Sign Method [13]
	BIM	基本迭代法 Basic Iterative Method[26]
	PGD	投影梯度下降法 Projected Gradient Descent [27]
	DF	深度欺骗法 DeepFool[6]
	JSMA	基于雅可比矩阵的显著图攻击 Jacobian-based Saliency Map Attack[28]
	CW	基于优化的攻击 Carlini and Wagner's attack [29]
黑盒攻击	OPA	单像素攻击法 One Pixel Attack[30]
	UAP	通用对抗扰动 Universal Adversarial Perturbation attack[31]

4 攻击结果与评价

4.1 实验设置

1) 数据集: 本文实验过程中使用的数据集为 Radio-ML2016.10a, 该数据集为布拉德利大学公开的调制信号数据集, 它使用 GNU Radio 合成 I/Q 信号样本, 包含 11 种调制类型, 信噪比范围从–20dB 到 18dB, 间隔 2dB 均匀分布。每个信号都有 128 个复杂的浮点时间样本。数据集大小为: 220000×128×2。I 和 Q 两路分别保存了这个 128 个信号点的实部和虚部的系数。

本文实验过程中采集了数据集中高于或等于 10dB 的高信噪比信号, 训练集样本数为 35200, 该测试集数据经过 ResNet 模型进行分类得到 91.01% 的精度。

2) 攻击参数设置: 各个不同攻击方法的常用参数值相同, 其他参数按照原作者默认值设置。

3) 其他设置: 数据集中选取的 35200 个信号样本的信噪比平均值为 15dB, 用于后续计算 SNRD。

4.2 实验结果及分析

攻击结果如表 2, 随机抽取第 5555 个信号样本波形图如图 3 所示。

1) 误分类: 现有的攻击方法仅从误分类角度来看, 都表现出强大的攻击能力。由表中可看出, 在 RadioML2016.10a 数据集上, 白盒攻击中迭代型的攻击方法攻击成功率普遍高于非迭代型的攻击方法, JSMA 和 CW 方法达到了 100%的攻击成功率, 其中原因非常直观, 与非迭代型攻击只需一步的扰动计算方法不同, 迭代型攻击通过多个复杂的迭代来寻找目标模型的最佳扰动, 因此能够达到更高的攻击成功率。但在黑盒攻击中, 迭代型攻击未在攻击成功率上体现明显优势, 这是由于黑盒攻击本身成功率较低。

表 2 攻击结果

Table 2 Attack results

攻击方法		误分类	不可感知性		信号特性						计算代价
		ASR(%)	L_0	L_2	VC	APD	ACR	PSR	PNR	SNRD	CC(s)
白盒攻击	FGSM	98.40	0.910	6.150	0.385	0.329	1.415	–12.981	0.439	–6.642	11811
	BIM	97.90	0.814	0.901	0.046	0.121	0.088	–23.383	–9.741	–1.977	79582
	PGD	99.97	0.680	1.018	0.051	0.137	0.117	–22.017	–8.375	–1.533	92852
	DF	99.78	0.910	0.452	0.019	0.066	0.043	–27.096	–13.477	0.565	25442
	JSMA	100	0.037	3.699	0.050	0.071	0.170	–7.164	6.481	–6.733	57911
	CW	100	0.910	1.602	0.078	0.212	0.153	–13.556	0.070	–2.305	89170
黑盒攻击	OPA	17.41	0.001	0.098	0.001	0.001	0.001	–7.277	–3.542	0.032	59033
	UAP	4.93	1.000	1.606	0.083	0.228	0.152	–14.917	0.083	–3.193	254.64

白盒攻击的攻击成功率明显高于黑盒攻击, 显而易见白盒攻击对模型的参数和结果都进行了深入了解, 攻击成功率自然也高于仅靠输出反馈调整输入数据的黑盒攻击。

2) 不可感知性: 本文通过 2 项指标对攻击方法的不可感知性进行了量化。大多数现有的攻击方法都使用 L_p 范数在目标函数中对攻击算法进行规范, 比如基于 L_0 范数的攻击方法, 能够在 L_0 失真指标上

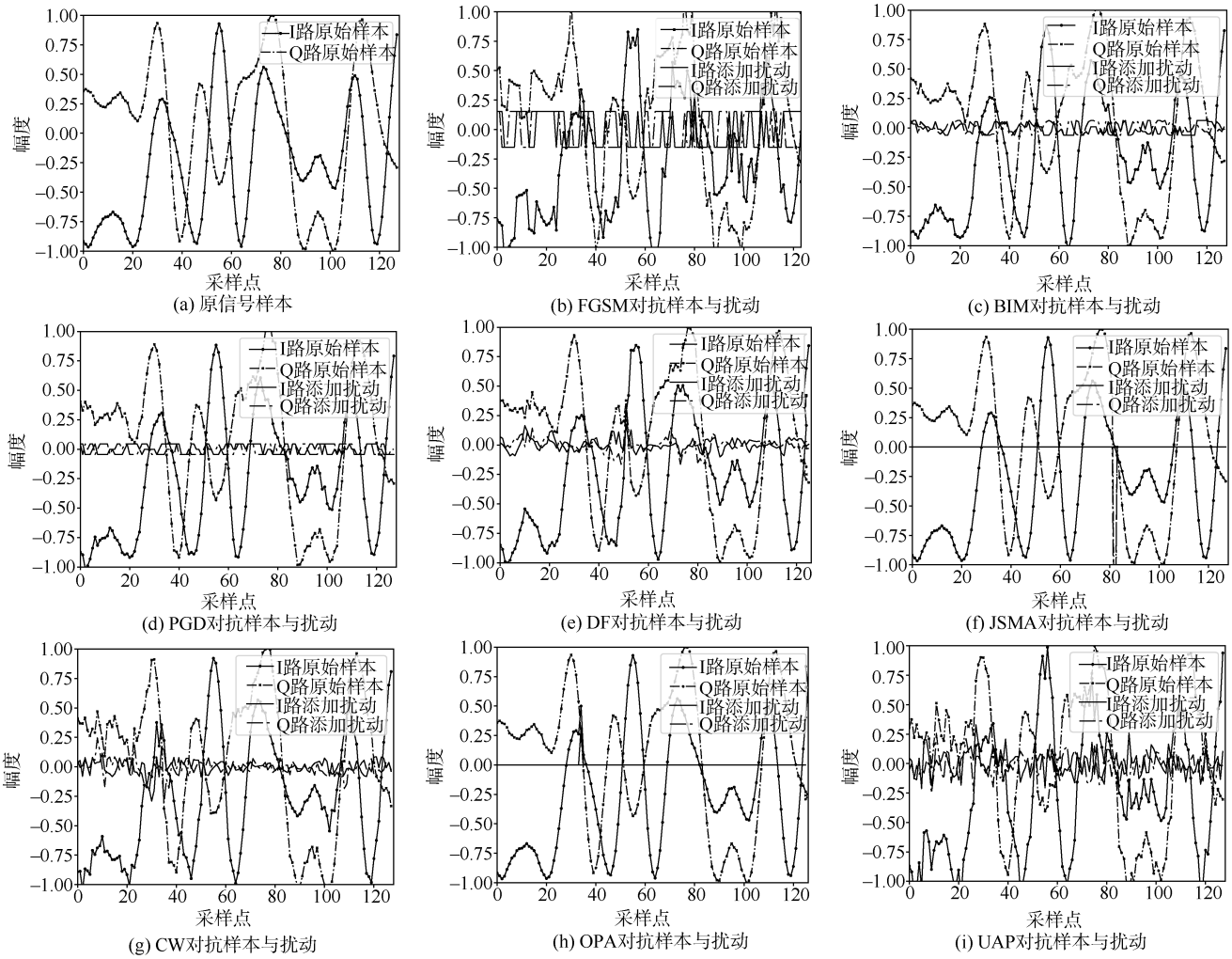


图 3　样本波形图

Figure 3　Waveform of the samples

表现更加良好, 但在 L_2 失真指标上便差强人意, 如 JSMA 方法本身比较简单直接, 在攻击过程中未限制扰动大小, 只控制了扰动个数足够少, 从表格中可看出该方法 L_0 失真非常小, 而 L_2 失真异常高。

迭代型攻击方法的不可感知性高于非迭代型攻击方法, 如非迭代攻击 FGSM 方法的 L_0 和 L_2 指标都是白盒攻击中最高的, 这就是非迭代攻击简单地一步生成对抗样本, 忽略寻找最佳扰动所致。通过表格观察可得, 误分类和不可感知性之间存在着密切联系, 两者计算的是基于两个不同目标的优化问题。误分类的优化目标是对抗样本的误分类率, 而不可感知性的优化目标为对抗样本和原样本的差异最小化。这两个目标无法保持一致, 因此误分类和不可感知性之间相互限制。白盒攻击的攻击成功率普遍较高, 但在不可感知性上的表现明显劣于黑盒攻击。

3) 信号特性: 信号是表示消息的物理量, 如电信号可以通过幅度、频率、相位等参数的变化来表示不同的消息。信号是消息的载体, 在信号攻击过程中, 信号幅度、相位等特性的变化极为重要, 过度失真将导致难以提取正确信息。本模块通过 6 项指标(VC, APD, ACR, PSR, PNR, SNRD)对攻击过程中的信号特性进行量化, VC, APD, ACR, PSR, PNR 值越小, SNRD 值越接近零, 信号特性维持程度越高。

误分类程度高的攻击方法生成的对抗样本, 维持信号特性的程度相对较低。攻击成功率达到 100% 的 JSMA 和 CW 的信号特性变化明显高于其他类别的攻击方法, 特别是 JSMA 的 PSR 和 PNR 都是所有攻击方法中数值最大的, 数值越大, 对抗样本相对原样本的信号特性变化就越大, 这也是 JSMA 追求修改数据点足够少忽略扰动幅度大小的特点所致。

黑盒攻击与白盒攻击相比, 其信号特性维持程度更高, 攻击成功率更低。

迭代型攻击方法在信号特性上表现也优于非迭代型攻击, 例如 FGSM 的每一项信号特性指标都劣于其他攻击方法, VC, APD, ACR, PSR, PNR 值都是所有方法中最高的, SNRD 值也是偏离零值程度最大的。这说明非迭代型攻击方法不仅在攻击图像时需要付出较大的攻击代价来换取误分类, 在对信号攻击时, 同样需要牺牲更多的信号特性换取更高的误分类。

信号特性中的VC指标计算方式类似于图像, 将I/Q两路信号的采样值当作独立的数据点对攻击代价进行量化, 该项指标可对对抗样本的信号特性进行描述, 却忽略了 I/Q 两路采样值之间的内在联系, 故不够全面, 需要添加其他几项信号特性指标来对攻击方法进行全面、综合的评价分析。由表 3 中数据分析可得, 其他几项信号特性指标并不完全与VC指标成正比, 这就是I/Q两路采样值之间的内在联系所致。本文所选的所有攻击方法在 VC 指标上都有较好表现, 但 FGSM 在其他几项指标中表现不足, JSMA 的 PNR 和 SNRD 也明显劣于其他攻击方法, 但其劣势却并未在VC指标中得到明显体现, 说明这两种攻击方法得到的对抗样本虽然与原样本的整体改变不大, 但其幅度、相位等信号特征却发生了较大失真。CW 方法的 VC 指标与其他攻击方法相比不算优秀, 但是其 PNR 值非常小, 说明该攻击方法添加的扰动与原样本中本就存在的噪声相比微不足道。这 6 项信号特性指标能够全面地评价攻击方法对于信号特性的影响。

4) 计算代价: 为了估计攻击的计算成本, 本文计算了各个攻击方法生成对抗样本耗费的时间。从表中观察可得, 一般情况下, 非迭代型攻击方法所用的计算时间明显低于迭代型攻击方法, 多次的迭代过程需以计算时间为代价获取更佳的攻击效果。

5 讨论与分析

本文提供的信号人工智能对抗攻击综合分析平台与其他平台相比, 不仅分析了错误分类的百分比, 无法从表面察觉的物理量的变化, 攻击者生成对抗样本的平均速度等其他平台常用因素, 同时还根据信号具有的相位、幅度、功率等特有属性提出了信号特性指标, 描述了攻击造成的信号特性变化, 为后续的解码等工作提供量化数据。此外。本平台的研究工作也为攻击和防御工作的开展提供了新的启发, 如能否针对信号特性提出一种新的攻击方法, 并针对该攻击方法提出新的防御方法。

不过, 本平台尚缺少对信号攻击的频域分析, 将在后续研究工作中继续改进; 此外, 本文只提供了号人工智能对抗攻击综合分析平台的理论分析方法, 并未搭建相应的物理系统, 因此我们将在未来的研究工作中不断完善, 并搭建物理访问平台。

6 结束语

本文提供了第一个信号人工智能对抗攻击综合分析平台, 实现了当下常用的 8 种攻击方法对信号数据集的攻击, 并通过本平台提供的 10 项攻击评价指标(误分类、不可感知性、信号特性、计算代价)对这 8 种攻击方法进行了全面、有效的评价分析, 希望本平台能为今后信号攻击和防御研究工作的推进提供帮助。

参考文献

[1] Huang Z Y. *Radio Signal Recognition Based on Deep Learning*[D]. Xi'an: Xidian University, 2018.
(黄震宇. 基于深度学习的无线电信号识别方法研究[D]. 西安: 西安电子科技大学, 2018.)

[2] LeCun Y, Bengio Y, Hinton G. Deep learning[J]. *Nature*, 2015, 521(7553): 436-444.

[3] Shen D, Wu G, Suk H I. Deep Learning in Medical Image Analysis[J]. *Annual Review of Biomedical Engineering*, 2017, 19: 221-248.

[4] Kermany D S, Goldbaum M, Cai W J, et al. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning[J]. *Cell*, 2018, 172(5): 1122-1131.e9.

[5] Owens A, Efros A A. Audio-Visual Scene Analysis with Self-Supervised Multisensory Features[C]. *Computer Vision - ECCV 2018*, 2018: 631-648.

[6] Li J C, Dai W, Metze F, et al. A Comparison of Deep Learning Methods for Environmental Sound Detection[C]. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017: 126-130.

[7] Young T, Hazarika D, Poria S, et al. Recent Trends in Deep Learning Based Natural Language Processing[J]. *IEEE Computational Intelligence Magazine*, 2018, 13(3): 55-75.

[8] 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.

[9] Mendis G J, Wei-Kocsis J, Madanayake A. Deep Learning Based Radio-Signal Identification with Hardware Design[J]. *IEEE Transactions on Aerospace and Electronic Systems*, 2019, 55(5): 2516-2531.

[10] Rajendran S, Meert W, Giustiniano D, et al. Deep Learning Models for Wireless Signal Classification with Distributed Low-Cost Spectrum Sensors[J]. *IEEE Transactions on Cognitive Communications and Networking*, 2018, 4(3): 433-445.

[11] Sundermeyer M, Schlüter R, Ney H. LSTM Neural Networks for Language Modeling[J]. *13th Annual Conference of the International Speech Communication Association 2012, INTERSPEECH 2012*, 2012, 1: 194-197.

[12] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[EB/OL]. 2013: ArXiv Preprint ArXiv:1312.6199.

[13] Goodfellow I J , Shlens J , Szegedy C. Explaining and harnessing adversarial examples[EB/OL]. 2014: ArXiv Preprint ArXiv:1412.6572.

[14] Nguyen A, Yosinski J, Clune J. Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images[C]. *2015 IEEE Conference on Computer Vision and Pattern Recognition*, 2015: 427-436.

[15] Sharif M, Bhagavatula S, Bauer L, et al. Adversarial Generative Nets: Neural Network Attacks on State-of-the-Art Face Recognition[EB/OL]. 2017

[16] Moosavi-Dezfooli S M, Fawzi A, Frossard P. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks[C]. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 2574-2582.

[17] Han X, Hu Y, Foschini L, et al. Deep Learning Models for Electrocardiograms are Susceptible to Adversarial Attack[J]. *Nature Medicine*, 2020, 26(3): 360-363.

[18] Sadeghi M, Larsson E G. Physical Adversarial Attacks Against End-to-End Autoencoder Communication Systems[J]. *IEEE Communications Letters*, 2019, 23(5): 847-850.

[19] Sagduyu Y E, Shi Y, Erpek T. Adversarial Deep Learning for Over-the-Air Spectrum Poisoning Attacks[J]. *IEEE Transactions on Mobile Computing*, 2021, 20(2): 306-319.

[20] Davaslioglu K, Sagduyu Y E. Trojan Attacks on Wireless Signal Classification with Adversarial Machine Learning[C]. *2019 IEEE International Symposium on Dynamic Spectrum Access Networks*, 2019: 1-6.

[21] Kim B, Sagduyu Y E, Erpek T, et al. Adversarial Attacks with Multiple Antennas Against Deep Learning-Based Modulation Classifiers[C]. *2020 IEEE Globecom Workshops*, 2020: 1-6.

[22] Kim B, Sagduyu Y E, Davaslioglu K, et al. Over-the-Air Adversarial Attacks on Deep Learning Based Modulation Classifier over Wireless Channels[C]. *2020 54th Annual Conference on Information Sciences and Systems*, 2020: 1-6.

[23] Sadeghi M, Larsson E G. Adversarial Attacks on Deep-Learning Based Radio Signal Classification[J]. *IEEE Wireless Communications Letters*, 2019, 8(1): 213-216.

[24] Ling X , Ji S , Zou J , et al. DEEPSEC: a uniform platform for security analysis of deep learning model[C]. *IEEE Symposium on Security and Privacy*, 2019:673-690.

[25] Gabor D. Theory of Communication. Part 1: The Analysis of Information[J]. *Journal of the Institution of Electrical Engineers - Part III*: *Radio and Communication Engineering*, 1946, 93(26): 429-441.

[26] Kurakin A, Goodfellow I J, Bengio S. Adversarial Examples in the Physical World[M]. Artificial Intelligence Safety and Security. First edition. | Boca Raton, FL: CRC Press/Taylor & Francis Group, Chapman and Hall/CRC, 2018: 99-112.

[27] Madry A, Makelov A, Schmidt L, et al. Towards Deep Learning Models Resistant to Adversarial Attacks[EB/OL]. 2017

[28] 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*, 2016: 372-387.

[29] Carlini N, Wagner D. Towards Evaluating the Robustness of Neural Networks[C]. *2017 IEEE Symposium on Security and Privacy*, 2017: 39-57.

[30] Su J W, Vargas D V, Sakurai K. One Pixel Attack for Fooling Deep Neural Networks[J]. *IEEE Transactions on Evolutionary Computation*, 2019, 23(5): 828-841.

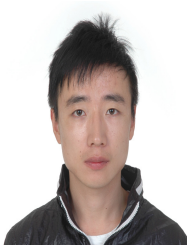
[31] Moosavi-Dezfooli S M, Fawzi A, Fawzi O, et al. Universal Adversarial Perturbations[C]. *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 86-94.

宣琦 于 2008 年在浙江大学控制科学与工程专业获得博士学位。现任浙江工业大学网络空间安全研究院教授。研究领域为人工智能安全、网络数据挖掘、信号智能。研究兴趣包括: 人工智能、信号分析、网络科学。Email: xuanqi@zjut.edu.cn

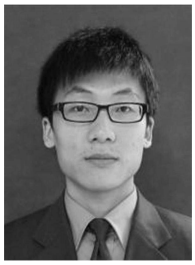
周晴 于 2020 年在浙江工业大学通信工程专业获得学士学位。现在浙江工业大学控制工程专业攻读硕士学位。研究领域为通信信号处理。研究兴趣包括: 对抗攻击、攻击检测与防御、深度学习。Email: zhouqingzjut@163.com

崔慧 于 2019 年在浙江工业大学通信工程专业获得学士学位。现在浙江工业大学控制工程专业攻读硕士学位。研究领域为深度学习、通信信号处理。研究兴趣包括: 人工智能、信号分析。Email: huicuizjut@qq.com

顾淳涛 于 2019 年在嘉兴学院南湖学院电气工程及其自动化专业获得学士学位。现在浙江工业大学控制工程专业攻读硕士学位。研究领域为人工智能安全。研究兴趣包括: 信号分析。Email: chuntaogu@zjut.edu.cn



徐东伟 于 2014 年在北京交通大学交通安全工程专业获得博士学位。现任浙江工业大学网络空间安全研究院讲师。研究领域为交通信息处理、交通复杂网络、机器学习。研究兴趣包括: 人工智能、信号分析。Email: dongweixu@zjut.edu.cn



朱佳伟 于 2012 年在电子科技大学通信与工程专业获得硕士学位。现任通信信息控制和安全技术重点实验室高级工程师。研究领域为信号处理。研究兴趣包括: 信号处理、外辐射源探测以及信号大数据。Email: 68350433@qq.com

王巍 于 2008 年在西安电子科技大学密码学专业获得博士学位。现任通信信息控制和安全技术重点实验室副主任, 研究员。研究领域为网络安全、网络通信。研究兴趣包括: 协议分析、人工智能。Email: wwzwh@163.com

杨小牛 于 1988 年在西安电子科技大学通信与电子系统专业获得硕士学位。现任通信信息控制和安全技术重点实验室主任。研究领域为通信信号处理与分析。研究兴趣包括: 软件无线电、智能信号处理、人工智能。Email: yxn2117@1126.com