面向智能无人通信系统的因果性对抗攻击生成算法

禹树文[1]，许威[1,2]，姚嘉铖[1]

（1. 东南大学移动通信全国重点实验室，江苏 南京 210096；2. 网络通信与安全紫金山实验室，江苏 南京 211111）

摘 要：考虑到基于梯度的对抗攻击生成算法在实际通信系统部署中面临着因果性问题，提出了一种因果性对抗攻击生成算法。利用长短期记忆网络的序列输入输出特征与时序记忆能力，在满足实际应用中存在的因果性约束前提下，有效提取通信信号的时序相关性，增强针对无人通信系统的对抗攻击性能。仿真结果表明，所提算法在同等条件下的攻击性能优于泛用对抗扰动等现有的因果性对抗攻击生成算法。

关键词：智能通信系统；对抗攻击；深度学习；因果系统；长短期记忆网络

中图分类号：TN911

文献标志码：A

DOI: 10.11959/j.issn.1000−436x.2024036

Causality adversarial attack generation algorithm for intelligent unmanned communication system

YU Shuwen[1], XU Wei[1,2], YAO Jiacheng[1]

1. National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China

2. Purple Mountain Laboratories, Nanjing 211111, China

Abstract: A causality adversarial attack generation algorithm was proposed in response to the causality issue of gradient-based adversarial attack generation algorithms in practical communication system. The sequential input-output features and temporal memory capability of long short-term memory networks were utilized to extract the temporal correlation of communication signals while satisfying practical causality constraints, and enhance the adversarial attack performance against unmanned communication systems. Simulation results demonstrate that the proposed algorithm outperforms existing causality adversarial attack algorithms, such as universal adversarial perturbation, under identical conditions.

Keywords: intelligent communication system, adversarial attack, deep learning, causal system, long short-term memory network

0 引言

随着科技的不断发展，新兴产业不断涌现，万物互联、虚拟现实等场景对无线通信质量提出了新的需求。6G 提出了“全频谱、全覆盖、全应用、强安全”的新发展范式，未来的无线通信系统被期望能够利用更高的频段，为广域范围内的终端设备提供极低时延、极高可靠性的高质量通信服务，以支撑 6G 时代的海量新兴应用需求[1-2]。相较于传统方案，受益于未来无线网络中的海量数据，基于数据驱动的人工智能（AI, artificial intelligence）技术能够更好地支撑 6G 的智能化需求。其中，基于深度神经网络（DNN, deep neural network）的深度学习（DL, deep learning）技术能够从大量数据中提取

收稿日期：2023−06−26；**修回日期：**2023−08−28

通信作者：许威，wxu@seu.edu.cn

基金项目：国家自然科学基金资助项目（No.62022026, No.62211530108）；中央高校基本科研业务费专项资金资助项目（No.2242022K60002, No.2242023K5003）

Foundation Items: The National Natural Science Foundation of China (No.62022026, No.62211530108), The Fundamental Research Funds for the Central Universities (No.2242022K60002, No.2242023K5003)

深层特征，因而极大地促进了面向 6G 的智能传输设计。近年来，通信领域内基于 DL 的优化算法不断涌现，在分布式信号处理[3-4]、无线传输方案[5-6]、物理层安全[7-8]、复杂架构下的资源分配[9-10]等方向都展现出巨大潜力。

在万物智联的趋势下，以自动驾驶、无人机等为代表的无人系统与通信网络的联系日趋紧密，有人系统与无人系统协同工作的模式在决策、行动上的优势也被进一步挖掘。然而，由于涉及无人系统自组网、有人无人系统接入协议、协同系统下的通感一体化设计等场景，有人无人协同系统对通信业务的可靠性、时延与速率提出了更严苛的要求。DL 算法的引入在提升无人系统性能的同时，也为无人系统与有人系统的协同工作提供更高质量的通信服务，实现 2 种系统的双向互补，有效提升协同系统在面对复杂任务时的表现[11]。尽管 DNN 有十分优越的性能，但是一个经过设计的很小幅度的输入扰动就能够使网络性能大幅恶化乃至失效[12]。同时，无线信道的共享和广播性质增加了基于 DL 的无线通信任务在空口传输过程中受到攻击的可能性[13]。因此，对于攻击的研究有助于分析智能通信系统在攻击下的运行表现，同时为智能通信系统的安全性改进提供支撑。

对抗攻击是一种通过在原始输入样本上叠加轻微扰动以使神经网络对其进行错误分类的攻击模式。所叠加的轻微扰动称为对抗扰动，它并不是单纯的高斯白噪声，而是根据原始输入样本和网络结构进行特定设计得到的。Goodfellow 等[14]提出了基于梯度的对抗攻击生成算法，即快速梯度符号法（FGSM, fast gradient sign method），将神经网络的损失函数在原始输入样本处的梯度符号作为对抗扰动。在此基础上，Kurakin 等[15]提出了基于多步迭代的对抗攻击生成算法，即基础迭代法（BIM, basic iterative method），其核心思想是在每一步迭代时采用缩小步长的 FGSM，再将结果投影至扰动的约束范围内。以上 2 种基于梯度的对抗攻击生成算法不仅依赖于目标网络结构，也依赖于原始输入样本，因此在实际的使用场景中有很大的限制。通用对抗扰动（UAP, universal adversarial perturbation）生成算法旨在寻找某一个特定的扰动，使其能够在不同的原始输入样本上都能取得较好的对抗攻击性能[16]。常见的 UAP 生成算法有基于迭代的 UAP 生成算法[16]、基于主成分分析（PCA, principal component analysis）的 UAP 生成算法[17]等。

类似地，在空口传输过程中，攻击者可以轻松截获传输信号并在其上添加恶意攻击信号，实现对接收端 DNN 的对抗攻击，进而影响有人无人协同系统的通信服务质量。从对抗攻击角度出发，现有文献进行了大量针对智能通信的攻击的研究。文献[17]针对基于 DL 的自动调制识别接收机进行攻击，采用 FGSM 和基于 PCA 的 UAP 生成算法进行对抗样本的生成，实现了比同等条件的加性白高斯噪声更好的攻击性能。文献[18]在文献[17]的基础上考虑了空口传输的信道影响，结合信道响应对文献[17]中生成的攻击进行优化，研究了在攻击者有完美信道状态信息（CSI, channel state information）和有限 CSI 情况下的攻击生成算法。除了分类网络，对抗攻击也同样适用于回归网络。文献[19]针对大规模多输入多输出（MIMO, multiple input multiple output）系统中的下行功率分配网络进行对抗攻击，通过修改 FGSM 中的损失函数使其能够处理回归网络的对抗样本生成。文献[20]则直接采用一个偏置网络替代上文提到的对抗攻击生成算法，对基于 CsiNet 的大规模 MIMO 信道状态信息反馈网络进行攻击，相较同等条件的加性白高斯噪声（AWGN, additive white Gaussian noise）有更好的攻击性能。

然而，上述的基于对抗攻击的攻击算法面临非因果性问题，难以在实际通信场景中应用。无论是 FGSM、BIM，还是偏置网络，都是在获得目标神经网络的全部输入后再基于算法计算对抗攻击信号，但是实际通信场景中的传输信号是时序的，实际的攻击系统无法在 $t+n$ 时刻对 t 时刻进行攻击，或者说无法基于 $t+n$ 时刻的信息生成 t 时刻的攻击。UAP 生成算法由于不依赖于神经网络的具体输入能够避免产生非因果性问题，但是忽略了通信信号的时间关联性，因而相比非因果性攻击有着较大的性能损失。

基于上述背景，本文主要研究了在因果性场景下的对抗攻击设计问题。本文主要的贡献如下。

1) 以典型的调制分类任务为例，提出一种新的攻击生成方法，在避免非因果性问题的同时，使用长短期记忆（LSTM, long short-term memory）网络，利用时序信号的时间关联性，增强攻击性能。

2) 给出将经典非因果性攻击算法适用于因果性场景下的修改方案，在不进行信号预测的情况下通过补全未知信号进行攻击信号设计。

3) 仿真结果表明，经典非因果性攻击算法在因果场景下的攻击性能大幅下降。同时，本文算法能够有效提取时序信号的时间关联性，相较于 UAP 生成算法和将非因果性攻击算法应用于因果性场景的修改方案都有着更优的攻击性能，且对于无监督异常检测算法有较好的鲁棒性。

1　系统模型

1.1　通信模型

本文考虑一个加性白高斯噪声信道下的自动调制分类（AMC, automatic modulation classification）任务。具体来说，考虑一个 AWGN 信道下的单输入单输出系统，发射机可能采用二进制相移键控（BPSK, binary phase shift keying）、正交相移键控（QPSK, quadrature phase shift keying）、8 移相键控（PSK, phase shift keying）、16 正交幅度调制（QAM, quadrature amplitude modulation）、64 QAM、连续相位频移键控（CPFSK, continuous phase frequency shift keying）、高斯频移键控（GFSK, Gauss frequency shift keying）、4 脉冲幅度调制（PAM, pulse amplitude modulation）、宽带频率调制（WBFM, wideband frequency modulation）、单边带调制（SSB, single-sideband modulation）、双边带调制（DSB, double-sideband modulation）等数字调制或模拟调制方式，接收机的任务是正确识别接收信号的调制方式。记发射机采用第 k 类调制方式，发射信号为 $\boldsymbol{s}_k$，则接收机的接收信号 $\boldsymbol{z}_k$ 为

$$\boldsymbol{z}_k = \boldsymbol{s}_k + \boldsymbol{n} \tag{1}$$

其中，$\boldsymbol{n}$ 为零均值加性白高斯噪声。接收机通过接收信号 $\boldsymbol{z}_k$ 判定发射信号 $\boldsymbol{s}_k$ 的调制方式。由于 $\boldsymbol{z}_k$ 为复数信号，为方便神经网络进行计算，需要对 $\boldsymbol{z}_k$ 进行预处理，表示为

$$\boldsymbol{z}'_k = \begin{bmatrix} \mathrm{Re}(z_{k,1}) & \cdots & \mathrm{Re}(z_{k,T}) \\ \mathrm{Im}(z_{k,1}) & \cdots & \mathrm{Im}(z_{k,T}) \end{bmatrix}^{\mathrm{T}} \tag{2}$$

其中，$z_{k,t} \in \mathbb{C}$ 表示 t 时刻的接收信号，T 表示发射信号长度，$\mathrm{Re}(\cdot)$ 表示取实部操作，$\mathrm{Im}(\cdot)$ 表示取虚部操作，$[\cdot]^{\mathrm{T}}$ 表示转置操作。为方便表示，下文中均以 $\boldsymbol{x}_k = \left[\boldsymbol{x}_{k,1}, \boldsymbol{x}_{k,2}, \cdots, \boldsymbol{x}_{k,T}\right]^{\mathrm{T}}$ 表示 $\boldsymbol{z}'_k$，其中

$$\boldsymbol{x}_{k,t} = \left[\mathrm{Re}(z_{k,t}), \mathrm{Im}(z_{k,t})\right]^{\mathrm{T}} \tag{3}$$

由于涉及的调制方式中存在模拟调制方式，因此经典的在星座图上采用聚类算法进行调制分类的方案[21]并不适用，本文采用 DL 的方法进行调制分类任务。

记一个 DNN 分类器为 $f\left(\cdot;\boldsymbol{\theta}_{\mathrm{AMC}}\right): \boldsymbol{\chi} \to \mathbb{R}^C$，其中，$\boldsymbol{\theta}_{\mathrm{AMC}}$ 表示网络模型参数；$\boldsymbol{\chi} \subset \mathbb{R}^p$ 表示网络输入域，p 为输入维度；C 表示分类数。对于每个输入 $\boldsymbol{x} \in \boldsymbol{\chi}$，分类器 $f(\cdot;\boldsymbol{\theta}_{\mathrm{AMC}})$ 给出一个标签，表示为

$$\hat{l}(\boldsymbol{x}, \boldsymbol{\theta}_{\mathrm{AMC}}) = \arg\max_k f_k(\boldsymbol{x}, \boldsymbol{\theta}_{\mathrm{AMC}}) \tag{4}$$

其中，$f_k(\boldsymbol{x}, \boldsymbol{\theta}_{\mathrm{AMC}})$ 代表与第 k 类有关的输出。根据这些定义，所考虑的调制分类系统的任务定义为

$$\max_{\boldsymbol{\theta}_{\mathrm{AMC}}} P\left(\hat{l}(\boldsymbol{x}_k, \boldsymbol{\theta}_{\mathrm{AMC}}) = k\right) \tag{5}$$

即最大化分类器 $f(\cdot;\boldsymbol{\theta}_{\mathrm{AMC}})$ 将接收信号 $\boldsymbol{x}_k$ 分类为第 k 类调制的概率。

文献[12]中的实验表明，目标分类网络复杂度对对抗攻击性能并无明显影响，因此，区别于文献[22]所提 VT-CNN2 深度卷积分类器，本文设计的调制分类网络结构如图 1 所示，网络主要由 3 个卷积层与 2 个线性层组成，具体网络参数在第 3 节仿真分析中给出。

1.2　攻击模型

本文所提算法采用离线训练的方式，在不考虑信号处理时间的情况下，假设攻击能够即刻实现，不需要额外时间。令攻击信号为 $\Delta\boldsymbol{z}$，则接收机实际接收信号为 $\boldsymbol{z}_k + \Delta\boldsymbol{z}$。由于 $\Delta\boldsymbol{z} \in \mathbb{C}$，与式(2)和式(3)类似，记

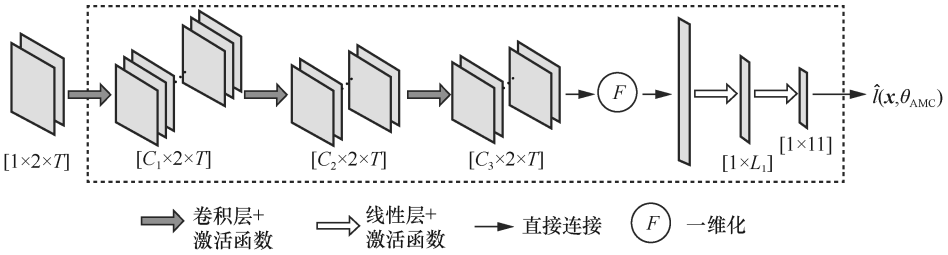


图 1　调制分类网络结构

$$\Delta \boldsymbol{z}' = \begin{bmatrix} \mathrm{Re}(\Delta z_1) & \cdots & \mathrm{Re}(\Delta z_T) \\ \mathrm{Im}(\Delta z_1) & \cdots & \mathrm{Im}(\Delta z_T) \end{bmatrix}^{\mathrm{T}} \tag{6}$$

同样地，为方便表示，下文中均以 $\Delta \boldsymbol{x} = \left[\Delta \boldsymbol{x}_1, \Delta \boldsymbol{x}_2, \cdots, \Delta \boldsymbol{x}_T\right]^{\mathrm{T}}$ 表示 $\Delta \boldsymbol{z}'$，其中

$$\Delta \boldsymbol{x}_t = \left[\mathrm{Re}\left(\Delta z_t\right), \mathrm{Im}\left(\Delta z_t\right)\right]^{\mathrm{T}},\ t = 1, 2, \cdots, T \tag{7}$$

在对抗攻击领域，攻击模型涉及攻击者知识、攻击能力、攻击强度、攻击目的等指标。在本文考虑的通信场景中，攻击模型配置如下。

1) 攻击者的知识包含网络结构与参数 $f(\cdot;\boldsymbol{\theta}_{\mathrm{AMC}})$，网络训练集与验证集，网络当前与之前时刻的输入以及对应的标签 k，即 t 时刻，攻击者的知识包含 $\boldsymbol{x}_{k,1}, \boldsymbol{x}_{k,2}, \cdots, \boldsymbol{x}_{k,t}$。

2) 攻击者仅能对当前时刻网络的输入进行更改，不能对网络本身参数或网络各层输出进行更改。

3) 攻击信号 $\Delta \boldsymbol{x}$ 满足无穷范数约束，即

$$\| \Delta \boldsymbol{x} \|_{\infty} \leqslant \epsilon \tag{8}$$

其中，ϵ 为约束边界。

4) 攻击者采用无目标攻击，即：使网络分类错误即可。对于分类任务式(5)的无目标对抗攻击建模如下

$$\begin{aligned} &\min_{\boldsymbol{\theta}} P\left(\hat{l}(\boldsymbol{x}_k + \Delta \boldsymbol{x}, \boldsymbol{\theta}_{\mathrm{AMC}}) = k\right) \\ &\text{s.t.} \| \Delta \boldsymbol{x} \|_{\infty} \leqslant \epsilon \end{aligned} \tag{9}$$

2　对抗攻击生成算法

2.1　现有对抗攻击生成算法

BIM 是 FGSM 的多步版本。FGSM 最早由 Goodfellow 等[14]于 2014 年提出，是一种高效的对抗样本生成算法。将目标网络的损失函数记为 $L\left(\boldsymbol{\theta}_{\mathrm{AMC}}, \boldsymbol{x}, \boldsymbol{y}\right)$，其中，$\boldsymbol{x}$ 为网络输入，$\boldsymbol{y} \in \{0,1\}^C$ 为标签向量。FGSM 将式(9)的优化问题转化为与网络损失函数相关的优化问题，表示为

$$\begin{aligned} &\max_{\boldsymbol{\theta}} L\left(\boldsymbol{\theta}_{\mathrm{AMC}}, \boldsymbol{x} + \Delta \boldsymbol{x}, \boldsymbol{y}\right) \\ &\text{s.t.} \| \Delta \boldsymbol{x} \|_{\infty} \leqslant \epsilon \end{aligned} \tag{10}$$

通过将样本 $\boldsymbol{x}$ 邻域内的损失函数进行线性近似，再结合式(8)的约束条件，FGSM 最终得到的对抗扰动为

$$\Delta \boldsymbol{x} = \epsilon \mathrm{sign}\left(\nabla_{\boldsymbol{x}_k} L\left(\boldsymbol{\theta}_{\mathrm{AMC}}, \boldsymbol{x}, \boldsymbol{y}\right)\right) \tag{11}$$

其中，sign(·) 函数为符号函数。对于线性模型，FGSM 能够获得最优解，但是对于非线性模型，由于在原始输入样本的邻域内进行了线性近似，FGSM 的性能并非最优。Kurakin 等[15]在后续的研究中提出了 BIM，具体流程如算法 1 所示。

算法 1　BIM

初始化　目标网络原始输入 $\boldsymbol{x}$，对应标签向量 $\boldsymbol{y}$，目标网络 $f\left(\cdot;\boldsymbol{\theta}_{\mathrm{AMC}}\right)$，对抗攻击满足的 L 无穷范数约束 ϵ，BIM 迭代次数 M，BIM 迭代步长 α

1)　$\boldsymbol{x}_0 = \boldsymbol{x}$

2) for　$m = 1 \to M$　do

3)　计算网络损失函数 $L\left(\boldsymbol{\theta}_{\mathrm{AMC}}, \boldsymbol{x}_{m-1}, \boldsymbol{y}\right)$

4)　$\boldsymbol{x}_m = \alpha \mathrm{sign}\left(\nabla_{\boldsymbol{x}_k} L\left(\boldsymbol{\theta}_{\mathrm{AMC}}, \boldsymbol{x}_{m-1}, \boldsymbol{y}\right)\right) + \boldsymbol{x}_{m-1}$

5)　$\Delta \boldsymbol{x} = \boldsymbol{x}_m - \boldsymbol{x}$

6)　截断 $\Delta \boldsymbol{x}$ 以满足 L 无穷范数约束

7)　$\boldsymbol{x}_m = \Delta \boldsymbol{x} + \boldsymbol{x}$

8) end for

如引言所述，BIM 是非因果性的，因为需要知道原始输入 $\boldsymbol{x}$ 的全部信息后才能生成相应的攻击信号 $\Delta \boldsymbol{x}$，这在无线传输系统中是不现实的。而 UAP 生成算法则是契合实际通信场景的对抗攻击生成算法，其目的是通过目标网络的部分或全部训练集样本，生成某一特定对抗攻击，使其能够对不同的样本都有较好的攻击性能。

文献[16]首次提出了 UAP 的概念并且给出了一种基于迭代的 UAP 生成算法，其核心思想是不断将攻击后的样本到决策边界的长度最短的向量与攻击向量相加以获得新的攻击向量。攻击者可获取的训练集样本数量增加，基于迭代的 UAP 生成算法所需的时间也随之增加，并且需要多次遍历样本以达到最优攻击性能，同时对于“到决策边界的长度最短的向量”的计算方式也需要进行额外的优化。因此，本文采用另一种基于 PCA 的 UAP 生成算法与本文所提算法进行性能对比，与基于迭代的 UAP 生成算法相比具有更低的复杂度和更优的性能[17]。算法流程如算法 2 所示。

算法 2　基于 PCA 的 UAP 生成算法

初始化　目标网络部分或全部训练集样本 $\boldsymbol{X} = \left[\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N\right]$，相对应的标签 $\boldsymbol{Y} = \left[\boldsymbol{y}_1, \boldsymbol{y}_2, \cdots, \boldsymbol{y}_N\right]$，目标网络 $f\left(\cdot;\boldsymbol{\theta}_{\mathrm{AMC}}\right)$，对抗攻击满足的 L 无穷范数约束 ϵ

1) 对于每组 $\left(\boldsymbol{x}_n, \boldsymbol{y}_n\right)$，通过 BIM 计算对抗攻击

$\Delta \boldsymbol{x}$，记 $\boldsymbol{N}=\left[\Delta \boldsymbol{x}_1, \Delta \boldsymbol{x}_2, \cdots, \Delta \boldsymbol{x}_N\right]^{\mathrm{T}}$

2) 计算 $\boldsymbol{N}$ 的第一主成分，记为 $\boldsymbol{v}_1$

3) $\Delta \boldsymbol{x}^{\mathrm{UAP}}=\dfrac{\epsilon \boldsymbol{v}_1}{\|\boldsymbol{v}_1\|_\infty}$

2.2　基于 LSTM 网络的对抗攻击生成算法

尽管 UAP 生成算法能够满足实际通信场景中的对抗攻击需求，但是其只依赖于网络训练集，未能利用当前网络已有的输入信息。本节提出一种既满足实际通信场景中对抗攻击的因果性要求，又能够利用当前网络已有的输入信息的攻击生成算法。

循环神经网络（RNN, recurrent neural network）在传统的前馈神经网络基础上进行拓展，引入了一个隐藏状态，以更好地提取序列数据的特征。对于一串时间序列输入 $\boldsymbol{x}=\left[\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_T\right]^{\mathrm{T}}$，在 t 时刻，RNN 根据当前隐藏状态 $\boldsymbol{h}_t$ 和输入 $\boldsymbol{x}_t$ 得到网络输出 $\boldsymbol{y}_t$，并且更新下一时刻隐藏状态 $\boldsymbol{h}_{t+1}$，表示为

$$\begin{aligned}\boldsymbol{h}_t&=\boldsymbol{U}\boldsymbol{x}_t+\boldsymbol{W}\boldsymbol{h}_{t-1}\\ \boldsymbol{y}_t&=\boldsymbol{V}\phi(\boldsymbol{h}_t)\end{aligned}\tag{12}$$

其中，$\boldsymbol{U}$、$\boldsymbol{W}$、$\boldsymbol{V}$ 为网络参数，$\phi(\cdot)$ 为隐藏层激活函数，$\boldsymbol{h}_0$ 为 RNN 初始状态。

可以看出，无论是序列性的输入输出的模式还是对于序列历史信息的利用，RNN 都符合本文对于对抗攻击生成的需求。但是 RNN 在处理长序列的时候由于面临梯度消失或者梯度爆炸的问题，不能有效地提取序列的长期关联性[23-24]，为了解决该问题，Hochreiter 等[23]在 1997 年首次提出 LSTM 网络的概念。如图 2 所示，LSTM 网络在隐藏状态 $\boldsymbol{h}_t$ 的基础上增加了细胞状态 $\boldsymbol{c}_t$，引入了遗忘门、输入门与输出门结构，有效地解决了长时依赖问题，其中乘法器为哈达玛积。LSTM 网络具体算法表示为

$$\begin{aligned}\boldsymbol{f}_t&=\mathrm{sigmoid}\left(\boldsymbol{W}_f\left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right]+\boldsymbol{b}_f\right)\\ \boldsymbol{i}_t&=\mathrm{sigmoid}\left(\boldsymbol{W}_i\left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right]+\boldsymbol{b}_i\right)\\ \tilde{\boldsymbol{c}}_t&=\tanh\left(\boldsymbol{W}_c\left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right]+\boldsymbol{b}_c\right)\\ \boldsymbol{o}_t&=\mathrm{sigmoid}\left(\boldsymbol{W}_o\left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right]+\boldsymbol{b}_o\right)\\ \boldsymbol{c}_t&=\boldsymbol{f}_t\circ\boldsymbol{c}_{t-1}+\boldsymbol{i}_t\circ\tilde{\boldsymbol{c}}_t\\ \boldsymbol{h}_t&=\boldsymbol{o}_t\circ\tanh\left(\boldsymbol{c}_t\right)\end{aligned}\tag{13}$$

其中，$\boldsymbol{f}_t$、$\boldsymbol{i}_t$、$\tilde{\boldsymbol{c}}_t$、$\boldsymbol{o}_t$ 分别为图 2 下侧 4 个神经元的输出，$\boldsymbol{W}_f$、$\boldsymbol{b}_f$、$\boldsymbol{W}_i$、$\boldsymbol{b}_i$、$\boldsymbol{W}_c$、$\boldsymbol{b}_c$、$\boldsymbol{W}_o$、$\boldsymbol{b}_o$ 为网络参数，$\circ$ 为哈达玛积。

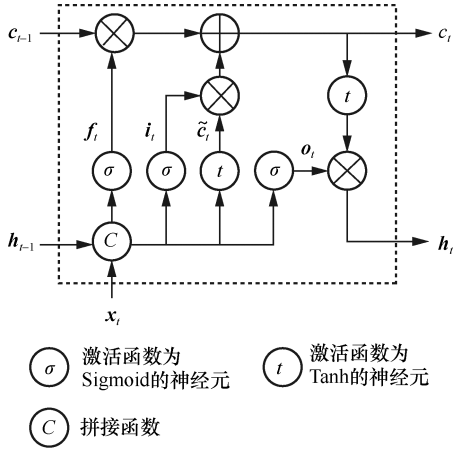


图 2　LSTM 网络结构

本文采用基于 LSTM 网络生成因果性对抗攻击。为了使生成的对抗攻击 $\Delta \boldsymbol{x}$ 满足 L 无穷范数约束，因果性对抗攻击生成网络的输出为 $\Delta \boldsymbol{x}$ 而并非 $\Delta \boldsymbol{x}+\boldsymbol{x}$，本文所设计的网络结构如图 3 所示，由 LSTM 网络、线性层与幅度调整模块构成。其中，LSTM 网络用于记忆并提取输入信号的时间相关

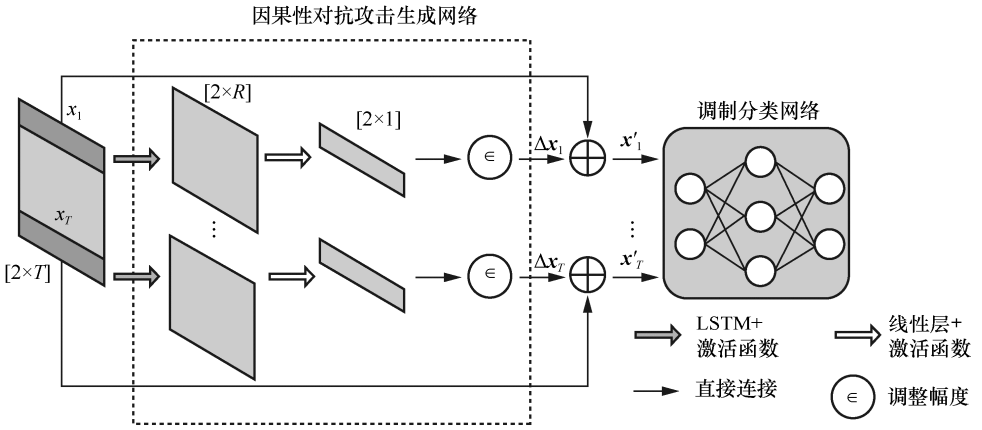


图 3　本文所设计的网络结构

性，每个时刻的线性层均相同，且激活函数采用 Tanh，目的是与幅度调整模块协同，使生成的对抗攻击 $\Delta \boldsymbol{x}$ 满足 L 无穷范数约束。

目标网络为已训练好的图 1 中的网络 $f(\cdot;\boldsymbol{\theta}_{\mathrm{AMC}})$，在训练攻击生成网络时固定网络模型参数，只参与梯度传播，不进行参数更新。记攻击生成网络为 $g(\cdot;\boldsymbol{\theta}_g)$，其中，$\boldsymbol{\theta}_g=(\boldsymbol{\theta}_{\mathrm{LSTM}};\boldsymbol{\theta}_{\mathrm{Linear}})$ 为网络模型参数，$\boldsymbol{\theta}_{\mathrm{LSTM}}$ 和 $\boldsymbol{\theta}_{\mathrm{Linear}}$ 分别为 LSTM 网络 $m(\cdot;\boldsymbol{\theta}_{\mathrm{LSTM}})$ 和线性层 $n(\cdot;\boldsymbol{\theta}_{\mathrm{Linear}})$ 的参数。网络 $g(\cdot;\boldsymbol{\theta}_g)$ 的优化目标为

$$\max_{\boldsymbol{\theta}_g} L_f\left(\boldsymbol{\theta}_{\mathrm{AMC}}, \boldsymbol{x}+g\left(\boldsymbol{x},\boldsymbol{\theta}_g\right), \boldsymbol{y}\right) \tag{14}$$

其中，L_f 和 $\boldsymbol{\theta}_{\mathrm{AMC}}$ 分别为图 1 网络的损失函数和模型参数，具体训练与测试流程如算法 3 所示。

算法 3　因果性对抗攻击生成算法

训练部分初始化　目标网络部分或全部训练集样本 $\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N]$，相对应的标签 $\boldsymbol{Y}=[\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_N]$，目标网络 $f(\cdot;\boldsymbol{\theta}_{\mathrm{AMC}})$，对抗攻击满足的 L 无穷范数约束 ϵ，待训练的 LSTM 网络 $m(\cdot;\boldsymbol{\theta}_{\mathrm{LSTM}})$，线性层 $n(\cdot;\boldsymbol{\theta}_{\mathrm{Linear}})$

测试部分初始化　目标网络原始输入 $\boldsymbol{x}=\left[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_T\right]^{\mathrm{T}}$，经过训练的网络 $m(\cdot;\boldsymbol{\theta}_{\mathrm{LSTM}})$，$n(\cdot;\boldsymbol{\theta}_{\mathrm{Linear}})$

训练过程

1) 对于每组 $(\boldsymbol{x}_n,\boldsymbol{y}_n)$，计算对抗攻击 $\Delta\boldsymbol{x}_n=\epsilon\tanh\left(n\left(\mathrm{ReLU}\left(m\left(\boldsymbol{x}_n;\boldsymbol{\theta}_{\mathrm{LSTM}}\right)\right);\boldsymbol{\theta}_{\mathrm{Linear}}\right)\right)$

2) 计算目标网络在对抗攻击下的损失函数 $L\left(\boldsymbol{\theta}_{\mathrm{AMC}},\boldsymbol{x}+\Delta\boldsymbol{x},\boldsymbol{y}\right)$

3) 以 $-L\left(\boldsymbol{\theta}_{\mathrm{AMC}},\boldsymbol{x}+\Delta\boldsymbol{x},\boldsymbol{y}\right)$ 作为损失函数更新 $\boldsymbol{\theta}_{\mathrm{LSTM}}$ 和 $\boldsymbol{\theta}_{\mathrm{Linear}}$

4) 重复步骤 1)～3)，直至 $-L\left(\boldsymbol{\theta}_{\mathrm{AMC}},\boldsymbol{x}+\Delta\boldsymbol{x},\boldsymbol{y}\right)$ 收敛

测试过程

1) for $t=1\to T$ do

2) $\Delta\boldsymbol{x}_t=\epsilon\tanh\left(n\left(\mathrm{ReLU}\left(m\left(\boldsymbol{x}_t;\boldsymbol{\theta}_{\mathrm{LSTM}}\right)\right);\boldsymbol{\theta}_{\mathrm{Linear}}\right)\right)$

3) end for

4) $\Delta\boldsymbol{x}^{\mathrm{LSTM}}=\left[\Delta\boldsymbol{x}_1,\Delta\boldsymbol{x}_2,\cdots,\Delta\boldsymbol{x}_T\right]^{\mathrm{T}}$

3　仿真分析

3.1　仿真参数与对比算法

本文采用 GNU radio ML 数据集 RML 2016.10a[21]，包含 220 000 个样本和 11 种不同的调制方案，分别为 BPSK、QPSK、8PSK、16QAM、64QAM、CPFSK、GFSK、4PAM、WBFM、SSB 和 DSB。样本产生于 20 种不同的信噪比（SNR, signal to noise ratio）水平，SNR 为−20～18 dB，步长为 2 dB。每个样本为 2×128 的矩阵，对应于 128 个同相分量和 128 个正交分量。本文取 SNR 为−10～18 dB 共 165 000 个样本，其中训练集包含 115 500 个样本，验证集包含 33 000 个样本，测试集包含 16 500 个样本，攻击者知识包含训练集与验证集。调制分类网络各层神经元个数分别为 $C_1=512$，$C_2=256$，$C_3=160$，$L_1=256$。除最后一层线性层的激活函数为 Softmax 外，其余网络层激活函数均为 ReLU。网络在所有 SNR 上训练；BIM 算法、基于 PCA 的 UAP 生成算法与本文所提的基于 LSTM 的因果性对抗攻击生成算法均在测试环境的 SNR 与干扰噪声比（JNR, jamming to noise ratio）上训练。本文所提因果性攻击生成网络中 LSTM 网络神经元个数 $R=128$，LSTM 网络与线性层激活函数均为 Tanh。

对于 BIM，迭代步长与迭代次数都会大幅影响算法性能。图 4 给出了 SNR = 0、JNR = −10 dB 时，采用不同迭代步长的 BIM 的准确率。其中，$\epsilon_{-10\mathrm{dB}}$ 为 JNR = −10 dB 时对应的 L 无穷范数约束。本文在该条件下，选取迭代次数为 20，迭代步长为 $\frac{3\epsilon_{-10\mathrm{dB}}}{10}$。其余 SNR 与 JNR 条件下 BIM 的迭代次数与迭代步长类似可得。SNR=0 场景下 BIM 算法的迭代次数与迭代步长如表 1 所示。

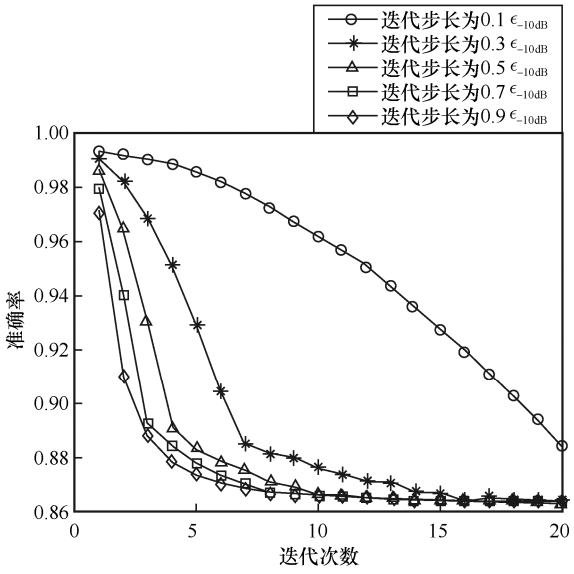


图 4　不同迭代步长的 BIM 的准确率

表 1　BIM 算法的迭代次数与迭代步长

JNR/dB	迭代次数	迭代步长
−30	1	$\epsilon_{-30\text{dB}}$
−26	1	$0.8\epsilon_{-26\text{dB}}$
−22	1	$\epsilon_{-22\text{dB}}$
−18	3	$\epsilon_{-18\text{dB}}$
−14	18	$0.2\epsilon_{-14\text{dB}}$
−10	20	$0.3\epsilon_{-10\text{dB}}$
−6	20	$0.5\epsilon_{-6\text{dB}}$
−2	17	$0.7\epsilon_{-2\text{dB}}$
2	19	$0.7\epsilon_{2\text{dB}}$
6	18	$0.9\epsilon_{6\text{dB}}$

3.2　算法性能

本节分析目标分类网络在不同攻击方式下的分类准确率随 JNR 变化的趋势，SNR = 0 时不同攻击方式对系统分类性能的影响如图 5 所示。其中，BIM 攻击仍基于网络的全部输入，即非因果性攻击；LSTM 攻击表示本文算法生成的攻击。为了使原本满足 L2 范数约束的高斯攻击满足 L 无穷范数约束，本文定义高斯攻击为

$$\Delta \boldsymbol{x}_{\text{gau}} = \epsilon \text{sign}(\boldsymbol{n}') \tag{15}$$

其中，$\boldsymbol{n}'$ 为均值为零的随机高斯向量。由式(15)可知，此时高斯攻击为固定长度、随机方向的攻击向量。随机分类指目标分类网络随机给出信号的调制分类，此时的准确率约为 0.090 9。可以看出，经过设计的攻击生成算法的攻击性能均优于高斯攻击，说明基于 DL 的智能通信对于对抗攻击更敏感，因此针对对抗攻击的研究是有意义的。同时，本文提出的因果性对抗攻击生成算法由于提取了信号的时间相关性特征，因此相比基于 PCA 的 UAP 生成算法性能更优。

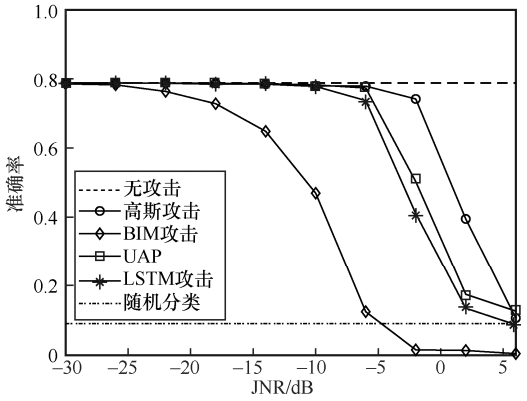


图 5　SNR = 0 时不同攻击方式对系统分类性能的影响

从图 5 可以看出，BIM 因为没有因果性约束，相较于服从因果性约束的算法有巨大的性能提升。为了证明 BIM 在因果系统下性能会受到限制，同时进一步分析本文所提攻击在因果性约束场景下的性能提升，本文提出 3 种将 BIM 应用于因果系统的方案：在每个时刻，通过已知信号（分别为 0 信号、高斯信号、已有信号）补全后续未知输入，再根据补全后的输入得到预测标签，最后根据补全输入与预测标签通过 BIM 得到当前时刻的攻击。具体如算法 4 所示，其中 $\textbf{ones}(1, i)$ 为长度为 i 的全 1 行向量，$\otimes$ 为克罗内克积。

算法 4　3 种因果性 BIM

初始化　目标网络原始输入 $\boldsymbol{x} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_T]$，目标网络 $f(\cdot;\boldsymbol{\theta})$，$\boldsymbol{x}$ 满足的约束条件 $a \leqslant h(\boldsymbol{x}) \leqslant b$

1) for $t = 1 \to T$ do
2)　//方案 1：采用 0 信号补全
3)　$\boldsymbol{x}' = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_t, 0, \cdots, 0]$
4)　//方案 2：采用高斯信号补全
5)　$\boldsymbol{x}' = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_t, m(\boldsymbol{n}'_{t+1}), \cdots, m(\boldsymbol{n}'_T)]$，其中，$\boldsymbol{n}'_k \sim \mathcal{CN}(0,1)$，$m(\cdot)$ 满足 $a \leqslant h(m(\boldsymbol{n}'_k)) \leqslant b$
6)　//方案 3：采用已有信号补全
7)　计算重复次数 $i = \left\lceil \dfrac{T}{t} \right\rceil$
8)　$\tilde{\boldsymbol{x}} = \textbf{ones}(1, i) \otimes [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_t]$
9)　$\boldsymbol{x}' = \tilde{\boldsymbol{x}}_{1:T}$
10)　$\boldsymbol{y}' = f(\boldsymbol{x}', \boldsymbol{\theta})$
11)　对于 $(\boldsymbol{x}', \boldsymbol{y}')$，采用 BIM 计算对抗样本 $\boldsymbol{x}'_{\text{adv}}$
12)　$\boldsymbol{x}_{\text{adv}}(t) = \boldsymbol{x}'_{\text{adv}}(t)$
13) end for

SNR = 0 时满足因果性的 BIM 攻击与其他类型攻击性能对比如图 6 所示。从图 6 可以看出，采用高斯信号进行补全的 BIM 性能甚至不如同等条件下的高斯攻击。采用 0 信号进行补全的 BIM 的性能虽然优于高斯攻击，但性能依旧不如基于 PCA 的 UAP 生成算法和本文提出的因果性对抗攻击。采用已有信号进行补全的 BIM 的性能优于基于 PCA 的 UAP 生成算法但是劣于本文所提攻击算法。这是因为该种补全方案相较其余 2 种补全方案更有效地利用了目标网络已有的输入信息，因此能够获得优于基于 PCA 的 UAP 生成算法的性能，但是由于缺乏对于输入信号的时间相关性提取，因而性能劣于本文所提攻击算法。

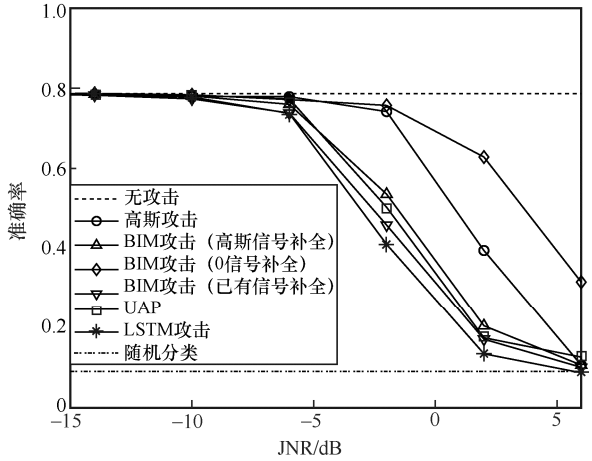


图 6　SNR = 0 时满足因果性的 BIM 攻击与其他类型攻击性能对比

由此可见，BIM 在因果系统下的攻击性能会被大幅抑制。同时，相较基于 PCA 的 UAP 生成算法，本文所提算法相较于以不同方式进行补全的 BIM 算法都有着更高的攻击性能增益。

3.3　不同 SNR 下算法适用性

不同 SNR 下 UAP 生成算法与本文所提因果性攻击生成算法的性能对比如图 7 所示，其中虚线分别为对应 SNR 下的无额外攻击的性能。从图 7 可以看出，在不同 SNR 条件下，本文所提算法均能有效提取输入信号的时序相关性，实现优于基于 PCA 的 UAP 生成算法的攻击性能。特别地，在低信噪比的情况下，由于噪声的影响使目标网络在无额外攻击的情况下的分类准确率较低（约 0.166 4），接近随机分类情况下的性能下界，因而额外攻击能够带来的性能增益有限，在此情况下，所提算法相较于基于 PCA 的 UAP 生成算法仍存在性能增益。

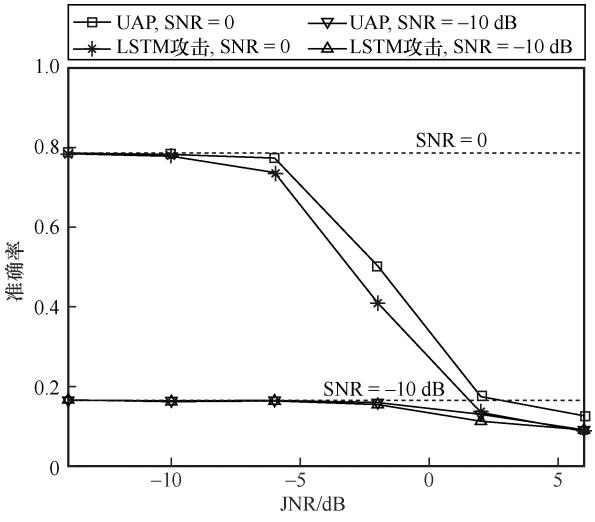


图7　不同SNR下UAP生成算法与本文所提因果性攻击生成算法的性能对比

3.4　在接收机存在防御算法下的攻击性能

常见的无监督异常检测算法有基于 PCA 的异常检测算法和 MAD-GAN（multivariate anomaly detection generative adversarial network）等[25]。在基于 PCA 的异常检测算法下，SNR = 0 时不同攻击方式的异常样本率如图 8 所示。其中，基于 PCA 的异常检测算法采用训练集数据训练，采用置信度为 0.9 的平方预测误差（SPE, squared prediction error）作为检测标准。从图 8 可以看出，随着 JNR 的增大，传统的高斯攻击被检测出的概率增大，而 UAP 生成算法和本文所提算法由于针对特征空间而非样本空间进行攻击，因此几乎不受影响。

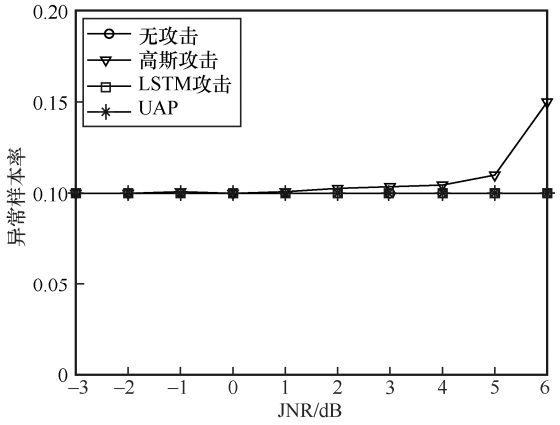


图 8　SNR = 0 时不同攻击方式的异常样本率

4　结束语

本文选取 AWGN 信道下的自动调制识别场景，对有人无人协同的智能通信的对抗攻击生成算法进行研究。通过利用 LSTM 网络的序列输入输出特性与记忆特性，在满足实际通信场景的因果性条件下，充分提取发射信号的时序关联性，设计了一种基于 LSTM 网络的因果性对抗攻击生成算法。仿真结果表明，相较于传统因果性对抗攻击生成算法，本文所提算法在因果性约束场景下具有更优的攻击性能。作为提取信号时序信息辅助攻击生成的启发式研究，本文假设攻击能够即刻实现，未来的研究中将会考虑攻击存在时延的情况，同时也将进一步针对防御算法下的对抗攻击性能开展研究。

参考文献：

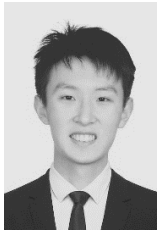
[1] YOU X H, WANG C X, HUANG J, et al. Towards 6G wireless communication networks: vision, enabling technologies, and new paradigm shifts[J]. Science China Information Sciences, 2020, 64(1): 1-74.

[2] XU W, HUANG Y M, WANG W, et al. Toward ubiquitous and intelligent 6G networks: from architecture to technology[J]. Science China Information Sciences, 2023, 66(3): 1-2.
[3] XU W, YANG Z H, NG D W K, et al. Edge learning for B5G networks with distributed signal processing: semantic communication, edge computing, and wireless sensing[J]. IEEE Journal of Selected Topics in Signal Processing, 2023, 17(1): 9-39.
[4] YAO J C, YANG Z H, XU W, et al. GoMORE: global model reuse for resource-constrained wireless federated learning[J]. IEEE Wireless Communications Letters, 2023, 12(9): 1543-1547.
[5] WEI K, XU J D, XU W, et al. Distributed neural precoding for hybrid mmWave MIMO communications with limited feedback[J]. IEEE Communications Letters, 2022, 26(7): 1568-1572.
[6] XIA W C, ZHENG G, ZHU Y X, et al. A deep learning framework for optimization of MISO downlink beamforming[J]. IEEE Transactions on Communications, 2020, 68(3): 1866-1880.
[7] SHI W, XU W, YOU X H, et al. Intelligent reflection enabling technologies for integrated and green internet-of-everything beyond 5G: communication, sensing, and security[J]. IEEE Wireless Communications, 2023, 30(2): 147-154.
[8] XIE R J, XU W, YU J B, et al. Disentangled representation learning for RF fingerprint extraction under unknown channel statistics[J]. IEEE Transactions on Communications, 2023, 71(7): 3946-3962.
[9] QI Q, WANG J Y, MA Z Y, et al. Knowledge-driven service offloading decision for vehicular edge computing: a deep reinforcement learning approach[J]. IEEE Transactions on Vehicular Technology, 2019, 68(5): 4192-4203.
[10] HUANG K, LUO Z Z, LIANG L, et al. Fast spectrum sharing in vehicular networks: a meta reinforcement learning approach[C]// Proceedings of the IEEE 96th Vehicular Technology Conference (VTC2022-Fall). Piscataway: IEEE Press, 2022: 1-5.
[11] 陈杰, 辛斌. 有人/无人系统自主协同的关键科学问题[J]. 中国科学: 信息科学, 2018, 48(9): 1270-1274.
CHEN J, XIN B. Key scientific problems in the autonomous cooperation of manned-unmanned systems[J]. Scientia Sinica (Informationis), 2018, 48(9): 1270-1274.
[12] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[J]. arXiv Preprint, arXiv: 1312.6199, 2013.
[13] YAO J C, YANG Z H, XU W, et al. Imperfect CSI: a key factor of uncertainty to over-the-air federated learning[J]. IEEE Wireless Communications Letters, 2023, 12(12): 2273-2277.
[14] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv Preprint, arXiv: 1412.6572, 2014.
[15] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial machine learning at scale[J]. arXiv Preprint, arXiv: 1611.01236, 2016.
[16] MOOSAVI-DEZFOOLI S M, FAWZI A, FAWZI O, et al. Universal adversarial perturbations[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2017: 86-94.
[17] SADEGHI M, LARSSON E G. Adversarial attacks on deep-learning based radio signal classification[J]. IEEE Wireless Communications Letters, 2019, 8(1): 213-216.
[18] KIM B, SAGDUYU Y E, DAVASLIOGLU K, et al. Channel-aware adversarial attacks against deep learning-based wireless signal classifiers[J]. IEEE Transactions on Wireless Communications, 2022, 21(6): 3868-3880.
[19] MANOJ B R, SADEGHI M, LARSSON E G. Downlink power allocation in massive MIMO via deep learning: adversarial attacks and training[J]. IEEE Transactions on Cognitive Communications and Networking, 2022, 8(2): 707-719.
[20] LIU Q, GUO J J, WEN C K, et al. Adversarial attack on DL-based massive MIMO CSI feedback[J]. Journal of Communications and Networks, 2020, 22(3): 230-235.
[21] TIAN J J, PEI Y Y, HUANG Y D, et al. Modulation-constrained clustering approach to blind modulation classification for MIMO systems[J]. IEEE Transactions on Cognitive Communications and Networking, 2018, 4(4): 894-907.
[22] O'SHEA T, WEST N E. Radio machine learning dataset generation with GNU radio[C]//Proceedings of the GNU Radio Conference. [S.l.:s.n.], 2016: 69-74.
[23] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.
[24] GERS F A, SCHMIDHUBER J, CUMMINS F. Learning to forget: continual prediction with LSTM[J]. Neural Computation, 2000, 12(10): 2451-2471.
[25] LI D, CHEN D C, SHI L, et al. MAD-GAN: multivariate anomaly detection for time series data with generative adversarial networks[C]//Proceedings of 28th International Conference on Atificial Neural Networks (ICANN). Berlin: Springer, 2019: 703-716.

[作者简介]

禹树文（1996– ），男，蒙古族，江苏盐城人，东南大学博士生，主要研究方向为智能通信、通感一体化。

许威（1982– ），男，江苏如皋人，博士，东南大学教授、博士生导师，主要研究方向为无线通信、智能通信等。



姚嘉铖（1999– ），男，江苏如东人，东南大学博士生，主要研究方向为无线分布式智能。