基于目标导向行为和空间拓扑记忆的视觉导航方法

阮晓钢 李 鹏 朱晓庆 刘鹏飞

(北京工业大学信息学部 北京 100124)

(计算智能与智能系统北京市重点实验室 北京 100124)

摘 要 针对在具有动态因素且视觉丰富环境中的导航问题,受路标机制空间记忆方式启发,提出一种可同步学习目标导向行为和记忆空间结构的视觉导航方法.首先,为直接从原始输入中学习控制策略,以深度强化学习为基本导航框架,同时添加碰撞预测作为模型辅助任务;然后,在智能体学习导航过程中,利用时间相关性网络祛除冗余观测及寻找导航节点,实现通过情景记忆递增描述环境结构;最后,将空间拓扑地图作为路径规划模块集成到模型中,并结合动作网络用于获取更加通用的导航方法.实验在 3D 仿真环境 DMlab 中进行,实验结果表明,本文方法可从视觉输入中学习目标导向行为,在所有测试环境中均展现出更高效的学习方法和导航策略,同时减少构建地图所需数据量;而在包含动态堵塞的环境中,该模型可使用拓扑地图动态规划路径,从而引导绕路行为完成导航任务,展现出良好的环境适应性.

关键词 目标导向行为;深度强化学习;碰撞预测;时间相关性网络;空间拓扑地图;动作网络

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2021.00594

A Visual Navigation Method Based on Goal-Driven Behavior and Space Topological Memory

RUAN Xiao-Gang LI Peng ZHU Xiao-Qing LIU Peng-Fei

(*Faulty of Information Technology, Beijing University of Technology, Beijing* 100124)

(*Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing* 100124)

Abstract Everyone knows it is impossible for agents to reach the goal efficiently until it has sufficiently explored the environment or constructed cognitive model of the world, but the essential question is how to generate goal-driven behaviour. Organisms can spontaneously explore the environment with rare or deceptive reward and build map-like representation to support subsequent actions, such as finding food, shelters or mates. What we want to know is whether the robot can imitate such cognitive mechanism to complete navigational tasks? Obviously, relying on high precision sensors as a source to recall the structure of environment is not practical in real world, so we perceive the state space and learn control policy with visual inputs. And to deal with the problems stem from dimension disaster, the deep learning is also used in our method. The navigation systems developed in robotics can typically be divided into two classes: one reach the goal by encoding the structure of environment, it can utilize multiple sensor information as input and provide high-quality environment maps; and the other one is map-less approach, which maintain a control policy in the learning process and use it to finish goal reaching tasks, each of them has their pros and cons. In this paper, we proposed a visual navigation method which can learn goal-driven behavior and encode space structure synchronously. Firstly, in order to learn

收稿日期:2020-05-02;在线发布日期:2020-11-07.本课题得到国家自然科学基金(61773027)、北京市自然科学基金(4202005)、北京市教育委员会科技计划一般项目(KM201810005028)资助.**阮晓钢**,博士,教授,主要研究领域为自动控制、人工智能及智能机器人.E-mail:adrxg@bjut.edu.cn.**李 鹏**,博士研究生,主要研究方向为深度强化学习及机器人导航.**朱晓庆**(通信作者),博士,讲师,主要研究方向为智能机器人及机器学习.E-mail:alex.zhuxq@bjut.edu.cn.**刘鹏飞**,硕士研究生,主要研究方向为人工智能及机器人导航.

control policy from raw visual information, we take deep reinforcement learning as basic navigation framework, it provides an end-to-end framework and allow our approach directly predict control signal from high-dimensional sensory inputs. Meanwhile, due to the environment contains a much wider variety of possible training signals, an auxiliary task named collision prediction is added to the model. Then, in the process of exploration, the agent throughout the environment numerous times and observe a lot of states, but much of them are repetitive, the temporal correlation network is used to remove these redundant observation and search for waypoints. Because the various perspective of agent, instead of using hand-designed features, we use temporal distance, which only related to environment steps to compute the similarity between states. And inspired by the researches about cognitive mechanism of animals, we learned that many mammals are able to utilize an observation, especially the one include landmarks, to represent a neighboring state space, thus encoding the environment in a simpler and efficient way. So we use waypoints, which discovered in exploration sequences and can represent an adjacent state space that within a certain temporal distance, to describe the structure of environment gradually. Finally, the space topological map is integrated into the model as a path planning module, and combines with locomotion network to obtain a more general navigation method. The experiment was conducted in 3D simulation environment DMlab. The experiment results show this navigation method can learn goal-driven behavior from visual inputs, and show more efficient learning approach and navigation policy in all test environments, and reduce the amount of data required to build map. Furthermore, by placing the agent in dynamically blocked environment, the model can take advantage of topological map to guide detour behavior and complete navigational tasks, showing better environmental adaptability.

Keywords　goal-driven behavior; deep reinforcement learning; collision prediction; temporal correlation network; space topological map; locomotion network

1　引　言

动物，包括人类在内，在空间认知和行动规划方面具有非凡的能力，与其对应的导航行为也在心理学[1]和神经科学[2]中得到广泛研究. 1948 年，Tolman[3]提出“认知地图(cognitive map)”概念用于说明物理环境的内在表达，自此，认知地图的存在和形式一直饱受争议. 近年来，通过将电极放置在啮齿类动物脑中及研究其电生理记录，位置细胞(place cells)，网格细胞(grid cells)和头朝向细胞(Head-Direction cells，HD cells)[4]等多种有关环境编码的细胞得以被人们熟知. 在空间认知过程中，每种细胞有其特定功能，它们相互合作完成对状态空间的表达，各类细胞连接如图 1 所示[5]. 此外，还有证据表明海马体-内嗅皮层脑区不仅参与空间记忆，在规划路径中也具有重要作用.

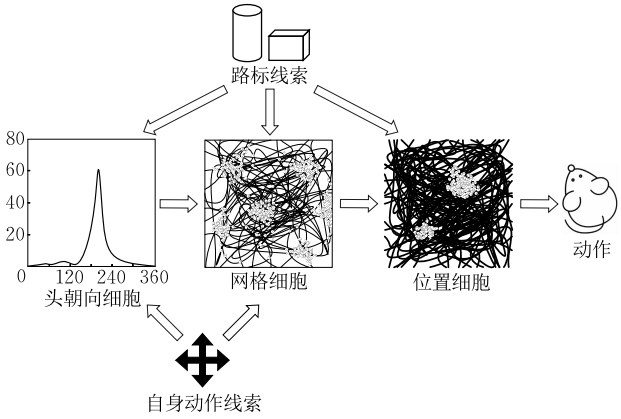


图 1　导航细胞连接图

相比之下，移动机器人导航系统通常以同步定位和建图(Simultaneous Localization and Mapping，SLAM)[6]为主要实现方式，该类方法可利用传感器数据(例如：激光、里程计、声呐、视觉等)并结合机器人自身运动信息构建未知环境的度量地图以实现自主导航. 在众多 SLAM 方法中，与本文工作最为相近

的是视觉 SLAM(Visual SLAM,VSLAM)模型[7],该方法主要以视觉感知环境信息,并通过摄像机姿态和多视角几何理论构建地图.为提高数据处理速度,一些 VSLAM 算法会优先提取图像特征点(例如:SIFT[8]、ORB[9]),然后通过匹配特征点完成帧间估计和闭环检测.基于 SLAM 的方法可提供高质量环境地图,但此类方法致力于位置推算和地图构建,往往需要额外的姿态或自身运动信息,且对动态环境缺乏适应性.

深度强化学习(Deep Reinforcement Learning,DRL)[10]由深度学习(Deep Learning,DL)[11]和强化学习(Reinforcement Learning,RL)[12]共同组成,它的出现在一定程度上推动了机器拟人化的发展.由于其具有端到端的学习框架,深度强化学习也被广泛应用于导航领域,并在高维空间中展现出良好的适应性.Zhu 等人[13]将预训练的 ResNet 与具有 siamese 架构的网络模型结合,实现以目标驱动的视觉导航,并在模型中增加目标适应性训练,使智能体对新目标具有更好的泛化能力.但这种方法本质上依赖于纯反应行为,在复杂环境中性能下降明显.Mnih 等人[14]提出一种基于策略的异步强化学习方法,并利用该方法训练结合长短时记忆网络(Long Short Term Memory Network,LSTM)的模型在 3D 迷宫中学习导航,实验结果表明该模型可存储环境相关信息并获得更加通用的控制策略.Jaderberg 等人[15]验证多种辅助任务对 CNN-LSTM 模型的影响,在 Atari 测试环境中,通过对 DQN(Deep Q-Networks,深度 Q 网络)[16]和 UNREAL Agent 的比较,进一步证明了 LSTM 的记忆功能.Mirowski 等人[17]构建一种具有堆叠 LSTM 架构的模型,在结合深度预测和闭环检测后,智能体学习速度和导航效率显著提高.同时在实验过程中,是否存在 LSTM 及 LSTM 层数对导航性能的影响也得到验证.模型中包含通用 LSTM 的系统可储存大量环境信息,即使是在回合间随机放置目标的 3D 环境中也能很好地完成导航任务.然而该类方法的控制策略只针对特定环境有效,当通路中出现堵塞或障碍物时,智能体需再次映射该路径,因此有很多研究人员试图通过对空间结构进行编码以更好地应对环境变化.于乃功等人[18]模仿海马结构空间认知机理构建细胞吸引子模型,从而实现构建精确环境认知地图.Parisotto 等人[19]使用二维记忆图储存环境信息,利用该抽象地图可完成路径规划任务.Gupta 等人[20]引入一种新颖的神经导航结构,该方法可从第一人称视角学习环境表征.Savinov 等人[21]则通过半参数拓扑记忆(Semi-Parametric Topological Memory,SPTM)构建未知环境的拓扑地图,并使用该地图驱使智能体寻找目标.以编码环境为导航实现方式的算法可通过构建空间类图表征引导目标导向行为,受堵塞和障碍物影响较小,但路径需针对每次任务进行规划,即使在全连通环境下也是如此,这无疑会降低算法的导航效率.

综上所述,深度强化学习为获取控制策略和编码环境结构提供了多种方法,本文在此基础上将两种导航形式结合,提出一种可在学习目标导向行为过程中构建空间拓扑地图的导航方法.其中,目标导向行为由具有深度强化学习架构的智能体在环境中学习所得,而拓扑地图则基于其情景记忆和观测之间的时间距离构建.运动网络是规划模块的补充,它可以帮助智能执行所规划的路径.

2 深度强化学习简介

深度强化学习将深度学习的视觉感知能力与强化学习的行动规划能力融为一体,构建了一种对视觉世界具有更高层次理解的端到端模型.在相关研究中,深度强化学习的基本架构包括 DQN[16]和深度递归Q网络(Deep Recurrent Q-Networks,DRQN)[22].

2.1 深度 Q 网络

DQN 是第一个被证明可在多种环境中直接通过视觉输入学习控制策略的强化学习算法,其模型如图 2 所示,输入为智能体观测到的连续 4 帧图像.

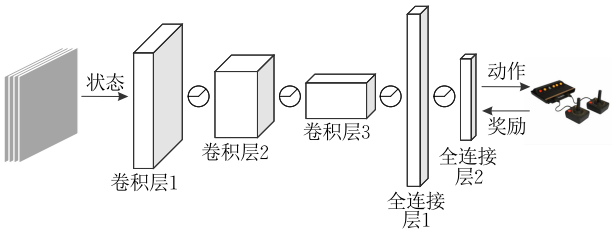


图 2 DQN 模型

标准强化学习算法假设智能体通过离散时间步与环境进行交互,其目标是学习回合内可最大化奖励的策略.在每一个时间步 t,智能体会根据当前状态 s_t 和策略 π 选择动作 a_t,在执行动作后获得奖励 r_t 并进入下一状态 s_{t+1},每一个时间步的回报 R_t 定义为累积折扣奖励:

$$R_t=\sum_{t'=t}^{T}\gamma^{t'-t}r_{t'} \tag{1}$$

其中,T 为回合最大时间步,t' 为当前时间步,t 为起始时间步,$\gamma\in[0,1]$为折扣因子,$r_{t'}$ 为当前时间步所获奖励.DQN 通过动作值函数 Q^{π} 学习控制策略,Q^{π} 定

义为给定策略 π 和状态 s 下执行动作 a 后的期望回报：

$$Q^{\pi}(s,a)=E[R_t|s_t=s,a_t=a] \tag{2}$$

其中，s_t 和 a_t 为当前时间步状态及动作. 在定义 Q^{π} 的同时定义最优动作值函数 Q^*，即 $Q^*(s,a)=\max\limits_{\pi} Q^{\pi}(s,a)$，借助贝尔曼方程可迭代更新动作值函数：

$$Q_{i+1}(s,a)=E_{s'}[r+\gamma \max\limits_{a'} Q_i(s',a')] \tag{3}$$

其中，s' 和 a' 为下一时间步状态及动作. 当 $i\to\infty$ 时，$Q_i\to Q^*$. DQN 使用参数为 θ 的非线性函数逼近器—卷积神经网络(Convolutional Neural Networks，CNNs)—拟合 Q 值. 此时同样可以利用贝尔曼等式更新参数 θ，定义均方误差损失函数：

$$L_t(\theta_t)=E_{s,a,r,s'}[(y_t-Q_{\theta_t}(s,a))^2] \tag{4}$$

其中，t 为当前时间步，$y_t=r+\gamma \max\limits_{a'} Q_{\theta_t}(s',a')$ 为目标，$Q_{\theta_t}(s,a)$ 为当前网络 Q 值. 通过微分损失函数可得梯度更新值：

$$\nabla_{\theta_t} L_t(\theta_t)=E_{s,a,r,s'}[(y_t-Q_{\theta_t}(s,a))\nabla_{\theta_t} Q_{\theta_t}(s,a)] \tag{5}$$

通过在环境中学习不断减小损失函数，使得 $Q(s,a;\theta)\approx Q^*(s,a)$. 实际上 DQN 并不是第一个尝试利用神经网络实现强化学习的模型，它的前身是神经拟合 Q 迭代(Neural Fitted Q-iteration，NFQ)[23]，其架构也与 Lange 等人[24]提出的模型密切相关. 而 DQN 性能之所以如此突出，目标网络和经验回放[25-26]有不可磨灭的贡献.

2.2　深度递归 Q 网络

DQN 已被证明能够在不同 Atari 游戏上从原始视觉输入学习人类级别的控制策略，正如它的名字一样，DQN 根据状态中每一个可能动作的 Q 值(或回报)选择动作，在 Q 值估计足够准确的情况下，可通过在每个时间步选择 Q 值最大的动作获取最优策略. 然而，由图 2 可知，DQN 的输入由智能体遇到的 4 个状态组成，这种从有限状态学习的映射，本身也是有限的，因此，它无法掌握那些要求玩家记住比过去 4 个状态更远事件的游戏. 当使用 DQN 在部分可见马尔可夫决策过程(Partially Observable Markov Decision Process，POMDP)[27]中学习控制策略时，由于无法结合过去的状态选择最优动作，DQN 在 POMDP 环境中的表现很不稳定. 为此，Hausknecht 等人[22]将具有记忆功能的 LSTM 与 DQN 结合，提出 DRQN 模型，其结构如图 3 所示.

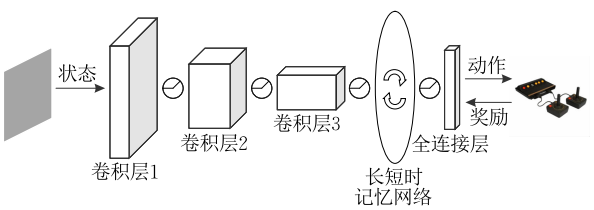


图 3　DRQN 模型

3　综合预训练模型

动作网络和时间相关性网络分别是执行绕路行为和构建拓扑地图的基础，两个网络在模型结构和训练方法上有很多相似之处，且都需要在智能体学习目标导向行为之前完成训练. 因此，构建综合预训练模型对两个网络同步进行训练，下面将对两个网络和训练模型进行详细介绍.

3.1　动作网络

动作网络被训练用于选取动作，这些动作可帮助智能体完成导航节点之间的移动，进而实现利用规划路径寻找目标. 动作网络以观测对 (o_i,o_j) 为输入，并以概率 $p(o_i,o_j)\in R^{|A|}$ 为输出，导航节点之间的动作可根据该概率选取.

在以图像作为输入进行预测的方法中，使用最为普遍的是帧间差方法，这是一种作用于像素级别的预测算法，其突出特点是实时性，但该方法始终存在准确率有限和易受干扰两个问题. 为提高预测精度及摆脱环境干扰物的影响，使用特征空间代替原始视觉感知作为网络输入. 由于动作网络是针对智能体观测之间的动作做出预测，因此可将网络编码的物体分为三类：(1) 可被智能体动作影响的物体；(2) 不受智能体动作影响，但其动作可影响智能体的物体；(3) 与智能体动作完全无关的物体. 本文致力于构建一个对(1)和(2)敏感，且不受(3)影响的特征空间，并利用其完成动作预测.

相较人为设计的特征，本文使用深度神经网络(Deep Neural Network，DNN)自动生成特征. 动作网络模型如图 4 所示，它具有端到端架构，在这种架构下特征不会与动作分离，而是在一起相互学习，从

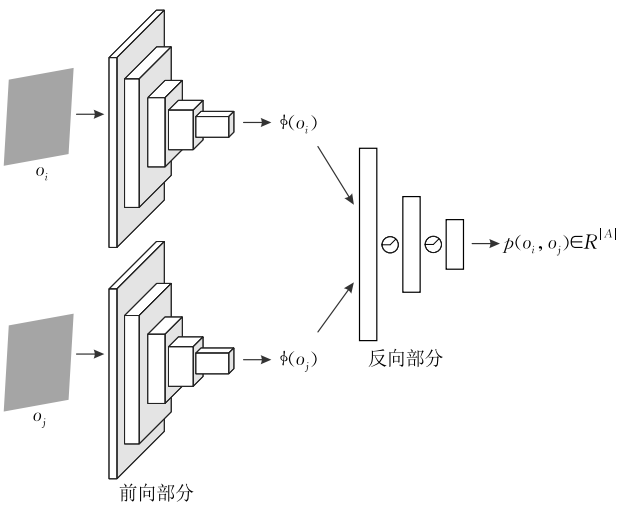


图 4　动作网络模型

而确保特征不会对任何不能影响或不受智能体动作影响的物体进行编码.动作网络模型中包含前向和反向两部分,其中,前向部分是基于 ResNet-18[28] 的深度卷积编码器,可将原始观测(o_i,o_j)编码为特征向量$[\phi(o_i),\phi(o_j)]$;反向部分则以特征向量作为输入并计算动作概率.

3.2 时间相关性网络

时间相关性网络的目标是通过时间距离寻找情景记忆中的导航节点,这对于避免存储冗余观测和构建拓扑地图至关重要.同时,本文的视觉感知任务(包括智能体定位及目标检测)也由时间相关性网络实现.

在探索过程和随后的目标导向行为中,智能体会多次遍历环境并储存大量情景观测数据.通过阅读有关哺乳动物空间认知方式的研究,了解到哺乳动物可利用一个观测,特别是包含路标的观测,映射一个邻近空间,以此高效认知环境[29].本文的拓扑地图构建方法也是受此启发,判断观测是否邻近可通过图像特征相似度法实现,但由于智能体视角的多变性,导致该类方法并不能很好地显示观测是否邻近.因此,为降低环境特征对算法性能的影响,舍弃图像相似度方法,采用在情景记忆中得到广泛研究的时间距离[21]判断观测是否邻近.

从概念上讲,时间相关性网络可被看成一个分类任务,它给予时间上邻近的观测较高的相似值,而给予时间上远离的观测较低的相似值.由于观测序列的连续性,较短的时间距离必然导致相邻的观测,且时间距离只与观测之间的步长有关,不受图像特征影响.时间相关性网络模型如图 5 所示,它包含嵌入和比较两部分:嵌入部分用于抽象化视觉输入$(o_i\rightarrow R^n_{o_i})$,其结构基于 ResNet-18;比较部分以特征作为输入并计算时间相关系数:

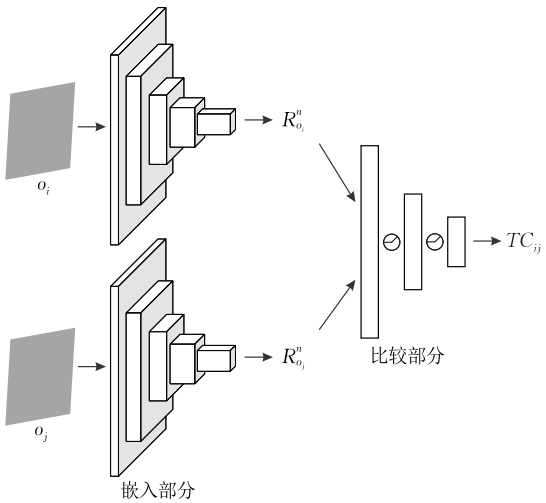


图 5 时间相关性网络模型

$$tc(o_i,o_j)=TC(E(o_i),E(o_j)) \tag{6}$$

其中,$tc(o_i,o_j)\in[0,1]$为 o_i 和 o_j 之间的时间相关系数,$E(o_i)$为观测 o_i 特征化过程,$TC(x,y)$用于计算特征间的时间相关系数.

3.3 训练模型

由 3.1 节和 3.2 节可知,动作网络和时间相关性网络有很多相似之处.第一,两个网络都使用 siamese 架构学习特征和进行预测,其卷积部分全部基于 RseNet-18.第二,虽然两个网络所使用的训练样本具有不同的形式,但其原始数据来源于同一随机探索环境的智能体.第三,两个网络都以自监督学习为训练方式,且使用相同训练方法和超参数.最后,对 R-network[30] 不同部分重要性的研究更是促使我们将两个网络放在同一模型中进行训练.考虑到特征对预测的影响,舍弃时间相关性网络的嵌入部分,保留动作网络的前向部分,并使用动作预测误差构建特征,综合预训练模型如图 6 所示.

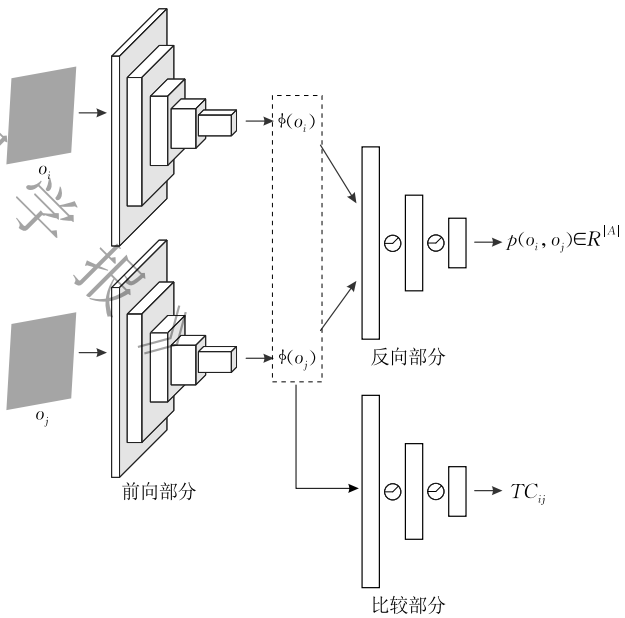


图 6 综合预训练模型

在使用该模型进行训练时,两个网络的损失函数分别计算.其中,动作网络通过监督学习进行训练,并使用交叉熵作为损失函数.训练样本形式为$((o_i,o_{i+k}),a_i)$,动作 a_i 对应式中第一个观测 o_i,该样本以情景记忆$\{o_1,o_2,\cdots,o_N\}$和动作序列$\{a_1,a_2,\cdots,a_N\}$为原始数据,并使用 k 个时间步分割而成.网络训练被定义为学习函数 L:

$$\hat{a}_i\Leftrightarrow p=L(o_i,o_{i+k};\theta_L) \tag{7}$$

其中,$\hat{a}_i$ 是 a_i 的预测值,p 为动作预测概率,o_i 和 o_{i+k} 为相隔 k 个时间步的两个观测,网络参数 θ_L 通过

式(8)进行优化：

$$\min_{\theta_L} loss(\hat{a}_i, a_i) \tag{8}$$

其中，$loss$ 用于衡量预测动作与实际动作之间的差异. 通过以随机运动的智能体轨迹作为原始训练数据，可习得有效的动作条件分布 $P(a|o_i, o_{i+k})$.

时间相关性网络的训练样本由两个观测和一个二进制标签组成：$\langle o_i, o_{i+k}, y_{ik} \rangle$，数据同样来源于随机探索环境的智能体. 如果两个观测值之间至多相隔 k 个时间步，则认为它们邻近($y_{ik}=1$). 负样本由两个至少相隔 $M \cdot k$ 个时间步的观测组成，M 用于扩大正负样本间差异. 最后，利用逻辑回归作为损失函数并输出邻近概率.

4　导航方法

智能体与新环境的交互分为两个阶段：在第一阶段内，智能体随机探索环境，并使用收集到的数据训练动作网络和时间相关性网络；在第二阶段内，智能体同步学习目标导向行为和构建空间拓扑地图，并将二者结合用于完成导航任务.

4.1　目标导向行为

目标导向行为可看作智能体在回合内学习最大化奖励策略时的副产物，而具有深度强化学习架构的系统更是在该领域取得了最先进的成果，所以本文模型也以深度强化学习为基本导航框架，并增加额外输入和辅助任务以提升学习效率.

为使智能体更高效地学习目标导向行为，导航框架以 DRQN 模型为基础，并针对本文任务做出以下调整：(1) 由于导航过程中使用辅助任务提升智能体学习效率，多余的卷积层会增加模型训练难度，因此将 DRQN 模型中卷积层由 3 层减少至 2 层；(2) 为缓解辅助任务带来的额外计算压力，对训练数据进行降维处理，即将 DRQN 模型中第一层和第二层卷积输出的 32 张和 64 张特征图分别减少至 16 张和 32 张特征图. 改进后的导航模型如图 7 所示，其输入包括：观测 $o_t \in R^{3\times W\times H}$(其中 W 和 H 为图像的宽度和高度)、上一时间步动作 $a_{t-1} \in R^{|A|}$ 和奖励 $r_{t-1} \in R$. 同时，使用模型后端分离的线性层计算策略 π 和值函数 V.

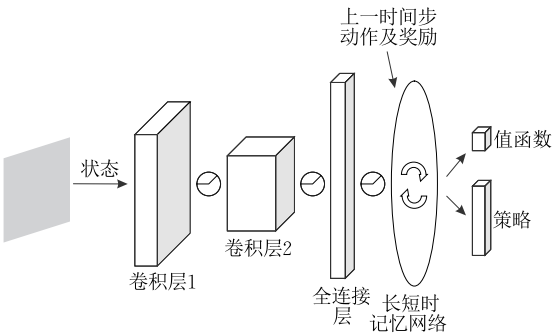


图 7　基本导航框架

在训练方法上，没有直接利用 DQN 所依赖的动作值函数 $Q(s_t, a_t; \theta)$ 和均方误差学习导航，而是使用异步优势 Actor-Critic(A3C)[14] 算法在给定状态 s_t 的情况下学习策略 $\pi(a_t | s_t; \theta)$ 和值函数 $V(s_t; \theta)$. 且在整个训练过程中，除仿真环境内可获得的奖励(苹果、目标)外，不增加动作或碰撞惩罚，所用奖励函数如式(9)所示：

$$R_{t:t+n} = \sum_{i=1}^{n} \gamma^i r_{t+i} + \gamma^n V(s_{t+n+1}, \theta) \tag{9}$$

其中，$R_{t:t+n}$ 为包含 n 个时间步的累积折扣奖励，r_{t+i} 为当前时间步所获奖励，$V(s_{t+n+1}, \theta)$ 为环境终端网络值函数，在损失函数中使用熵正则化处罚代替均方误差：

$$L_{A3C} \approx L_{VR} + L_\pi - E_{s\sim\pi}[\alpha H(\pi(s, \cdot, \theta))] \tag{10}$$

其中，$L_{VR} = E_{s\sim\pi}[(R_{t:t+n} + \gamma^n V(s_{t+n+1}, \theta^-) - V(s_t, \theta))^2]$，$L_\pi = -E_{s\sim\pi}[R_{1:\infty}]$，$H(\pi(s, \cdot, \theta))$ 为策略 π 的熵，α 为熵系数. 在训练过程中，多个智能体与多个环境并行交互. 尽管后续实验证明该模型可从原始视觉输入中学习目标导向行为，但部分数据显示智能体学习效率与拓扑地图构建速度密切相关. 也就是说，导航策略越快趋于稳定，地图就越快覆盖整个空间. 因此，为提高智能体学习效率和减少构建地图所需数据量，在模型中结合一个名为碰撞预测的辅助任务，其实现方法如图 8 所示.

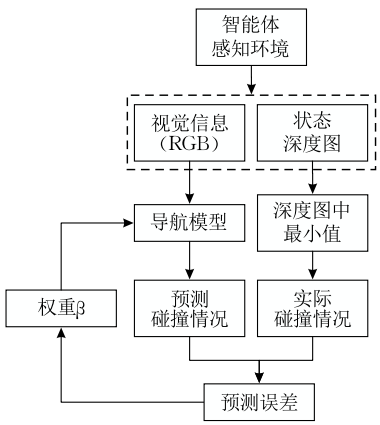


图 8　碰撞预测数据处理流程

其中，碰撞概率由 LSTM 后的单层感知器输出，预测误差 L_{CP} 通过实际和预测情况比较所得，并结合权重 β 应用于损失函数：

$$L_{A3C} \approx L_{VR} + L_\pi + \beta L_{CP} - E_{s\sim\pi}[\alpha H(\pi(s, \cdot, \theta))] \tag{11}$$

其中，L_{VR}、L_π 及 H 计算方法与式(10)相同. 不难发

现,本文模型中使用的辅助任务实际上利用了空间深度信息.但与大多数算法不同,我们没有将深度图直接作为模型输入以寻求更好效果,而是以损失函数的形式呈现环境结构信息,并利用其提供的密集训练信号加速引导学习.此外,碰撞预测为在线(对于当前帧)辅助任务,不依赖任何形式的回放机制.

4.2 空间拓扑记忆

拓扑地图是一种记忆空间结构的方法,文中使用导航节点对其进行填充.在每一探索回合结束后,结合时间相关性网络和智能体观测序列对地图进行更新,从而实现利用情景记忆递增地描述状态空间.

构建拓扑地图包括两个阶段:(1)初始阶段.此时模型内没有任何有关环境的记忆,输入的观测序列将作为智能体对环境的第一认知,因此需简化序列本身.假设智能体在环境中运行 T 个时间步得到情景记忆$(o_1,o_2,\cdots,o_T)$,以首次简化为例,通过时间相关性网络计算序列内第一个观测 o_1 与其他观测 o_i 的时间相关系数:

$$tc^1(o_1,o_i)=TC(E(o_1),E(o_i)) \quad (12)$$

其中,tc^1 为第一次简化的时间相关系数,$i=2,3,\cdots,T$.根据阈值 tc_t,省略与 o_1 邻近的观测,简化示意图如图 9 所示.这是简化的第一次迭代,观测 o_1 将作为第一个导航节点 w_1 储存在拓扑地图中,然后使用随后的观测和同样的方法持续简化序列直到最后一个观测.

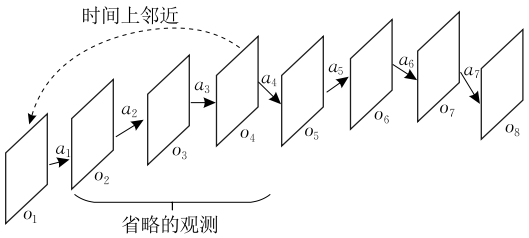


图 9 情景记忆简化示意图

简化过程按情景记忆内观测的先后顺序进行,所以地图中的导航节点递增储存且在理论上连通.但在规划路径时,需考虑动作网络的预测能力,因此,使用式(13)检测导航节点是否可达:

$$e_{ij}=1\Leftrightarrow TC(E(w_i),E(w_j))>tc_r \vee |i-j|\leq k \quad (13)$$

其中,e_{ij} 为导航节点间连接关系,w_i 和 w_j 为地图中导航节点,$tc_r\in(0.5,tc_t)$ 为可达性阈值,i 和 j 为导航节点脚标,式中包含时间距离和空间关系两种判别方法.

(2)扩张阶段.此时模型中已包含部分环境拓扑地图,智能体需通过集成每个观测序列不断扩充地图.因此,当前情景记忆$(o_1,o_2,\cdots,o_T)_c$ 中的每一个观测都需要与地图中的每一个导航节点进行比较以得到它们之间的时间相关系数:

$$tc^c(o_i,w_x)=TC(E(o_i),E(w_x)) \quad (14)$$

其中,tc^c 为当前情景记忆与拓扑地图间的时间相关系数,o_i $(i=1,2,\cdots,T)$ 为当前序列中的观测,w_x $(x=1,2,\cdots,n)$ 为拓扑地图中的导航节点.如果当前情景记忆中的观测全部与拓扑地图邻近,则不需要更新地图.相反,如果在当前序列中的观测,即使只有一个观测不能使用拓扑地图进行映射,那么该观测将作为新的导航节点添加到地图中.此时,需要创建与其对应的连接:

$$e_{ix}=\begin{cases}1, & TC(E(o_{i-1}),E(w_x))>tc_t\\ 0, & i=1\end{cases} \quad (15)$$

其中,o_{i-1} 为 o_i 前一时间步的观测,o_i $(i\in[2,T])$ 为当前情景记忆中新发现的导航节点,w_x $(x\in[1,n])$ 为拓扑地图中的导航节点,tc_t 为邻近阈值.

4.3 导航流程

导航任务以回合制进行,每个回合持续固定的时间步或直到找到目标为止.在回合内,智能体起始位置固定,通过目标导向行为或规划的路径完成导航任务.

由于控制策略在无障碍环境中获得,因此当不确定环境中是否存在堵塞时,可使用具有目标导向行为的智能体进行试探性导航.如果智能体在一定时间步内到达目标,则证明环境中没有堵塞,导航任务可通过该策略完成.相反,如果智能体在一定时间步内无法接触目标,则证明环境中存在堵塞,单纯的目标导向行为已不再适用,导航任务需结合拓扑地图和路径规划完成.

在重新规划路径之前,需确定智能体停滞和目标所属导航节点,并将它们作为路径的起点和终点.该视觉感知过程由时间相关性网络实现:对于确认智能体停滞位置,可使用当前观测与拓扑地图内导航节点进行比较,并根据时间相关系数确定智能体所属导航节点;对于目标检测,本文仿真环境中的目标有其固定的形状和颜色,并可在学习目标导向行为中收集获得,利用该类图片和时间相关性网络可定位目标位置.定位方法如图 10 所示,图中黑色圆

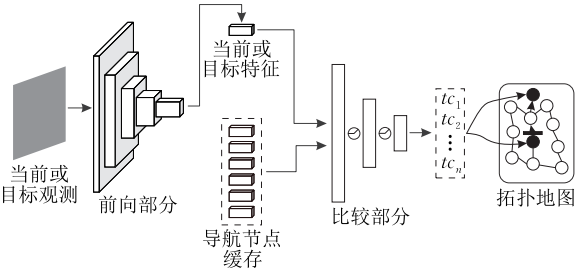


图 10 定位方法

形分别代表当前目标和智能体所属导航节点，黑色线段代表堵塞位置.

在得到起始和目标位置后，根据迪杰斯特拉算法[31]寻找导航节点 w^a 和 w^g 之间的最优路径：

$$\langle w_0^r, w_1^r, \cdots, w_n^r \rangle,\ w_o^r = w^a,\ w_n^r = w^g \quad (16)$$

其中，w^a 为起始节点，w^g 为目标节点. 然而从图 10 可以看出，由于拓扑地图是在全连通环境下构建的，规划的路径(黑色路径)可能包含跨越堵塞的连接，而这在实际导航中并不可行. 类似的不可用连接应被发现，并避免在接下来的路径规划中使用. 因此，一旦发现智能体长时间停留在一个位置，就证明路径中包含跨越堵塞的连接. 此时，应将该连接的路径代价设置为无穷大，并使用修正的拓扑地图重新规划路径. 由于导航节点之间相互连接，且环境中的堵塞可能不止一处，所以路径规划是一个迭代调整的过程，整个导航流程如图 11 所示.

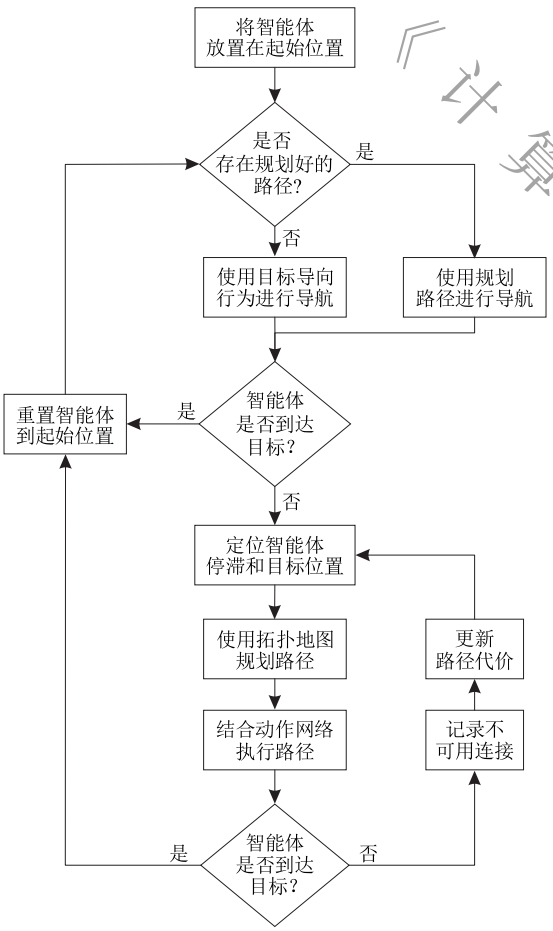


图 11　导航流程

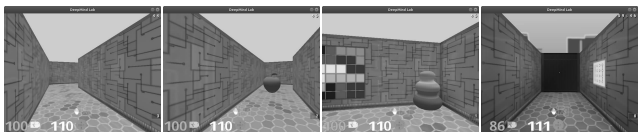
5　实　验

实验中通过定性导航任务评估本文模型，并与相关基线方法进行比较. 学习过程主要以奖励/时间(reward/time)图的形式呈现，图中每一时间点对应的奖励值为一小时内(虚拟时间)智能体所获奖励与完成回合数的平均值，智能体在每个回合内执行 4500 步动作.

5.1　实验设置

5.1.1　实验平台

实验在 3D 仿真环境 DMlab 中进行[32]，平台运行示意图如图 12 所示. 在该环境内，智能体执行离散动作，可实现小范围转向，加速前进后退或转弯. 奖励由智能体在迷宫中接触目标获得，在接触目标后，智能体将被重置到起始位置. 每个回合都提供充足的时间步，保证智能可多次到达目标. 仿真环境以 60 帧/秒的速度运行，并在环境中放置奖励刺激探索行为，其中，苹果奖励为+1，目标奖励为+10.



(a) 前进　(b) 苹果　(c) 目标　(d) 门

图 12　仿真环境

5.1.2　基线方法

在学习目标导向行为实验中，使用在深度强化学习领域具有代表性的前馈(FeedForward，FF)[16]模型和结合长短时记忆网络(LSTM)[22]的模型与本文模型(Navigation+Collision Prediction，Nav+CP)进行比较. 其中，FF 模型由 3 层卷积和 1 层全连接构成，每一层后都配有 ReLU 非线性单元，策略和值函数由单独的输出层计算. LSTM 模型结构与 FF 模型类似，只是在全连接层后增加 1 层 LSTM. 此外，没有结合碰撞预测的本文方法(Nav)也在迷宫中进行测试.

在具有动态堵塞的环境中进行测试时，共有三种方法用来验证拓扑地图在模型中的作用. 其中，第一种只使用学习到的导航策略(Nav+CP)在环境寻找目标，第二种将目标导向行为与空间拓扑地图相结合(Nav+CP+Space Topological Map，Nav+CP+STM)用于完成导航任务，第三种是基于智能体坐标位置(x, y)的最优路径(Optimal Path，OP)方法，由于使用了环境特权信息，这种方法可将环境直接离散化为二维地图从而选择最优路径.

5.1.3　模型实现细节

由图 4 可知，本文导航模型与 FF 和 LSTM 模型结构不同，它是由 2 层卷积、1 层全连接和 1 层 LSTM 组成. 其中，第一层卷积核尺寸为 8×8，跨度为 4×4，输出 16 张特征图；第二层卷积核尺寸为 4×4，跨度为 2×2，输出 32 张特征图；全连接层具

有 256 个神经元,前三层神经网络都配有 ReLU 非线性单元.在得到卷积编码器输出后,将其与智能体上一时间步的动作和奖励串联作为 LSTM 层的输入,LSTM 层与全连接层具有相同神经元数且配有遗忘单元[33].策略和值函数由 LSTM 层输出线性预测所得,碰撞概率则由单层感知器预测所得.

综合预训练模型的输入为两个观测,两者都要经过 ResNet-18 编码器处理并生成 512 维特征向量.在模型内部,动作网络首先将两个观测的特征串联,然后结合 2 层全连接(每层具有 256 个神经元)和 Softmax 层输出动作概率.而时间相关性网络则是分别对两个观测的特征进行处理,并使用 4 层全连接(每层具有 512 个神经元)计算两个观测是否邻近.除输出层外,每一层神经网络后都配有 ReLU 非线性单元.

5.1.4 超参数

为展现各模型在视觉丰富环境中的导航性能,不对智能体观测进行黑白化预处理,而是直接以 84×84 RGB 图像作为模型输入.与之前方法[16,22]相比,彩色图像可提供更多环境信息,但也在一定程度上增加了模型训练难度.学习过程中,借鉴文献[14]中所介绍的 A3C 范例引导强化学习,使用 8 线程及没有动量和方差干预的 RMSProp 算法训练神经网络,每个动作依然重复 4 次.学习率从$[1\times10^{-4},5\times10^{-3}]$区间内按对数均匀分布采样,熵代价从$[5\times10^{-4},1\times10^{-2}]$区间内按对数均匀分布采样,权值 β 从$[1\times10^{-2},1\times10^{-1}]$区间内按对数均匀分布采样.

综合预训练模型的输入是分辨率为 160×120 像素的两个 RGB 图像,该训练数据通过一个随机探索环境的智能体产生.训练过程中使用学习率 $\lambda=0.0001$ 的 Adam 优化器[34]进行学习,近期数据储存在容量为 B 的缓冲区,每次从缓冲区中随机采样 64 对观测更新网络参数.

5.2 参数选择实验

本文致力于研究更加通用的导航策略,在测试模型性能之前,需预先设定一些参数.这些参数主要涉及两方面:一是辅助任务中的碰撞阈值;另一个是动作网络和时间相关性网络中的训练细节,这些参数将在如图 13 所示的迷宫中进行定性确认.

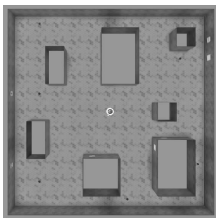
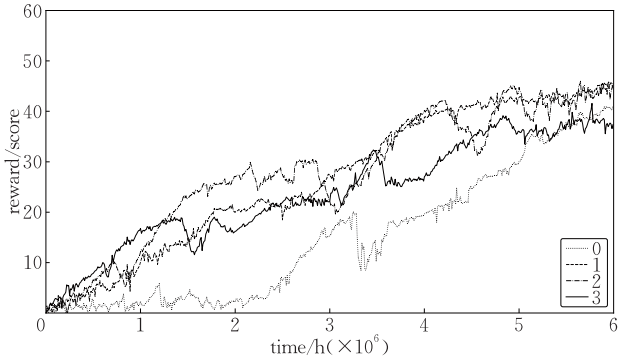


图 13 参数选择环境

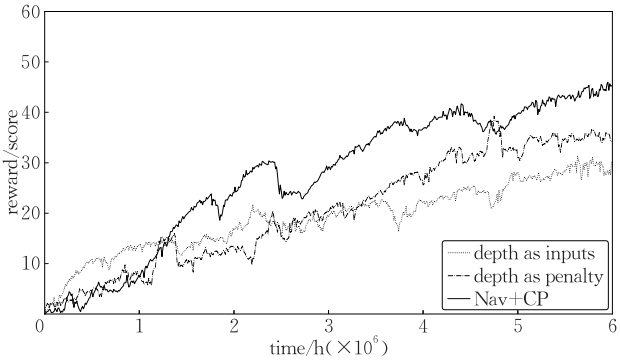
5.2.1 碰撞阈值实验

在本文导航模型中,使用一种名为碰撞预测的辅助任务,而碰撞发生与否取决于智能体与障碍物之间的最小距离.因此,需确定不同约束值对导航性能的影响,同时对利用不同类型深度信息的导航方法进行比较.在测试期间,只执行目标导向行为,不构建环境地图.

实验结果如图 14 所示,数据为 5 个具有最佳性能的智能体平均所得.从图 14(a)可以看出,当阈值在[0,3]内采样时,智能体探索效率和学习效果各不相同.如果阈值设置为 0,也就是说,只有在智能体撞到障碍物后才认为碰撞发生,会导致探索效率低下.相反,如果阈值较大,智能体则会过早执行避障动作,从而间接干扰导航策略,导致需要更多的时间步才能到达目标.当阈值为 1 或 2 时,智能体不仅可有效躲避障碍物,还能保持高效的目标导向行为.然而,考虑到策略稳定性问题,碰撞阈值在本文中设置为 1.



(a) 不同碰撞阈值实验结果



(b) 不同类型深度信息方法实验结果

图 14 碰撞阈值实验结果

从图 14(b)可以看出，当智能体直接将深度信息作为碰撞判别依据时，在探索环境过程中可有效躲避障碍物，但此时碰撞预测误差仅用于动作惩罚，对导航策略没有实质性的帮助. 通过将深度图作为输入，智能体同样可学习到控制策略，并使用其快速遍历环境，但这种方法忽略了环境的颜色特征，使智能体无法进一步理解状态空间. 而对于 Nav+CP 模型，虽然其探索效率不如将深度图作为输入的方法，但辅助任务的使用给予智能体更多的环境结构信息，从而实现更高效的目标导向行为.

5.2.2　分割阈值实验

在训练时间相关性网络过程中，需要时间步间隔 k 分割正负样本，动作网络的训练样本同样使用阈值 k 进行区分. 由于网络性能与 k 值的选取密切相关，现将不同分割阈值对动作网络、时间相关性网络和导航节点所占比例的影响总结为表 1，表中数据为 5 个具有最优超参数的智能体平均所得. 由表 1 可知，随着 k 的增加，动作网络性能逐渐下降，尤其是在 $k>4$ 后，预测精度下降明显. 时间相关性网络训练效果与正负样本间的差异成反比，在起始阶段，由于正负样本几乎邻近，故时间相关性网络的预测准确率较低，而随着 k 值的增加，其性能同步提升. 但当 $k>4$ 后，由于神经网络本身的限制，时间相关性网络的预测能力再次下降. 此外，虽然测试环境特征较为单一，但时间相关性网络准确率依然可达到 90%以上，表明其性能并不依赖于环境特征. 导航节点所占比例在理论上与 k 值成正比，即相隔的时间步越大，导航节点所占比例越低，在实验数据中也体现出相似规律. 但由于时间相关性网络影响，在 $k>5$ 后导航节点比例有所增加. 由于时间相关性网络为本文视觉感知实现方式，涉及智能体定位和目标检测，其预测精度对导航性能至关重要. 因此，本文设定阈值 $k=4$，此时时间相关性网络准确率达到 93.56%，满足视觉导航要求，且动作网络预测精度也在 90%以上，导航节点比例也处于较低水平.

表 1　分割阈值实验结果

k	动作网络/%	时间相关性网络/%	导航节点比例/%
1	96.65	89.25	33.71
2	95.14	92.37	21.46
3	93.06	93.11	15.58
4	**90.82**	**93.56**	**11.78**
5	86.73	91.02	10.29
7	81.15	87.98	13.54
10	72.86	82.43	12.01

5.2.3　环境交互量实验

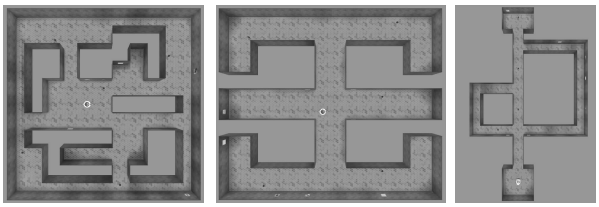
在整个学习过程中，训练样本由两部分组成：预训练和在线学习. 目标导向行为通过在线学习完成，因此不必关心样本数量问题. 而动作网络和时间相关性网络则需针对特定环境进行训练，为节省训练时间，有必要确定所需环境交互量. 训练数据量对网络性能影响如表 2 所示，表中数据为 5 个具有最优超参数的智能体平均所得. 由表 2 可知，随着交互量的增加，动作网络预测准确性持续上升，但增长比率逐渐下降；时间相关性网络的性能也随训练数据的增加而提高，但当网络处于过拟合状态时，预测准确率会有所下降. 经过综合考虑，将预训练部分与环境交互量设置为 2.5 M，此时两个网络预测精度都达到 90%以上，可满足算法要求.

表 2　环境交互量实验结果

交互量	动作网络/%	时间相关性网络/%
300 K	83.55	79.26
500 K	88.64	82.03
1 M	92.75	87.29
2.5 M	**93.83**	**92.41**
5 M	94.14	90.32

5.3　静态迷宫实验

本文所用测试环境如图 15 所示. 其中，Maze-1 为常规迷宫，其内部包含形状各异的障碍物和多条通路；Maze-2 设计灵感来源于 T 型迷宫[35]，它具有对称的空间结构，目标位于 4 个分支的尽头；Maze-3 最初用于验证认知地图理论，其环境由 3 条不同长度的通路组成. 在 Maze-1 和 Maze-3 中，目标和水果位置固定，而智能体起始位置在回合间随机变化. 但由于 Maze-2 空间结构的特殊性，其环境设置方式与前两者相反，即智能体起始位置固定，目标在回合间随机重置. 此类环境鼓励智能体学习一种探索-利用策略，即在探索迷宫过程中记忆目标位置，以便于在每次重置后更快速地找到目标.



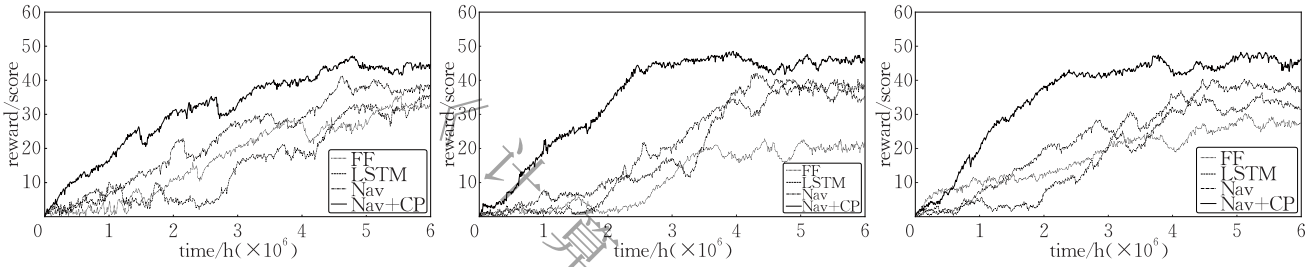
(a) Maze-1　(b) Maze-2　(c) Maze-3

图 15　静态迷宫测试环境

在不同测试环境中，目标导向行为学习曲线(数据为 5 个具有最佳性能的智能体平均所得)展现出一些特殊的结果. 首先，由于单一观测很难决定全局最优动作，智能体往往需要记住过去的状态才能维持导航功能. 但如图 16(a)所示，FF 模型在 Maze-1

中具有良好的动作规划能力，表明可能存在不涉及记忆的目标导向行为，即由编码器控制的纯反应式行为.然后，由图 16(b)可知，在明显需要记忆功能的 Maze-2 中，LSTM 模型所获奖励是 FF 模型的近两倍.这充分说明具有记忆功能的智能体可在探索环境过程中编码目标位置，并在随后的时间步内加以利用.单纯依赖反应式行为的智能体也可找到目标，但无法标记目标位置并多次使用.最后，图 16(c)清晰地显示出增加速度和动作作为额外输入以及使用碰撞预测作为辅助任务的影响.虽然 LSTM 模型在所获奖励上优于 FF 模型，但其训练速度仍然相对较慢.这主要是由传统强化学习算法导致，在增加碰撞预测后，Nav+CP 模型实现在所有环境中加速学习.此外，利用额外的输入和深度信息，智能体可更好地认知环境，同时获得更高效的导航策略.

在环境适应性方面，通过对智能体观测进行黑白化预处理，DQN 已被证明可在多种 Atari 游戏中学习人类级别的控制策略，展现出良好的环境适应性.而本文以 RGB 图像作为环境感知信息，为各模型学习控制策略提供了丰富的环境视觉信息，但这也对各模型的适应能力提出挑战.由图 16 可知，FF 及 LSTM 模型可通过 RGB 图像学习目标导向行为，特别是具有记忆功能的 LSTM 模型，在 3 个迷宫中均具有良好的学习能力.但无论是在学习效率上还是在所获奖励上，FF 和 LSTM 模型与本文模型还存在一定差距，表明 Nav+CP 模型更适用于色彩环境，且可更高效地将视觉信息转化为算法性能上的提升.



(a) Maze-1学习曲线 (b) Maze-2学习曲线 (c) Maze-3学习曲线

图 16 导航奖励时间(reward/time)图

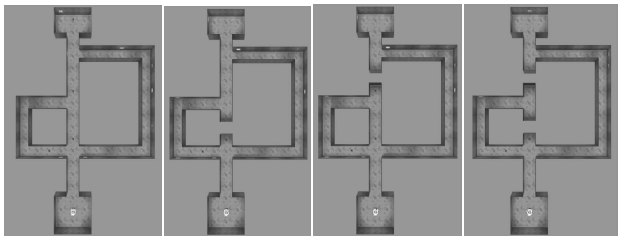
表 3 总结了各模型在不同环境内的性能参数及构建地图所需时间，表中数据为 5 个具有最优超参数的智能体平均所得.其中，学习曲线下面积(Area Under learning Curve，AUC)是一种比较学习效率的方法，学习曲线覆盖的面积越大，表明学习效率越高.构建地图所需时间是通过将学习过程中构建的拓扑地图与预采集的整个环境特征进行比较，当覆盖范围超过一定值后，则认为地图构建完成，并将此时的训练数据量记为地图完成时间.由表 3 可知，构建拓扑地图所用时间与智能体学习效率密切相关，学习曲线覆盖面积越大，构建地图所用时间越短.因此，需尽可能地提高智能体学习效率.FF 和 LSTM 模型以图像作为输入，Nav 模型在此基础上增加了动作和奖励，但这都不足以克服传统强化学习的影响.而与碰撞预测相结合的 Nav+CP 模型则利用了学习过程中的额外损失，通过这些包含环境结构信息的训练信号加速引导学习，同时减少构建拓扑地图所需数据量.

表 3 静态迷宫实验结果

环境	模型	AUC	奖励	地图完成时间(hour/1e6)
Maze-1	FF	42.5	33.24	3.52
	LSTM	40.6	35.28	4.36
	Nav	47.8	38.55	2.14
	Nav+CP	**63.4**	**45.63**	**1.17**
Maze-2	FF	27.9	21.97	3.84
	LSTM	45.3	37.92	3.16
	Nav	42.1	35.76	3.56
	Nav+CP	**74.2**	**46.65**	**1.22**
Maze-3	FF	38.6	28.41	3.14
	LSTM	39.5	34.64	3.06
	Nav	51.3	39.08	2.07
	Nav+CP	**80.4**	**47.53**	**1.04**

5.4 动态堵塞实验

在上一章节中，已获得全连通环境下的目标导向行为和拓扑地图，接下来的实验将测试该模型在包含动态堵塞环境中的性能.测试所用环境与 Maze-3 相同，但在通路中增加堵塞，堵塞位置如图 17 所示.



(a) 无堵塞 (b) A处堵塞 (c) B处堵塞 (d) AB处堵塞

图 17 动态堵塞测试环境

在实验过程中，首先固定智能体起始位置，然后使用在 Maze-3 中训练完成的智能体依次进行无堵塞、A 处堵塞、B 处堵塞、AB 处同时堵塞的实验.

未使用及使用拓扑地图导航实验结果如图 18 所示，数据为 5 个具有最佳性能的智能体平均所得，并以值函数/时间步(value-function/step)图表示，图中虚线为目标. 如图 18(a)、(b)所示，在没有堵塞的情况下，拓扑地图的存在与否并不影响模型导航性能，两种方法都使用中间路径完成导航任务，它们的值函数变化趋势相似，且均到达目标 4 次. 在接下来的实验中，两种方法的差异越加明显，从图 18(c)、(d)可知，Nav+CP 模型在 A 处停留，且不能绕过堵塞，导致智能体无法到达目标和值函数持续下降. Nav+CP+STM 模型同样遇到堵塞 A，但与前一种方法不同，它结合拓扑地图利用左侧路径到达目标，且由于左侧路径比中间路径长，所以智能体到达目标的频率比在无堵塞环境中低. 对于相对较远的堵塞，从图 18(e)、(f)可以看出，Nav+CP 模型仍然使用中间路径来引导目标导向行为，导致智能体停留在 B 处，而 Nav+CP+STM 模型则通过重新规划路径到达目标. 由于堵塞 B 的特殊位置，在使用拓扑地图规划路径时，智能体并没有试图使用左侧路径绕过堵塞 B，而是直接使用右侧路径. 同时，由于右侧通路是三者中最长的路径，智能体需要更多的动作才能到达目标，导致接触目标的次数进一步减少. 最后，从图 18(g)、(h)可知，即使在包含两个堵塞的环境中，Nav+CP+STM 模型仍然可以找到一条可行的路径到达目标，但探索过程较为复杂. 具体来说，智能体首先会停留在堵塞 A 处，然后使用左侧路径绕过堵塞 A 并遇到堵塞 B，最后通过右侧路径绕过堵塞 B 到达目标，这也是智能体在 5000 个时间步内只到达目标一次的原因. 对于 Nav+CP 模型，与第二次实验类似，智能体会始终停留在堵塞 A 处.

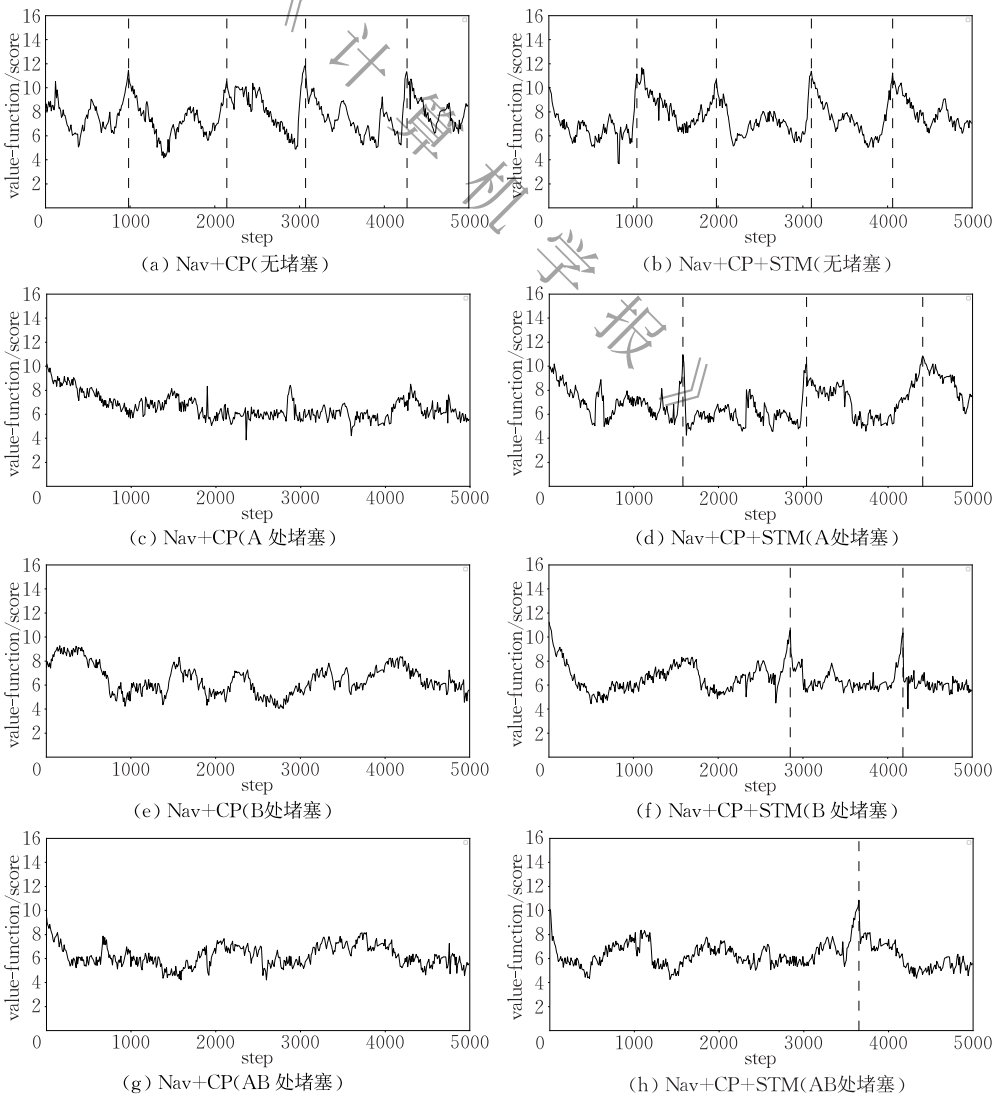


图 18　值函数时间步(value-function/step)图

为更好展现拓扑地图在模型中的作用,在实验过程中收集了更为详细的数据并总结为表 4,表中数据为 5 个具有最优超参数的智能体平均所得.其中,包括智能体在 5000 个时间步内到达目标的次数和所获奖励,延迟为智能体首次找到目标与随后找到目标所用时间步之比.从表 4 可以看出,随着堵塞位置从 A 移动到 B 及堵塞数量的增加,Nav+CP+STM 模型需要更多的时间步才能到达目标,导致智能体到达目标次数和所获奖励的减少.但无论堵塞的位置和数量如何变化,集成拓扑地图的智能体始终可以找到目标.而对于 Nav+CP 模型,由于不能动态规划路径,一旦环境中出现堵塞,智能体将长时间停滞在堵塞位置.这进一步证明拓扑地图可作为路径规划模块集成到模型中,并用于引导动态环境下的绕路行为.

表 4 动态堵塞实验结果

环境	模型	目标次数	奖励	延迟
无堵塞	Nav+CP	4.6	48.2	0.99
	Nav+CP+STM	**4.5**	**47.5**	**1.02**
	OP	5.2	54.3	1.01
A 处堵塞	Nav+CP	0	3.3	
	Nav+CP+STM	**3.1**	**32.4**	**1.21**
	OP	3.7	39.2	0.99
B 处堵塞	Nav+CP	0	5.6	∞
	Nav+CP+STM	**2.1**	**23.4**	**1.16**
	OP	2.4	27.1	1.02
AB 处堵塞	Nav+CP	0	3.4	∞
	Nav+CP+STM	**1.7**	**21.6**	**1.38**
	OP	2.5	27.8	1.01

6 结 论

针对动态环境中的导航问题,本文提出一种可同步学习目标导向行为和构建空间拓扑地图的视觉导航方法.为在具有复杂结构且丰富视觉的状态空间中学习目标驱动的导航策略,以深度强化学习为基本框架,并在模型中结合碰撞预测提供密集训练信号,以实现加速学习和提升导航性能.对于编码环境,利用图像之间的时间相关性祛除冗余观测和寻找导航节点,并通过集成情景记忆描述环境结构.实验结果表明,本文方法可从原始传感器输入中学习目标导向行为,同时构建空间拓扑地图,即使在包含动态堵塞的环境中也可实现高效导航.在接下来的研究中,我们将进一步优化本文模型,并力求在真实环境中验证其性能.除此之外,也将基于本文模型对非常大或终身学习场景中的导航方法进行深入探讨.

参 考 文 献

[1] Tommasi L, Chiandetti C, Pecchia T, et al. From natural geometry to spatial cognition. Neuroscience & Biobehavioral Reviews, 2012, 36(2): 799-824

[2] Moser E I, Kropff E, Moser M B. Place cells, grid cells, and the brain's spatial representation system. Annual Review of Neuroscience, 2008, 31: 69-89

[3] Tolman E C. Cognitive maps in rats and men. Psychological Review, 1948, 55(4): 189-208

[4] Madl T, Franklin S, Chen K, et al. Bayesian integration of information in hippocampal place cells. PLoS One, 2014, 9(3): e89762

[5] Yonelinas A P, Otten L J, Shaw K N, et al. Separating the brain regions involved in recollection and familiarity in recognition memory. The Journal of Neuroscience, 2005, 25: 3002-3008

[6] Cesar C, Luca C, Henry C, et al. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. IEEE Transactions on Robotics, 2016, 32(6): 1309-1332

[7] Sun Y X, Liu M, Meng M Q H. Improving RGB-D SLAM in dynamic environments: A motion removal approach. Robotics and Autonomous Systems, 2017, 89: 110-122

[8] Song Hai-Tao, He Wen-Hao, Yuan Kui. A stereo vision system based on SIFT feature for robot environment perception. Control and Decision, 2019, 34(7): 1545-1552(in Chinese)
(宋海涛, 何文浩, 原魁. 一种基于 SIFT 特征的机器人环境感知双目立体视觉系统. 控制与决策, 2019, 34(7): 1545-1552)

[9] Mur-Artal R, Tardos J D. ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras. IEEE Transactions on Robotics, 2017, 33(5): 1255-1262

[10] Liu Quan, Zhai Jian-Wei, Zhang Zong-Zhang, et al. A survey on deep reinforcement learning. Chinese Journal of Computers, 2018, 41(1): 1-27(in Chinese)
(刘全, 翟建伟, 章宗长等. 深度强化学习综述. 计算机学报, 2018, 41(1): 1-27)

[11] Yann L, Yoshua B, Geoffrey H. Deep Learning. Nature, 2015, 521(7553): 436-444

[12] Oh J, Chockalingam V, Singh S P, et al. Control of memory, active perception, and action in Minecraft. https://arxiv.org/pdf/1605.09128.pdf. 2016, 05, 01

[13] Zhu Y, Mottaghi R, Kolve E, et al. Target-driven visual navigation in indoor scenes using deep reinforcement learning. https://arxiv.org/pdf/1609.05143.pdf. 2016, 09, 16

[14] Mnih V, Badia A P, Mirza M, et al. Asynchronous methods for deep reinforcement learning. https://arxiv.org/abs/1602.01783.pdf. 2016, 06, 16

[15] Jaderberg M, Mnih V, Czarnecki W M, et al. Reinforcement learning with unsupervised auxiliary tasks. https://arxiv.org/abs/1611.05397.pdf. 2016, 11, 16

[16] Volodymyr M, Koray K, Silver S D, et al. Human-level control through deep reinforcement learning. Nature, 2015, 518: 529-533

[17] Mirowski P, Pascanu R, Viola F, et al. Learning to navigate in complex environment. https://arxiv.org/pdf/1611.03673.pdf. 2017, 01, 13

[18] Yu Nai-Gong, Yuan Yun-He, Li Ti, et al. A cognitive map building algorithm by means of cognitive mechanism of hippocampus. Acta Automatica Sinica, 2018, 44(1): 52-73 (in Chinese)
(于乃功，苑云鹤，李倜等. 一种基于海马认知机理的仿生机器人认知地图构建方法. 自动化学报，2018，44(1)：52-73)

[19] Parisotto E, Salakhutdinov R. Neural map: Structured memory for deep reinforcement learning. https://arxiv.org/abs/1702.08360.pdf. 2017, 02, 27

[20] Gupta S, Davidson J, Levine S, et al. Cognitive mapping and planning for visual navigation. https://arxiv.org/abs/1702.03920.pdf. 2019, 02, 07

[21] Savinov N, Dosovitskiy A, Koltun V. Semi-parametric topological memory for navigation. https://arxiv.org/abs/1803.00653.pdf. 2018, 03, 01

[22] Hausknecht M, Stone P. Deep recurrent Q-learning for partially observable MDPs. https://arxiv.org/abs/1507.06527.pdf. 2017, 01, 11

[23] Martin R. Neural fitted Q iteration-first experiences with a data efficient neural reinforcement learning method// Proceedings of the European Conference on Machine Learning (ECML 2005). Berlin, Heidelberg, Germany, 2005: 317-328

[24] Lange S, Riedmiller M, Voigtländer A. Autonomous reinforcement learning on raw visual input data in a real world application//Proceedings of the 2012 International Joint Conference on Neural Networks (IJCNN). Brisbane, Australia, 2012: 1-8

[25] Thomas G T, Egidio F, Federico R, et al. Model-based reinforcement learning for closed-loop dynamic control of soft robotic manipulators. IEEE Transactions on Robotics, 2019, 35(1): 124-134

[26] Liu Jian-Wei, Gao Feng, Luo Xiong-Lin. Survey of deep reinforcement learning based on value function and policy gradient. Chinese Journal of Computers, 2019, 42(6): 1406-1438(in Chinese)
(刘建伟，高峰，罗雄霖. 基于值函数和策略梯度的深度强化学习综述. 计算机学报. 2019，42(6)：1406-1438)

[27] Kaelbling L P, Littman M L, Cassandra A R. Planning and acting in partially observable stochastic domains. Artificial Intelligence, 1998, 101(1): 99-134

[28] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. https://arxiv.org/abs/1512.03385.pdf. 2015, 12, 10

[29] McNaughton B L, Battaglia F P, Jensen O, et al. Path integration and the neural basis of the 'cognitive map'. Nature Reviews Neuroscience, 2006, 7: 663-678

[30] Savinov N, Raichuk A, Marinier R, et al. Episodic curiosity through reachability. https://arxiv.org/abs/1810.02274.pdf. 2019, 08, 06

[31] Thomash H C, Charles E L, Ronald L R, et al. Introduction to Algorithms. Cambridge, USA: MIT Press, 2005

[32] Beattie C, Leibo J Z, Teplyashin D, et al. DeepMind Lab. https://arxiv.org/abs/1612.03801.pdf. 2016, 12, 12

[33] Gers F A, Schmidhuber J, Cummins F. Learning to forget: Continual prediction with LSTM. Neural Computation, 2000, 12(10): 2451-2471

[34] Diederik P K, Jimmy B. Adam: A method for stochastic optimization. https://arxiv.org/abs/1412.6980.pdf. 2017, 01, 30

[35] David O S, James T B, Gail E H. Hippocampus, space, and memory. Behavioral and Brain Science, 1979, 2(3): 313-322

RUAN Xiao-Gang, Ph.D., professor. His research interests include automatic control, artificial intelligence and intelligent robot.

LI Peng, Ph. D. candidate. His research interests include deep reinforcement learning and robot navigation problem.

ZHU Xiao-Qing, Ph. D., lecturer. His research interests include intelligent robot and machine learning.

LIU Peng-Fei, M. S. candidate. His research interests include artificial intelligence and robot navigation problems.

Background

Learning to navigate in complex environment with dynamic elements is a challenge in developing AI agent and most of today's robot algorithm struggle with such condition. Inspired by the researches about cognitive behavior in animals, there are many navigation methods design the agent to encode environmental structure during exploration process, and the typically used approach is SLAM. This kind of approach builds metric map of unknown environment by using sensory information from laser, odometer, sonar or vision. Through application motion information of robot and features of observation, the agent can get accurate estimation of environment and use it to realize autonomous navigation. Particularly

relevant to our work is visual SLAM(VSLAM) that use image as the main perception of state space, and aim to reconstruct 3D map by camera pose and multi-view geometry theory. In order to improve the speed of data processing, some VSLAM algorithms extract feature points of observation firstly, and then perform inter-frame estimation and closed-loop detection through matching these points. The SLAM-based approaches can provide high-quality environment map, but they are explicit focus on position inference and mapping, and need externally provided camera pose or ego-motion information, and do not naturally accommodate dynamic environment.

More recently, many researchers have noted the outstanding ability of deep learning (DL) in overcome the problems stem from dimensional disaster, and try to take advantage of it to help navigation in high-dimensional state space. So we considered using the deep reinforcement learning (DRL), which consist of DL and reinforcement learning (RL) and apply to navigation vary both in learning method and memory representation, as basic learning framework to get goal-driven behaviour and memory spatial structure.

In this paper, we proposed a novel architecture of navigation which can build space topological map during learning navigational policy. To directly perceive environmental information from visual inputs, using an agent with DRL framework to learn control policy, and the map is formed based on temporal correlation. Crucially, the temporal correlation is a predictive value which shows whether the pairs of observation temporally close or not. This allows us to find navigational nodes through comparing the trajectory recording with the map, and incrementally describe the environment by integrating every observation sequence.

This work is support by the National Natural Science Foundation of China (No. 61773027), the Natural Science Foundation of Beijing (No. 4202005), and the Project of S&T Plan of Beijing Municipal Commission of Education (No. KM201810005028).