DOI:10.19651/j.cnki.emt.2312715

面向抗干扰跳频通信的混合改进 DQN 决策算法*

夏重阳[1] 张剑书[1,2] 吴晓富[1] 靳 越[1]
(1.南京邮电大学通信与信息工程学院 南京 210003; 2.南京工程学院计算机工程学院 南京 211167)

摘 要: 针对复杂电磁环境下的跳频抗干扰通信决策问题,提出了一种新的混合深度循环 Q 网络(MixDRQN)决策算法。该深度决策算法有效集成了双深度 Q 网络(DoubleDQN)和对决深度 Q 网络(DuelingDQN)两种决策机理的优点,并在信号处理前端引入长短时记忆(LSTM)层,以增强决策网络对输入频谱瀑布信号的时间相关特征提取能力。研究表明,所提出的混合决策算法通过引入 DoubleDQN 解决了基于 ε-greedy 算法导致的 Q 值估计偏高的问题,同时通过 DuelingDQN 和前端增加的 LSTM 层,能有效学习输入频谱瀑布信号的时间相关特征。实验结果显示,所提方法在多种干扰信号下的收敛速度及抗干扰性能均显著提升,收敛速度较已有算法提升 8 倍以上。

关键词: 通信抗干扰;强化学习;深度 Q 网络;长短时记忆
中图分类号: TN973.3 **文献标识码:** A **国家标准学科分类代码:** 510.1050

Novel mixed DQN reinforcement learning algorithm for frequency hopping anti-jamming communications

Xia Chongyang[1] Zhang Jianshu[1,2] Wu Xiaofu[1] Jin Yue[1]
(1. College of Communication and Information Engineering, Nanjing University of Posts and Telecommunication, Nanjing 210003, China;
2. School of Computer Engineering, Nanjing Institute of Engineering, Nanjing 211167, China)

Abstract: This paper investigates the problem of anti-jamming communications with intelligent frequency hopping in complex electromagnetic environment. Essentially, this paper proposes a new mixed deep recurrent Q-learning network (MixDRQN) for reinforcement learning (RL) of the optimal anti-jamming strategy. The proposed deep RL algorithm effectively combines double deep Q-learning network (DoubleDQN) and dueling deep Q-learning network (DuelingDQN), and further introduces long short-term memory (LSTM) layer for preprocessing the time-sensitive inputs. With the use of DoubleDQN, the proposed RL algorithm solves the problem of Q-value over-estimation caused by ε-greedy algorithm. In the mean time, the use of DuelingDQN and the LSTM layer has been proved to be very efficient for learning the time-correlated feature of inputs. Extensive experimental results show that both the convergence speed and anti-jamming performance are significantly improved, and in particular, the convergence speed of the proposed RL algorithm is more than 8 times higher than that of the existing RL algorithms.

Keywords: communication anti-jamming; reinforcement learning; DQN; LSTM

0 引 言

随着无线通信与人工智能的快速发展,通信对抗技术也在向智能化对抗的方向转变。人工智能 与电子对抗的结合使得传统通信对抗(包括干扰与抗干扰)技术的有效性面临严峻的挑战。目前,通信抗干扰体制主要依赖于跳频[1]通信技术,其抗干扰能力具有静态体制性,并主要取决于跳频的速度、伪随机跳频图案的难破译性,通信抗干扰决策还无法真正做到动态实时决策[2]。随着干扰技术的智能化,固定策略的静态抗干扰技术已无法取得动态抗干扰性能[3-4],不足以应对干扰样式智能演变的场景。

为使得抗干扰决策能有效应对动态干扰场景,各种深度强化学习技术被引入到抗干扰决策中[5],典型如深度 Q 学习网络[6-8](deep Q-learning network, DQN)、双深度 Q 网络(double deep Q-learning network, DoubleDQN)、对决深度 Q 网络(dueling deep Q-learning network, DuelingDQN)等深度强化学习算法。其中 DQN 算法是深度强化学习的典型算法之一,但收敛速度不理想。文献[9]在 DQN 算法

收稿日期:2023-02-02
* 基金项目:国家自然科学基金(61771256)项目资助

架构下引入动态 ε 机制,提出了动态 ε-DQN 智能决策抗干扰算法,通过决策网络状态来动态选择合适的 ε 值,最终提高了决策网络训练的收敛速度和抗干扰成功率。DoubleDQN 和 DuelingDQN[10]是 DQN 的两种典型改进算法。其中,DoubleDQN 利用双 Q 网络解决了 DQN 中 Q 值过估计的问题。DuelingDQN 通过引入对决网络有效提升了网络模型的收敛速度。文献[11]利用 DoubleDQN 有效地学习了不同节点的通信模式,并在没有先验知识的情况下获得了接近最优的性能。文献[12]提出了一种基于 DuelingDQN 的多用户强化学习的分布式动态频谱快速接入算法,以找到一种多用户策略,实现多信道无线网络中效用的最大化。深度强化学习中的决策网络一般为卷积神经网络,但卷积神经网络对时序信号的处理并不擅长,为此文献[13]将 DQN 与长短时记忆(long short-term memory, LSTM)[14]相结合,提出了 DRQN 决策算法,并用 9 种 Atari 游戏验证了该算法的优异性能。

由于动态干扰场景下智能抗干扰决策的性能取决于深度强化学习算法的训练收敛性及其最终稳态性能,寻求合适的深度强化学习算法是抗干扰智能决策的一个重要课题。鉴于 DoubleDQN 和 DuelingDQN 都能有效提升 DQN 算法的性能,本文将两种算法进行融合,在使用双 Q 网络的同时引入对决网络,在深度神经网络方面进一步引入了 LSTM 预处理层,最终提出了混合深度循环 Q 网络(mixed deep recurrent Q-learning network, MixDRQN)抗干扰决策算法。研究表明,该算法相较于传统 DQN[15]能够更好地提取干扰信号时频图的频率特征,并能充分挖掘输入信号的时间相关特性,显著提升了决策网络的训练收敛速度及抗干扰性能。

1　系统模型

1.1　信号传输模型

如图 1 所示,考虑一个合法通信用户对(发送-接收方)与干扰机(Jammer)对抗的场景,强化学习智能体根据当前时刻接收到的信号做出决策,指导发射机在下一时刻选择合适的载波频率发射信号。

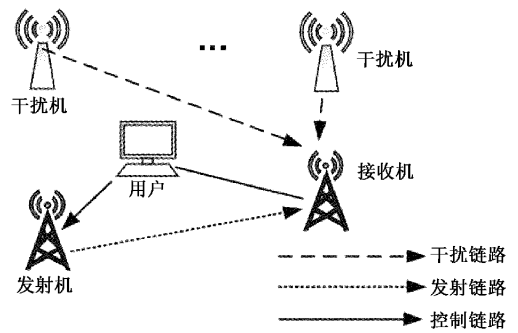


图 1　通信抗干扰模型

在跳频通信模式下,合法用户的发送方在可选频点集合中选择一个频点(用 $f_t \in [f_L, f_U]$ 表示,f_L 和 f_U 分别表示用户通信频带的起始和终止频率),以式(1)所示的给定的功率发送信号:

$$p_u = \int_{-b_u/2}^{b_u/2} U(f)\mathrm{d}f \tag{1}$$

由发射机(Transmitter)发送信号经信道传输至接收机(Receiver),其中 $U(f)$ 和 b_u 分别表示用户基带信号的功率谱密度函数(PSD)和带宽。而干扰机可以任意选择多个频率和波形的干扰信号来影响合法信号的有效接收。

用户与干扰机进行持续交互,利用环境反馈的奖励值,根据给定的决策算法,不断更新算法模型参数直至收敛,以抢占通信频带。考虑到用户信号、干扰和噪声的共存,智能体接收信号的 PSD 函数可以表示为:

$$S_t(f) = g_u U(f - f_t) + \sum_{j=1}^{J} g_j J(f - f_t^j) + n(f) \tag{2}$$

其中,g_u 和 g_j 分别表示信号和干扰在智能体接收端的增益,$J(f)$ 和 $n(f)$ 分别表示干扰和噪声的功率谱密度。

在实际频谱感知中,智能体接收端感知计算如下离散 PSD 样本值:

$$\boldsymbol{S}_t^i = 10\log\left(\int_{i\Delta f}^{(i+1)\Delta f} S_t(f + f_L)\right)\mathrm{d}f \tag{3}$$

其中,Δf 是频谱分析的分辨率。由式(3)可得频谱向量 $\boldsymbol{S}_t^i = \{s_t^1, s_t^2, \cdots, s_t^N\}(i = 1, 2, \cdots, N)$。

1.2　合法用户信号模型

在干扰对抗环境下,合法用户通过 PSD 环境矢量不断学习干扰方的干扰策略,因此,合法用户的抗干扰问题可以建模为马尔可夫决策过程(Markov decision process, MDP)。

将环境状态定义记忆长度为 T 的 PSD 环境矢量,也即 $\boldsymbol{S}_T$,其是大小为 $T \times N$ 的二维矩阵。$\boldsymbol{S}_T$ 的热力图表示一般称为频谱瀑布,包含频域和时域的能量信息。具体而言,$\boldsymbol{S}_T$ 可以写成以下的矩阵形式:

$$\boldsymbol{S}_T = \begin{bmatrix} S_{t-T+1}^1 & \cdots & S_{t-T+1}^N \\ \vdots & \ddots & \vdots \\ S_t^1 & \cdots & S_t^N \end{bmatrix} \tag{4}$$

合法用户根据 $\boldsymbol{S}_T$ 中的频谱向量 $\boldsymbol{S}_t^i$ 决策下一时刻跳频频点,并通过控制链路将结果传给发射机。

在抗干扰 MDP 中,假设合法用户的可用频点有 A 个,记为 $f_1, f_2, \cdots, f_A$,即动作空间可用定义为 $\mathcal{A} = \{f_1, f_2, \cdots, f_A\}$,用户在每个时刻选择一个动作 $a \in \mathcal{A}$ 进行跳频通信。

设合法用户的接收端 SINR 为 $\beta(f_t)$,合法信号在接收端能够成功恢复所需的 SINR 阈值为 β_{th},则归一化的信息传输率可以表示为 $\mu(f_t) = \delta(\beta(f_t) \geqslant \beta_{th})$,即当接收 SINR 超过阈值时,$\mu(f_t) = 1$,表示信息成功传输,否则传输失败。因此,即时奖励函数 r,可以定义为:

$$r(a_t)=\mu(a_t)-\lambda\delta(a_t\neq a_{t-1}) \tag{5}$$

式中：λ 表示用户中心频率切换的成本。

1.3　干扰信号模型

为从多角度论证本文所提抗干扰决策算法的普适性和有效性，本文基于跳频通信模式设置了两种通信场景[15-16]。

场景 1)设置可选频率集，用户只能在该集合中选择预设的频率进行跳频通信，且干扰信号也只能在预设的频率点上进行干扰。

场景 2)用户在可选频率集合中进行跳频通信，而干扰信号在不同时刻进行全频带干扰。

根据上述两种通信场景，本文设置了 3 种干扰信号：梳状干扰(comb)、扫频干扰(sweeping)和动态干扰(dynamic)，具体参数如下：

1)梳状干扰：干扰机选择 2、10 和 18 MHz 组成梳状谱干扰；

2)扫频干扰：干扰机按照频率大小顺序顺次干扰，扫频速度为 1 GHz/s；

3)动态干扰：干扰机每 100 ms 以 0.5 的概率随机选择梳状干扰或扫频干扰并持续 100 ms。

对于上述 3 种干扰类型，每个干扰信号的瞬时带宽设置为 4 MHz，发射功率为 30 dBm。对应的频谱瀑布如图 2 所示。

2　混合改进 DQN 决策算法

2.1　信号传输模型和用户信号模型设置

在信号传输模型中，对于式(3)，设置频谱分析分辨率 $\Delta f=100$ kHz，用户可用的频率范围 $f_t\in[0\ \text{MHz},20\ \text{MHz}]$，用户基带信号带宽 $b_u=4$ MHz，则 $N=(f_U-f_L)/\Delta f=20\ \text{MHz}/100\ \text{kHz}=200$，即 $\boldsymbol{S}_t^i=\{s_t^1,s_t^2,\cdots,s_t^{200}\}$。

对于用户信号模型，设置 $T=200$ ms，根据式(4)，频谱矩阵如下：

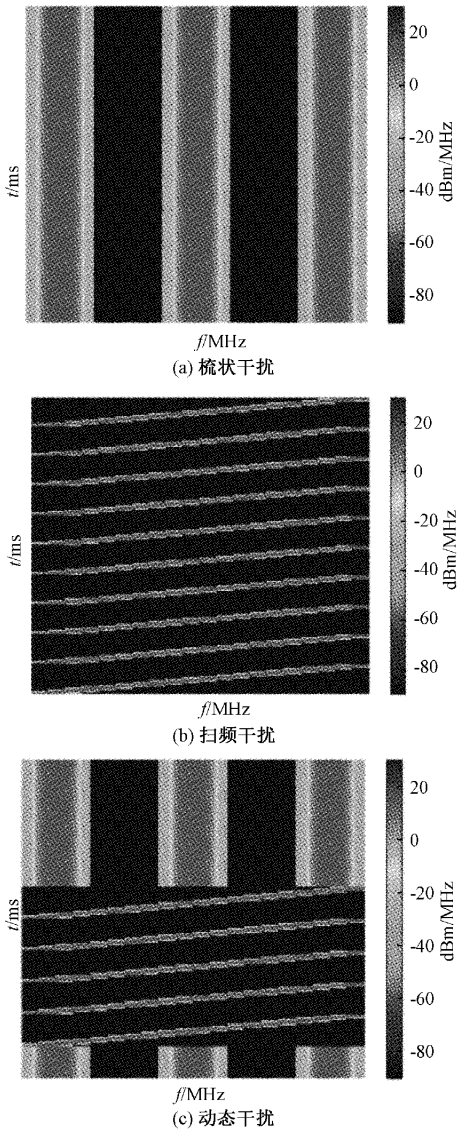
$$\boldsymbol{S}_T=\begin{bmatrix} s_{t-199}^1 & \cdots & s_{t-199}^{200} \\ \vdots & \ddots & \vdots \\ s_t^1 & \cdots & s_t^{200} \end{bmatrix}$$

用户信号的带宽是 4 MHz，中心频率以 2 MHz 的步长每 10 ms 变化一次，其对应的频谱瀑布如图 3 所示。

用户选择频率的带宽为 20 MHz，用户信号的带宽是 4 MHz，中心频率以 2 MHz 的步长每 10 ms 变化一次，因此对应智能体的动作空间大小为 9，即：

$a\in\{2\ \text{MHz},4\ \text{MHz},\cdots,18\ \text{MHz}\}$

设置用户中心频率的切换成本 λ 为 0.2，所以由式(5)可得 reward 有 4 个可能的值，分别是 1.0、0.8、0.0 和 −0.2。其中 1.0 表示该时隙频率未切换且成功避开干扰，0.8 表示该时隙频率切换且成功避开干扰，0.0 表示该时隙频率未切换且受到干扰，−0.2 表示该时隙频率切换且受到干扰。



(a) 梳状干扰

(b) 扫频干扰

(c) 动态干扰

图 2　干扰信号频谱瀑布

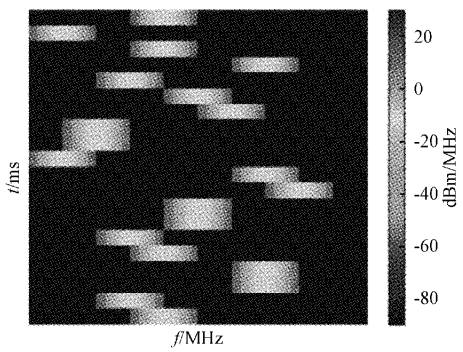


图 3　用户信号的频谱瀑布

2.2　LSTM 层

LSTM 是一种特殊的循环神经网络(recur-rent neural network，RNN)，可以利用历史信息对序列数据进行处理

和预测,并解决了 RNN 的长期依赖和梯度消失问题。LSTM 以序列数据为输入,在序列的前进方向上进行递归处理,可以处理序列结构的信息,并提取时间相关特征。LSTM 单元结构图如图 4 所示。

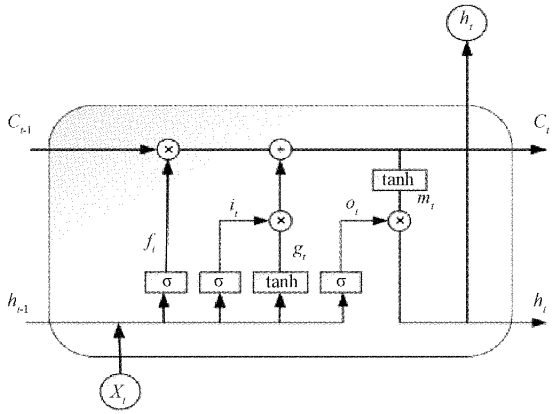


图 4　LSTM 单元结构

LSTM 网络在常规循环神经网络的基础上引入了遗忘门、输入门、输出门来控制 LSTM 单元的状态迭代。C_{t-1} 和 C_t 分别为 $t-1$ 时刻和 t 时刻的细胞记忆,h_{t-1} 和 h_t 分别是 $t-1$ 时刻和 t 时刻的细胞状态,本文中其大小为 $200\times(1\times200)$。$\boldsymbol{f}_t$ 本质上是一个由 0 和 1 组成的向量,与 C_{t-1} 按位相乘后,乘 1 的部分保留,乘 0 的部分遗忘。i_t 称为输入门,本文中输入 X_t 是大小为 $200\times(1\times200)$的干扰信号时频序列,会经过 tanh 函数对输入信息进行选择,避免全部信息输入造成的信息冗余。O_t 称为输出门,其与 t 时刻细胞状态 C_t 和 tanh 相乘,共同决定单元的输出信息。

图 4 中 f_t 部分叫做遗忘门,i_t 部分称为输入门,O_t 部分称为输出门,各个门的计算原理如下:

$$\boldsymbol{f}_t = \mathrm{sigmoid}(W_f X_t + W_f h_{t-1} + b_f) \tag{6}$$

$$i_t = \mathrm{sigmoid}(W_i X_t + W_i h_{t-1} + b_i) \tag{7}$$

$$g_t = \tanh(W_g X_t + W_g h_{t-1} + b_g) \tag{8}$$

$$C_t = C_{t-1} \times f_t + g_t \times i_t \tag{9}$$

$$o_t = \mathrm{sigmoid}(W_o X_t + W_o h_{t-1} + b_o) \tag{10}$$

$$h_t = o_t \times \tanh(C_t) \tag{11}$$

式中:W 均为权重,b 均为偏置。

2.3　MixDQN 算法

1)DuelingDQN 算法

DuelingDQN 引入了一种新的神经网络结构——对决网络,其输出包括两个分支,分别是状态价值函数(value,V)网络和每个动作的优势函数(advantage,A)网络。对决网络的结构如图 5 所示。

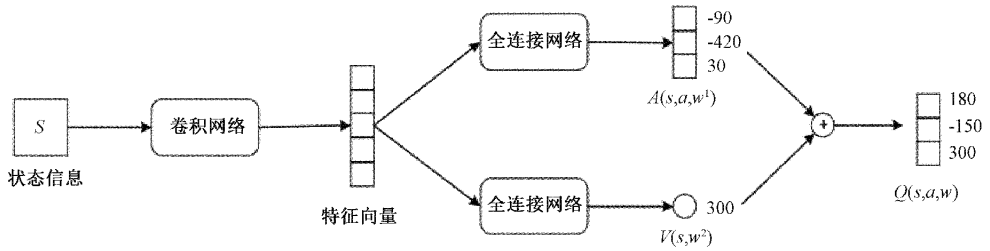


图 5　DuelingDQN 结构

网络的输出 Q 值计算公式如下:

$$Q(s,a;w^1,w^2) = V(s,w^2) + A(s,a,w^1) - \max_a A(s,a,w^1) \tag{12}$$

式中:w^1 和 w^2 分别是 A 和 V 两个网络的参数。

两个全连接网络 $A(s,a,w^1)$ 与 $V(s,w^2)$ 共享卷积层,这些卷积层把输入的状态 S 映射成特征向量,特征向量是 A 与 V 的输入。V 的输出是一个标量,A 的输出是一个向量,其维度是用户动作空间的大小 $|\ a\ |$,每个元素对应一个动作。为方便理解,举例如下。假设动作空间是 $\{L,R,U\}$,A 的输出是如下 3 个值:$A(s,L,w^1) = -90$,$A(s,R,w^1) = -420$,$A(s,U,w^1) = 30$。V 的输出是一个实数 $V(s,w^2) = 300$。首先计算 $\max_a A(s,a,w^1) = 30$,然后根据式(12),得到 $Q(s,L,w) = 180$,$Q(s,R,w) = -150$,$Q(s,U,w) = 300$,即每个动作的 Q 值。

对决网络的优势在于:对于不同动作,无需重新估计每个动作的 Q 值,引入优势函数后,对于新动作可以基于当前状态价值函数快速估计 Q 值。如上述例子,更新某动作 Q 值的同时也会更新其他动作的 Q 值,因此可以在更少的迭代次数里让更多的 Q 值得到更新。

2)DoubleDQN 算法

在 DQN 中,目标 Q 值通过 ε-greedy 算法选取最大 Q 值得到,所以会出现 Q 值估计偏高的问题。由于 DQN 是一种 off-policy 的策略,在每次学习时,没有使用下一次交互时的真实动作,而是使用当前策略认为的价值最大的动作,所以会出现对 Q 值的过高估计。为解决该问题,DoubleDQN 引入了双 Q 网络,对 Q 值计算过程中的动作选择做了调整,分为以下两步:

(1)在决策 Q 网络中找出最大 Q 值对应的动作,即 $a^* = \mathrm{argmax}_a Q(s_{j+1},a,w)$;

(2)用选择出来的动作 a^*,在目标网络中计算目标 Q 值,即 $y_j = r_j + \gamma Q(s_{j+1},a^*,w^-)$,以评估该动作的价值。其中 w^- 是目标网络的参数。

DoubleDQN 通过动作选择和动作评估的分别计算来避免过高估计的问题。

2.4　MixDRQN 算法

本文基于 LSTM 和 MixDQN 提出了混合深度循环 Q 网络决策算法(MixDRQN),其伪代码如算法 1 所示。

算法 1　MixDRQN 算法伪代码

1：初始化网络参数
2：*for epoch* $=1,2,\cdots,E$ *do*
3：*for* $t=0,1,2,\cdots,\infty$ *do*
4：　将混合信号频谱瀑布生成时间序列,输入 LSTM 层
5：　将 LSTM 层输出的数据输入到 MixDQN 网络中
6：　正向传播,得到策略网络中的 Q 值
7：　选择策略网络中最大 Q 值对应的动作：

$$a^{*}=\mathrm{argmax}_{a}Q(s_{j+1},a,\omega)$$

8：　将 a^{*} 代入目标网络中根据式(12)计算动作目标 Q 值

$$y=r(a^{*})+\gamma\max_{a}Q(s,a^{*};w^{1},w^{2})$$

9：　计算损失函数并进行梯度下降,更新 A 和 V 两个网络的参数 ω^{1} 和 ω^{2}
10：　*end for*
11：*end for*

MixDRQN 的输入是时频瀑布图 $\boldsymbol{S}_{T}$，鉴于 MixDQN 仅提取了频率特征,而没有充分挖掘时频瀑布图时间上的相关性,导致性能受限。为此,本文引入了 LSTM 结构层用于处理时间序列相关特性。

如图 6 所示,本文提出的 LSTM 网络由 2 个 LSTM 层组成。网络的输入为时频瀑布图,大小为 200×200,将其转化成 200 个 200×1 的时间序列作为输入向量。每一层 LSTM 由 200 个 LSTM 单元组成,LSTM 单元中隐藏层节点长度为 200。将 200 个向量输入到第 1 个 LSTM 层的 200 个 LSTM 单元中,第 2 个 LSTM 层的输入为第 1 层的输出。最后,将第 2 层最后一个时间节点的特征作为最终输出,将经过 LSTM 层处理过的 200 个 200×1 的向量拼接成一个大小为 200×200 的矩阵,并输入到卷积层处理。

LSTM 处理时间序列时,很难进行并行优化,且时间和内存开销会随神经元数量的增加而增大。此外,当层数过多时,层间梯度消失会加重,导致第 1 个 LSTM 层更新迭代放缓,收敛速度和收敛性能下降。综上考虑,本文采用 2 层 LSTM 层。

本文构建的基于 LSTM 和 MixDQN 的抗干扰决策网络的结构如图 7 所示。

其工作过程如下：

步骤 1)将大小为 200×200 的干扰与信号的混合采样时频图,按照频率轴分割,转化成 200 个大小为 200×1 时间序列向量,作为 LSTM 层的输入。输出是将经过 LSTM 层处理过的 200 个 200×1 的向量拼接成一个大小为 200×200 的矩阵,作为卷积层的输入。

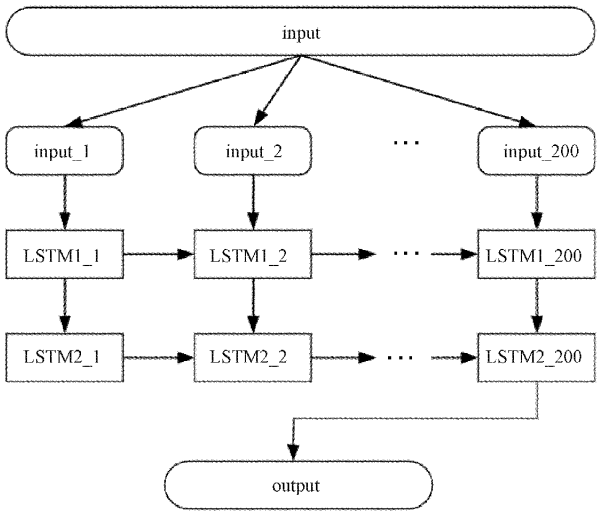


图 6　LSTM 层结构

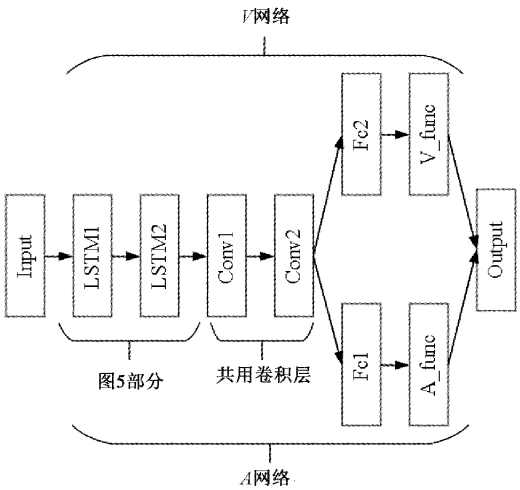


图 7　MixDRQN 深度学习网络模型

步骤 2)通过两次卷积操作,再通过一个 Flatten 层将多维数据展平成一维数据,作为两个全连接层的输入。

步骤 3)分别通过两个全连接层,再分别经 A 和 V 得到最终的输出,是大小为 9 的 Q 值向量。

图 7 中 Input 即图 6 中的输入,LSTM 层的输出即图 6 中的 Output。数据处理过程如图 5 所示,其中全连接层 Fc1 和 Fc2 参数完全相同,激活函数均采用 Sigmoid,网络结构参数如表 1 所示。

3　仿真实验

3.1　实验流程

基于 MixDRQN 的训练和测试以及网络模型的搭建、训练和测试均在 Linux 服务器上进行,语言环境为 Python3.7,CPU 为 Inter(R) Xeon(R) Silver 4210 2.80 GHz,GPU 为 NVIDIA Corporation GK210GL [Tesla K80]。

3.2　仿真条件

本文实验利用 Python 和 MATLAB 工具进行系统仿

表 1　网络结构参数

Layer	参数	输出	Activation
LSTM1	Cells: 200	200×200	Sigmoid,Tanh
LSTM2	Cells: 200	200×200	Sigmoid,Tanh
Conv1	Kernel: 8	16×50×50	ReLU
Conv2	Kernel: 4	32×25×25	ReLU
Flatten	—	20 000	
Fc1	—	256	Sigmoid
Fc2	—	256	Sigmoid
A	—	9	—
V	—	1	—

真和实验分析。实验中的场景是 2.3 节中的干扰模型。本文将 5 种抗干扰决策方法进行对比,分别是 DoubleDQN、DuelingDQN、DRQN、MixDQN 和 MixDRQN。其中,DRQN 是引入了 LSTM 层,但未采用双 Q 网络和对决网络的非混合网络结构;MixDQN 采用了混合网络结构,但未引入 LSTM 层;MixDRQN 则同时采用混合网络结构和 LSTM 层。实验参数设置如表 2,其中 ϵ_{max} 和 ϵ_{min} 分别是 ϵ 贪心算法 ϵ 取值的上下限。

表 2　实验参数

实验参数	数值
ϵ_{max}	0.9
ϵ_{min}	0.01
学习率 α	0.000 01
折扣因子 γ	0.95
batch_size	64
epoch	50
iteration	100

每次实验训练 50 个回合(epoch),每回合从经验回放池中采样一个批次(batch_size)大小的数据并进行 100 次迭代(iteration)。

3.3　性能仿真分析

本文中的抗干扰性能主要从两个方面衡量,分别是吞吐量和累积奖励。算法的收敛性能可以用误差函数衡量。实验中神经网络每更新一次网络参数,立即用当前模型去评估,再更新网络参数,再做评估,重复上述过程直至达到截止条件,这样设置可以看到模型的实时训练效果以及收敛速度。

将 5 次独立实验的结果取平均值得到实验对比图,由于最终算法均达到收敛,故为了便于观察只选取前 30 个 epoch 的曲线。

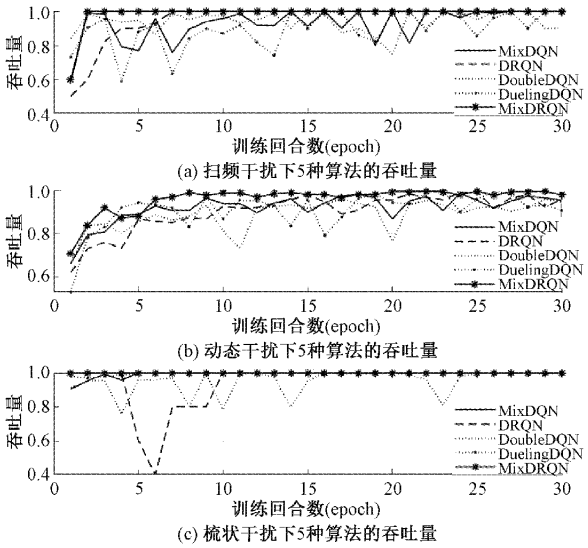
1)抗干扰性能分析

吞吐量表示每回合中抗干扰方躲避干扰并成功通信的次数,本文中采用归一化吞吐量 T_n 来衡量抗干扰性能。定义吞吐量为模型与干扰环境交互时 reward>0 的次数(reward>0 表示信号传输成功)。实验中为了方便观察与分析,采用归一化的吞吐量:将一个 epoch 中 reward>0 的次数除以总迭代次数,即:

$$T_n = \frac{N_{reward>0}}{N_{all}}$$

式中:$N_{reward>0}$ 表示一个 epoch 中 reward>0 的次数,N_{all} 表示一个 epoch 总迭代次数,本实验中 $N_{all}=100$。

如图 8 所示,是在 3 种干扰信号下不同抗干扰算法的吞吐量对比。



(a) 扫频干扰下5种算法的吞吐量

(b) 动态干扰下5种算法的吞吐量

(c) 梳状干扰下5种算法的吞吐量

图 8　不同抗干扰算法性能对比

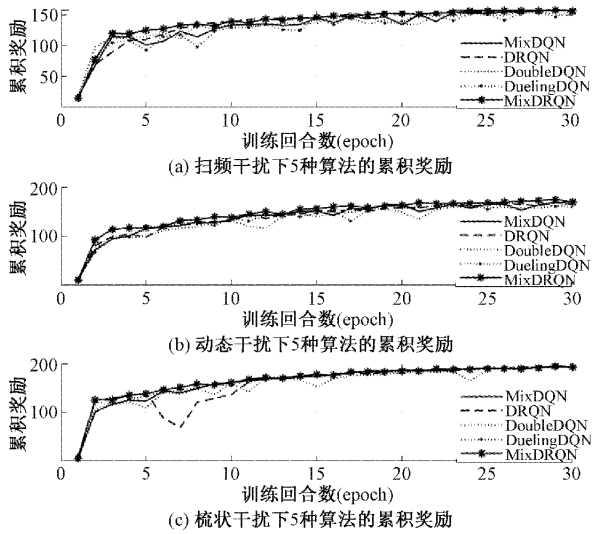
图 8 中从上至下分别对应扫频干扰、动态干扰和梳状干扰。以扫频干扰为例,DoubleDQN 和 DuelingDQN 的性能都差于其他 3 种,在训练 30 个 epoch 后才收敛。没有加入 LSTM 层处理的 MixDQN 算法性能有所提升,在训练 25 个 epoch 后模型才收敛。加入了 LSTM 层却没有用 MixDQN 的 DRQN 算法性能大幅提升,在第 7 个 epoch 即达到收敛。而所提方法 MixDRQN 性能最佳,在第 3 个 epoch 即收敛至最佳,性能较 DoubleDQN 和 DuelingDQN 提升 10 倍多,较 MixDQN 提升 8 倍多,较 DRQN 提升 2 倍多。对比其他干扰场景,MixDRQN 均取得最优效果,比其他四种算法至少提升 8 倍以上。

图 9 为 3 种干扰信号下,不同算法在训练过程中的累积奖励对比图。强化学习模型的训练以最大化累积奖励为目标,累积奖励越大,模型的整体受益越大,性能越好。本文中对累积奖励的定义为:训练 50 个 epoch,每个 epoch 进行 100 次 iteration,共计 5 000 次 iteration 所得 reward 的总和。

从图 9 中可以看出,MixDRQN 算法较其他 4 种算法均有较高的累积奖励,说明其性能最佳。

2)收敛性能分析

算法中的损失函数是目标 Q 值与预测 Q 值的均方误



(a) 扫频干扰下5种算法的累积奖励

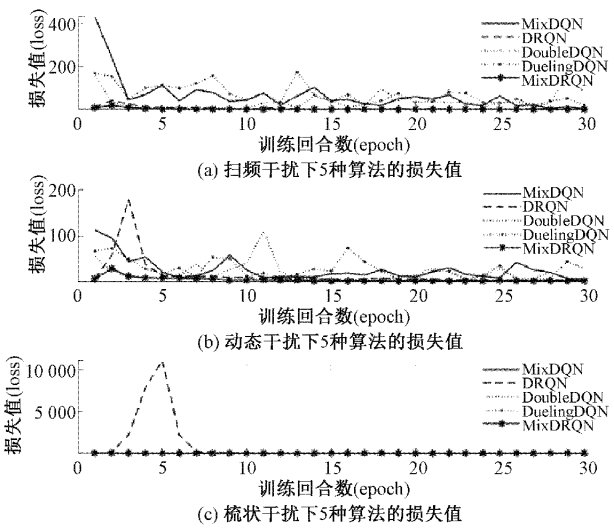
(b) 动态干扰下5种算法的累积奖励

(c) 梳状干扰下5种算法的累积奖励

图 9　训练时累积奖励对比

差，误差值越小，说明模型的预测越接近真实情况，算法收敛性能越好。为了比较不同算法的收敛性，对不同算法的误差函数进行对比。

不同干扰信号下，不同算法的误差函数曲线如图 10 所示。



(a) 扫频干扰下5种算法的损失值

(b) 动态干扰下5种算法的损失值

(c) 梳状干扰下5种算法的损失值

图 10　不同算法误差曲线对比

从图 10 中可以看出，以扫频干扰为例，DoubleDQN 和 DuelingDQN 均在 30 个 epoch 后才收敛；MixDQN 在第 25 个 epoch 后收敛；DRQN 在第 5 个 epoch 后收敛；而所提 MixDRQN 在第 2 个 epoch 后即收敛，性能大幅提升。对比其他干扰场景，MixDRQN 均取得最优效果。

4　结　　论

本文主要研究复杂电磁环境下跳频通信场景中的抗干扰决策问题。通过集成 DoubleDQN 和 DuelingDQN 两种决策机理的优点，并引入 LSTM 结构，提出了 MixDRQN 算法，用于多种干扰模式下的跳频通信抗干扰。仿真结果表明，在不同的干扰模式下，该算法表现出优异的训练收敛速度和抗干扰性能，适合动态抗干扰通信决策。

参考文献

[1] WANG, ZHANG F, ZHAO J, et al. Application of HBM2 data storage in time and frequency hopping network communication system[C]. Proceedings of the 2020 IEEE 6th International Conference on Computer and Communications, Chengdu, China, 2020: 1799-1803.

[2] 荆俊明，石建明，张秀蓉，等. 舞台载波通信自适应跳频抗干扰仿真分析[J]. 电子测量与仪器学报，2023，35(10)：145-152.

[3] PAN W, XIE L, XIA X, et al. A digital predistortion method for fast frequency-hopping systems[C]. 2019 IEEE International Symposium on Signal Process-ing and Information Technology(ISSPIT),2019: 1-3.

[4] LUO M. Analysis of wireless communication anti-jamming technology[J]. China New Telecommunications, 2020, 22(12):10-11.

[5] 杨鸿杰，张君毅. 基于强化学习的智能干扰算法研究[J]. 电子测量技术，2018，41(20)：49-54.

[6] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.

[7] SLIMENI F, CHTOUROU Z, SCHEERS B, et al. Cooperative Q-learning based channel selection for cognitive radio networks [J]. Wireless Networks, 2018, 25(4): 1-11.

[8] XU L, ZHAO Z. A distributed CRN resource allocation algorithm based on CBR and cooperative Q-learning [J]. Telecommunications Science, 2019, 35(2): 35-42.

[9] 宋佰霖，许华，蒋磊，等. 一种基于深度强化学习的通信抗干扰智能决策方法[J]. 西北工业大学学报，2021，39(3)：641-649.

[10] WANG Z, SCHAUL T, HESSEL M, et al. Dueling network architectures for deep reinforcement [C]. International Conference on Machine Learning. PMLR, 2016:1995-2003.

[11] XU Y, YU J, HEADLEY W C, et al. Deep reinforcement learning for dynamic spectrum access in wireless networks [C]. MILCOM 2018-2018 IEEE Military Communications Conference (MILCOM) IEEE, 2018: 207-212.

[12] NAPARSTEK O, COHEN K. Deep multi-user reinforcement learning for distributed dynamic spectrum access[J]. IEEE Transactions on Wireless

Communications, 2018, 18(1): 310-323.

[13] HAUSKNECHT M, STONE P. Deep recurrent Q-learning for partially observable mdps[C]. 2015 Aaai Fall Symposium Series,2015.

[14] MENG L, QU W, MA S, et al. Radar PRI modulation pattern recognition method based on LSTM[J]. Modern Radar, 2021, 43(1): 50-57.

[15] LIU X, XU Y, JIA L, et al. Anti-jamming communications using spectrum waterfall: A deep reinforcement learning approach[J]. IEEE Communications Letters, 2018, 22(5): 998-1001.

[16] LI Y, XU Y, XU Y, et al. Dynamic spectrum anti-jamming in broadband communications: A hierarchical deep reinforcement learning approach [J]. IEEE Wireless Communications Letters, 2020, 9 (10): 1616-1619.

作者简介

夏重阳(通信作者),硕士研究生,主要研究方向为通信对抗和强化学习。

E-mail:xia13218214523@163.com