DOI: 10.13382/j.jemi.B2205242

融合 LSTM 和 PPO 算法的移动机器人视觉导航*

张 仪[1,2] 冯 伟[1,2,3] 王卫军[1,2] 杨之乐[1,2,3] 张艳辉[1,2,3] 朱子翰[1,2] 谭 勇[4]

(1. 中国科学院深圳先进技术研究院 深圳 518055;2. 中国科学院大学 北京 100049;3. 广东省机器人与智能系统重点实验室 深圳 518055;4. 上海诺倬力机电科技有限公司 上海 200000)

摘 要:为提高移动机器人在无地图情况下的视觉导航能力,提升导航成功率,提出了一种融合长短期记忆神经网络(long short term memory, LSTM)和近端策略优化算法(proximal policy optimization, PPO)算法的移动机器人视觉导航模型。首先,该模型融合 LSTM 和 PPO 算法作为视觉导航的网络模型;其次,通过移动机器人动作,与目标距离,运动时间等因素设计奖励函数,用以训练目标;最后,以移动机器人第一视角获得的 RGB-D 图像及目标点的极性坐标为输入,以移动机器人的连续动作值为输出,实现无地图的端到端视觉导航任务,并根据推理到达未接受过训练的新目标。对比前序算法,该模型在模拟环境中收敛速度更快,旧目标的导航成功率平均提高 17.7%,新目标的导航成功率提高 23.3%,具有较好的导航性能。

关键词: 近端策略优化算法;长短期记忆神经网络;视觉导航

中图分类号: TN8;TP242 **文献标识码**:A **国家标准学科分类代码**: 510.40;510.80

Visual navigation of mobile robots based on LSTM and PPO algorithms

Zhang Yi[1,2] Feng Wei[1,2,3] Wang Weijun[1,2] Yang Zhile[1,2,3] Zhang Yanhui[1,2,3] Zhu Zihan[1,2] Tan Yong[4]

(1. Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China; 2. University of Chinese Academy of Sciences, Beijing 100049, China; 3. Guangdong Provincial Key Lab of Robotics and Intelligent System, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China; 4. Shanghai Nozoli Machine Tools Technology Co., Ltd., Shanghai 200000, China)

Abstract: In order to improve the visual navigation ability of mobile robots without maps and improve the success rate of visual navigation, a visual navigation model of mobile robots is proposed that integrates long short term memory (LSTM) and proximal policy optimization (PPO) algorithms. Firstly, the model integrates LSTM and PPO as a network model for visual navigation. Secondly, a new reward function is designed to train the target through factors such as the action of mobile robots, the distance between the robots and the target, and the running time of robots. Finally, the RGB-D image obtained from the first perspective of mobile robots and the polar coordinates of the target in mobile robots coordinate system are used as the model input, and the continuous motion of mobile robots is used as the model output to realize the task of end-to-end visual navigation without maps, and the new target that has not been trained is reached according to the model inference. Compared with the pre-order algorithms, the model has an average increase of 17.7% in the navigation success rate of the old target and 23.3% of the new target in simulated environments, which has better navigation performance.

Keywords: proximal policy optimization algorithms; long short term memory; visual navigation

收稿日期: 2022-03-08 Received Date: 2022-03-08

*基金项目:国家自然科学基金联合基金项目(U20A20283)、工信部冰雪器材加工成套装备项目(TC190H47P)、国家自然科学基金联合基金项目(U1813222)、深圳市国际合作研究项目(GJHZ20200731095009029)、广东特支计划科技创新青年拔尖人才项目(2019TQ05Z654)资助

0　引　言

随着移动机器人技术的迅速发展,移动机器人在各行各业中的使用也越发普遍。其中,移动机器人的自主导航能力十分关键,它根据目标位置的不同生成任意的无碰撞路径[1]。传统的移动机器人导航多为基于地图的方法,包括同时定位和建图(simultaneous localization and mapping,SLAM)[2]及路径规划[3],通常先通过 SLAM 方法构建有关环境的地图,再采用路径规划方法生成一条从起点到终点的无碰撞路径。其中 SLAM 又分为两种类别:激光 SLAM 和视觉 SLAM[4]。激光 SLAM 是基于激光传感器建立环境地图,目前应用较广,成果较多,但激光传感器造价高昂,在雨雪等天气下表现较差;视觉 SLAM 是基于视觉传感器构建地图[5],视觉传感器造价低,也可保留有关环境的语义信息,但在光线条件差的情况下,会出现纹理特征缺失等情况,多用于室内光线较好的场景中[6]。传统的导航方法对地图依赖程度较高,若 SLAM 对环境的认知有偏差,在路径规划阶段极有可能出现导航失败的情况。因此,在无地图情况下实现移动机器人的自主导航显得非常重要。而传统的路径导航算法无法有效的通过对环境的观察实现无地图的导航。而随着 2015 年深度强化学习的标志性成果——深度 Q 学习(deep Q network,DQN)[7]在 Nature 发表,通过原始高维环境输入例如视觉,使无地图导航成为可能。随后,有关深度强化学习的算法被不断提出。主要分为基于值函数与策略函数两种。基于值函数的方法以 DQN 为代表,后在此基础上提出了 Double DQN(DDQN)、Dueling DQN 等算法,这些方法在 Atari、导航、游戏等一系列任务中达到了人类专家水平。基于策略函数的深度强化学习算法主要为深度确定性策略梯度算法(deep deterministic policy gradient,DDPG)[8],异步优势评论者算法(asynchronous advantage actor critic,A3C)[9],置信区域策略优化算法(trust region policy optimization,TRPO)[10],近端策略优化算法(proximal policy optimization,PPO)PPO[11]等方法。另外,通过对算法进行并行化处理,为深度强化学习算法在实际环境中的应用提供了可能。

由于深度强化学习的发展,基于深度强化学习算法的移动机器人在无地图情况下的视觉导航研究逐渐增多,也是目前研究的热点问题[12]。Zhu 等[13]率先将深度强化学习应用到移动机器人的视觉导航中,只需将目标图像及移动机器人当前观测到的第一视角图像输入该视觉导航框架中,即可输出移动机器人的动作直到到达目标位置,实现端到端的无地图视觉导航。Nwaonumah 等[14]基于机器人操作系统(robot operating system,ROS),将移动机器人观测到的 RGB-D 图像作为输入,在 3 种仿真环境中对比了 DQN 和 A3C 两种深度强化学习算法在实现端到端的目标驱动视觉导航时的性能差异,发现 A3C 模型在奖励估计和学习趋势方面表现优于 DQN,另外,A3C 中线程越多导航性能越好。Ma 等[1]提出一种基于堆叠长短期记忆网络(long short term memory network,LSTM)的无地图视觉导航方法,它不仅可以记住训练过的目标,更重要的是,还可经过推理到达未训练过的目标,并在模拟和真实环境中进行了实验。Devo 等[15]提出了一种基于对象本地化和导航网络的网络框架模型,对象本地化网络将目标图像和当前观测图像作为输入,并输出对应的六维特征向量,再将六维特征向量和当前观测图像输入导航网络,并输出动作,该模型可在无需调整的情况下映射到真实环境。Yokoyama 等[16]提出了一个使用预先训练的深度预测模型估计单眼相机图像的深度数据的系统,再将深度数据转换为距离数据,并通过 DDQN 算法输出移动机器人的动作。

本文在文献[1]的深度强化学习视觉导航模型上改进,创新性的提出了一种融合 LSTM 和 PPO 算法的移动机器人视觉导航模型。该模型以移动机器人第一视角获得的 RGB-D 图像及在移动机器人坐标系下目标点的极性坐标为输入,在现有的移动机器人视觉导航模型基础上,融合 LSTM 和 PPO 算法改进网络结构和奖励函数来训练目标,以移动机器人的连续动作值为输出,实现无地图的端到端视觉导航任务,使移动机器人无碰撞的到达目标位置,并通过推理到达未经过训练的目标,提升视觉导航模型的导航成功率。

1　相关原理

1.1　LSTM 网络

LSTM 是在循环神经网络(recurrent neural networks,RNN)基础上改进的网络结构。RNN 是一种用于处理序列数据的神经网络,通过对现有的信息进行记忆去推测未知的信息,结构如图 1 所示。

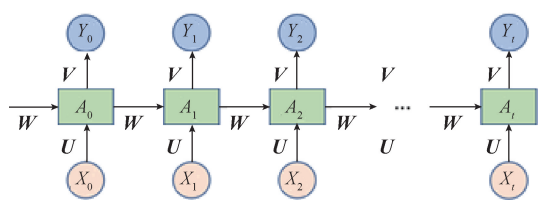


图 1　RNN 网络结构

Fig. 1　RNN structure

RNN 主要分为输入层、隐藏层及输出层。其中 $\boldsymbol{X}_t$ 是输入 RNN 网络中的特征向量,$\boldsymbol{U}$ 是输入层到隐藏层的参数矩阵,$\boldsymbol{A}_t$ 是隐藏层中的向量,$\boldsymbol{W}$ 是各时间点的权重矩

阵，$\boldsymbol{V}$ 是隐藏层到输出层的参数矩阵，$\boldsymbol{Y}_t$ 是输入 RNN 网络中的特征向量[12]。其中输出 $\boldsymbol{Y}_t$ 和隐藏层向量 $\boldsymbol{A}$ 的更新公式为：

$$\boldsymbol{Y}_t = g(\boldsymbol{V} \cdot \boldsymbol{A}_t) \tag{1}$$

$$\boldsymbol{A}_t = f(\boldsymbol{U} \cdot \boldsymbol{X}_t + \boldsymbol{W} \cdot \boldsymbol{A}_{t-1}) \tag{2}$$

RNN 中隐藏层的输入由该时刻的输入层及上一时刻隐藏层的输出同时决定，因此 RNN 可以记忆并处理序列数据。但 RNN 对于较长时间的信息会出现遗忘的情况，因此无法利用该网络继续预测或判断新的信息，这种问题称为“长依赖”。

为了解决 RNN 网络出现的“长依赖”问题，提出了 LSTM[17]，该网络通过增加“遗忘门”改进 RNN 结构。LSTM 在隐含层计算中增加了细胞状态和门机制，RNN 中的每个隐藏单元在 LSTM 变成一个有记忆功能的细胞，可以记忆长时间信息，LSTM 中由遗忘门、输入门和输出门 3 种控制门，结构如图 2 所示。

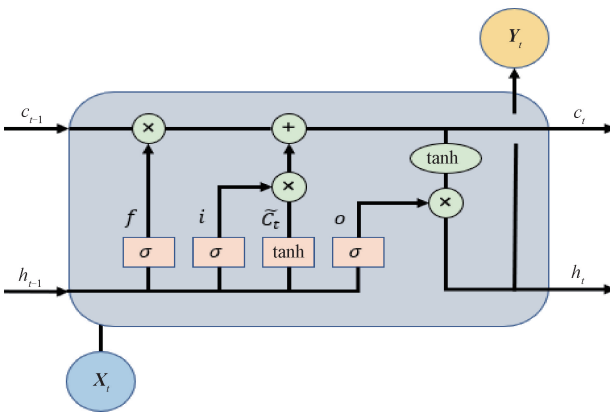


图 2 LSTM 网络结构图

Fig. 2 LSTM structure

遗忘门负责调控上一时刻细胞状态 C_{t-1} 传递到当前时刻细胞状态 C_t 的比例，输入门负责调控当前时刻网络的输入 $\boldsymbol{X}_t$ 传递到当前时刻细胞状态 C_t 的比例，输出门调控当前时刻细胞状态 C_t 传递到当前输出值 h_t 的比例，各状态更新公式如下[18]：

门控单元：

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{3}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{4}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{5}$$

存储单元：

$$\widetilde{C_t} = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{6}$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \widetilde{C_t} \tag{7}$$

输出状态：

$$Y_t = o_t \cdot \tanh(C_t) \tag{8}$$

其中，W 表示输入量的权重，$\boldsymbol{b}$ 为偏置向量。

LSTM 中的细胞状态和控制门的使用可以对时间序列中较早的信息有良好的记忆，也能掌握在序列信息中存在的长距离相关性，另外，细胞状态的增加也可以防止出现梯度消失的情况。

1.2 深度强化学习算法

强化学习主要解决的问题是：通过设计奖励函数，使智能体在与环境交互的过程当中不断学习策略，最终实现智能体奖励最大化的过程[19]。深度强化学习通过将深度学习方法和传统强化学习结合，利用深度神经网络对强化学习的状态、动作、价值等函数进行拟合，提高强化学习在高维场景下的应用能力[20]。

1）策略梯度

策略函数 π_θ 表示智能体在当前状态下采取各种动作的概率，是一个有关动作选择的概率密度函数[21]。将 $\tau = (s_0, a_0, \cdots, s_t, a_t,)$ 表示根据完成策略 π_θ 得到的一条轨迹，$r(\tau)$ 为这条轨迹在与环境交互时得到的奖励，将 $J(\theta)$ 表示为该轨迹的期望奖励，通过梯度上升来寻找最优参数 θ，将期望奖励 $J(\theta)$ 最大化作为优化目标，即 $\max J(\theta)$。

当前策略网络参数为 θ_t，则下一刻参数更新为：

$$\theta_{t+1} \leftarrow \theta_t + \beta \cdot \nabla J(\theta_t) \tag{9}$$

其中，β 为学习率，$\nabla J(\theta_t)$ 为策略梯度，$J(\theta)$ 可表示为：

$$J(\theta) = E_{\tau \sim \pi_\theta(\tau)}(r(\tau)) \tag{10}$$

则策略梯度为：

$$\nabla J(\theta_t) = E_{\pi_\theta}(\nabla_\theta \log \pi_\theta(s,a) r(\tau)) \tag{11}$$

利用最大化轨迹奖励来更新策略参数，最终结果是准确无偏的，但在实际训练过程中，由于与环境交互数据量不够，会出现理论数据与真实数据之间有所偏差的情况。

2）AC（Actor-Critic）算法

Actor-Critic[22] 包括两个网络：一个网络是策略网络，用 $\pi_\theta(s_t, a_t)$ 表示；一个是价值网络，用 $Q_\omega(s_t, a_t)$ 表示。其中，θ，ω 为网络参数，s_t 为智能体在 t 时刻下的状态，a_t 为智能体在 t 时刻下的动作。两个网络有一个共同点，可通过输入智能体状态 s_t：策略网络输出策略 π_θ，负责选择动作 a_t，称为 Actor；价值网络计算动作 a_t 得到的奖励 Q_ω，称为 Critic。

AC 算法采用策略梯度进行参数更新，公式为：

$$\theta_{t+1} \leftarrow \theta_t + \beta \cdot \nabla \log \pi_\theta(s_t, a_t) Q_\omega(s_t, a_t) \tag{12}$$

其中，β 为学习率，$\nabla \log \pi_\theta(s_t, a_t) Q_\omega(s_t, a_t)$ 为策略梯度。

由于 AC 算法属于在策略方法，在真实环境中学习效率低，因此网络参数更新较慢。

2　方法

2.1　问题定义

本文的主要任务是实现移动机器人在地图未知的情况下,通过输入移动机器人当前时刻以第一视角观测到的 RGB-D 图像及在移动机器人坐标系下目标点的极性坐标,融合 LSTM 和 PPO 算法改进网络结构和奖励函数来训练目标,以移动机器人的连续动作值为输出,实现无地图的端到端视觉导航任务,使移动机器人无碰撞的到达目标位置,并通过推理到达未经过训练的目标。因此,该问题可被定义为[1]:

$$V_t = f(I_t, T_t, V_{t-1}) \tag{13}$$

其中,I_t 为移动机器人观测到的 RGB-D 图像特征提取后的特征值,T_t 为当前移动机器人与目标位置的相对位置,V_{t-1} 为移动机器人上一时刻的动作,其中包括移动机器人的线速度与角,速度 f 为映射函数。

2.2　堆叠 LSTM

单个 LSTM 层始终不优于堆叠的 LSTM 结构[1],本文用两层堆叠 LSTM 网络记录移动机器人在环境交互过程中获得的经验,其中,每个 LSTM 层有 256 个单元。在第 1 层 LSTM 中,将移动机器人第一视角观测到的图像经卷积后得到的图像特征以及目标相对于移动机器人当前位置的极性作为输入,再将该层的输出以及上一时刻移动机器人的奖励和速度输入第 2 层 LSTM 中,堆叠 LSTM 网络结构如图 3 所示。

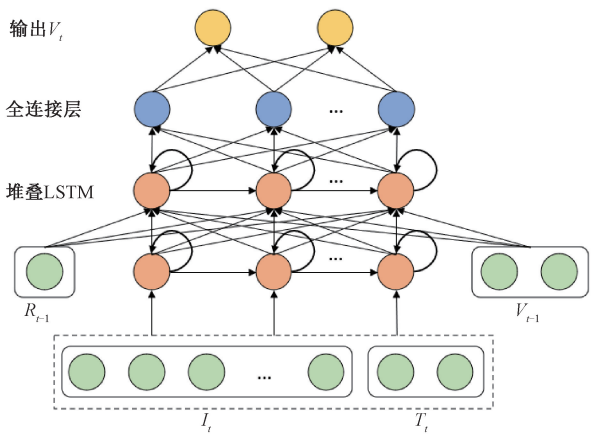


图 3　堆叠 LSTM 网络结构图

Fig. 3　Stacked LSTM structure

其中,I_t 为移动机器人观测到的 RGB-D 图像特征信息,T_t 为有关目标点的信息,V_{t-1} 为上一时刻移动机器人的动作,R_{t-1} 为上一时刻移动机器人执行的动作在环境中获得的奖励。

2.3　PPO 算法[23]

PPO 算法由 TRPO 算法简化而来,是一种离策略方法,可以用较少的经验使模型快速收敛。TRPO 算法的核心思想是在一定的置信域区间上对复杂函数做近似,再求解近似函数的最大化[10]。TRPO 的核心思想是让每一次的策略更新在一个置信区域内,保证 policy 的单调提升。

TRPO 算法定义如下:

$$\max E_{s\sim\rho_{\theta_{old}},a\sim\pi_{\theta_{old}}}\left[\frac{\pi_\theta(a\mid s)}{\pi_{\theta_{old}}(a\mid s)}A_{\theta_{old}}(s,a)\right]$$

$$\text{s.t. } \bar{D}_{KL}^{\rho_{\theta_{old}}}(\pi_{\theta_{old}},\pi_\theta)\leqslant\delta \tag{14}$$

其中,$\pi_\theta(a\mid s)$ 为训练过程中更新的新策略,$\pi_{\theta_{old}}(a\mid s)$ 为旧策略,$A_{\theta_{old}}(s,a)$ 为价值网络输出的优势函数,新策略中的参数 θ 经一段时间更新后便取代旧策略中的 θ_{old}。约束条件的目的是为了控制新策略的概率分布与旧策略的概率分布差别不能太大,以保证算法更新的稳定性。

PPO 算法将 TRPO 算法中的约束条件转化为新旧策略的比值约束,目标函数为:

$$L(s,a,\theta_{old},\theta)=\min\begin{cases}\dfrac{\pi_\theta(a\mid s)}{\pi_{\theta_{old}}(a\mid s)}A(s,a)\\ clip\left(\dfrac{\pi_\theta(a\mid s)}{\pi_{\theta_{old}}(a\mid s)},1-\varepsilon,1+\varepsilon\right)A(s,a)\end{cases} \tag{15}$$

其中,$clip$ 函数是将新旧策略的比值约束控制在 $[1-\varepsilon,1+\varepsilon]$,确保参数更新的稳定性。

其中参数更新公式为:

$$\theta\leftarrow\arg\max_\theta E_{s,a\sim\pi_{\theta_{old}}}[L(s,a,\theta_{old},\theta)] \tag{16}$$

2.4　移动机器人视觉导航模型

移动机器人视觉导航模型以移动机器人第一视角观测到的 RGB-D 图像及在移动机器人坐标系下的目标点有关距离和角度的极性坐标为输入,获得移动机器人的连续动作值,实现端到端的视觉导航。通过对目标进行训练,更新视觉导航模型中的各项参数,再对未进行过训练的目标进行推理,通过新目标到达成功率来决定视觉导航模型是否继续更新,新目标到达成功率越高,视觉导航模型对环境信息掌握程度越高。

视觉导航模型如图 4 所示,RGB 图像尺寸为 48×64×3,深度图像尺寸为 24×32×1,分别经过不同的 3 层卷积神经网络得到对应的特征值,输入到含有 32 个神经元的全连接层中,将得到的输出与目标信息输入到第 1 层 LSTM 网络中,再与速度 V_{t-1} 及奖励 R_{t-1} 输入到第 2 层 LSTM 网络中,再输入包含 256 个神经元的全连接层中,根据 PPO 算法确定移动机器人当前时刻的动作 V_t。其

中，方块代表目标点信息，V_{t-1} 为上一时刻移动机器人的动作，包括移动机器人的线速度 V_{t-1}^{l} 与角速度 V_{t-1}^{w}，R_{t-1} 为上一时刻移动机器人执行的动作在环境中获得的奖励。

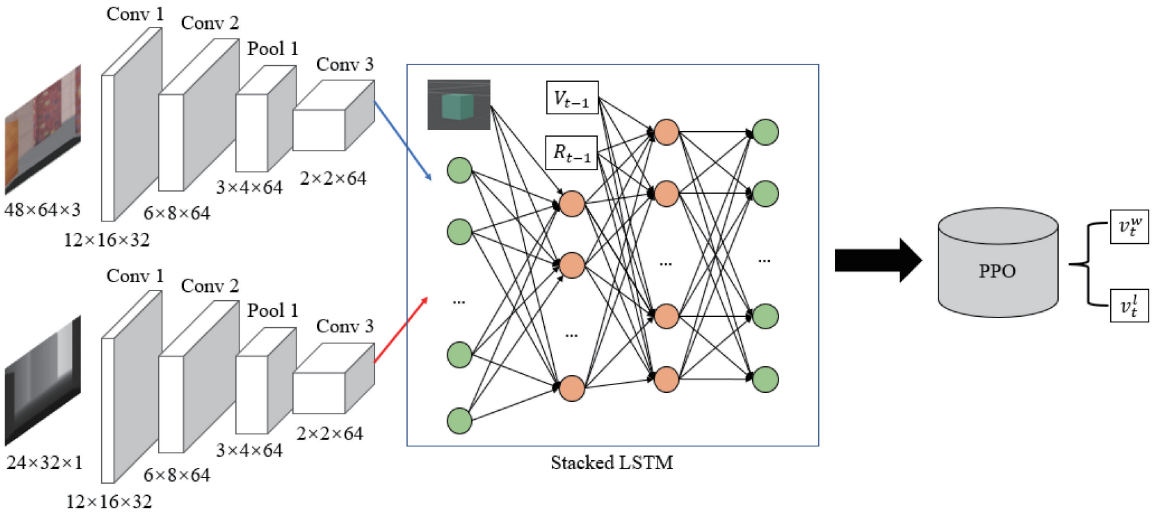


图4　融合 LSTM 和 PPO 算法的移动机器人视觉导航模型

Fig. 4　Visual navigation model for mobile robots integrating LSTM and PPO algorithms

2.5　动作与奖励

在三维环境中，移动机器人的线速度范围在 0～0.3 m/s 之间连续选取，角速度范围在 0～1 rad/s 之间连续选取，移动机器人通过在环境中收集经验训练视觉导航模型，得到从起点到训练目标的无碰撞路径，并根据推理到达未训练过的推理目标，实现端到端的视觉导航。

深度强化学习中的奖励函数设计十分关键，奖励函数的好坏决定了智能体能否有效的学习环境，能否快速完成任务。在视觉导航模型中，移动机器人能否安全快速平滑的到达目标，取决于移动机器人是否发生碰撞，是否速度够快，是否转弯较少。

因此，若想使移动机器人安全无碰撞的到达目标位置，需要在移动机器人发生碰撞时给与负奖励 r_c；在与目标位置的距离小于 c_d 时，认定移动机器人已经到达目标位置，给予正奖励 r_a；在其他情况下，为了鼓励移动机器人探索环境，加入有关移动机器人线速度的正奖励；为了使移动机器人到达目标的轨迹更平滑，加入有关移动机器人角速度的负奖励；为了使移动机器人朝着目标所在方向移动，加入有关移动机器人与目标距离的正向奖励；为了使移动机器人快速接近目标，给与有关时间的负奖励。因此，移动机器人视觉导航模型的奖励函数被定义为：

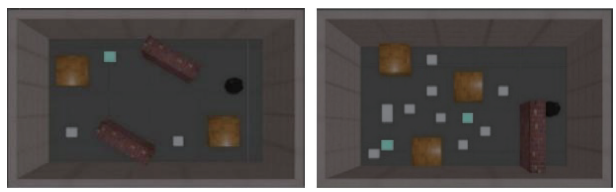
$$f(x)=\begin{cases} r_c, \text{发生碰撞} \\ r_a, d_t < c_d \\ c_r \cdot (d_{t-1}-d_t) + c_l \cdot v_t^l + c_a \cdot (v_t^a)^2 + c_t, \text{其他} \end{cases} \tag{17}$$

其中，r_c 为碰撞奖励，d_t 是 t 时刻下移动机器人与目标距离，若该距离小于阈值 c_d，移动机器人则获得奖励 r_a，c_r 为距离参数，c_l 为线速度参数，c_a 为角速度参数，c_t 为时间参数，v_t^l 为移动机器人在 t 时刻下的线速度，v_t^a 为移动机器人在 t 时刻下的角速度。

3　实验

3.1　实验环境

实验环境为 Ubuntu16.04 下的 ROS 操作系统中的 gazebo 仿真环境，仿真环境与文献[1]一致，如图5所示。移动机器人所在位置是起始位置，白色方块为训练目标，通过对这些目标的训练来更新视觉导航模型，深色方块是用于评估视觉导航模型的推理性能的未经过训练的新目标。移动机器人上安装了深度相机，以接收实时的 RGB-D 图像。



（a）环境1地图　（b）环境2地图

（a）Map 1　（b）Map 2

图5　环境地图

Fig. 5　Environment map

3.2　环境 1 仿真结果

环境 1 新旧目标的成功率对比了最后 100 个回合的

导航任务的成功率,如图 6 所示。当移动机器人可以以 60%的成功率到达新目标时,到达旧目标得成功率为 81%。训练过程中,奖励持续上涨,如图 7 所示。

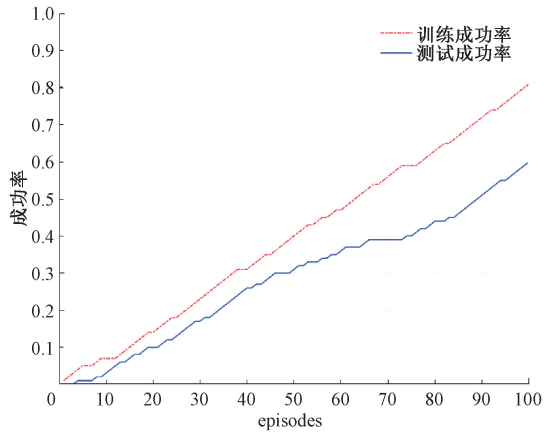


图 6　环境 1 改进模型成功率

Fig. 6　Improve model success rate in Map 1

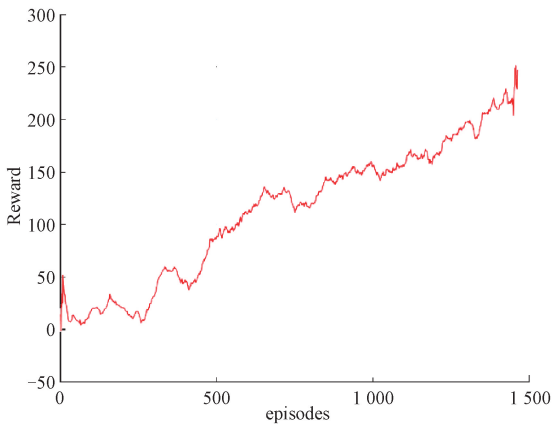


图 7　环境 1 改进模型奖励曲线图

Fig. 7　Improve model reward in Map 1

3.3　环境 2 仿真结果

环境 2 新旧目标的成功率对比了最后 100 个回合的导航任务的成功率,如图 8 所示。环境 2 比环境 1 更复杂,因此设置了更多的训练目标,以确保导航模型对环境有充分的了解。当移动机器人可以以 60%的成功率到达新目标时,到达旧目标得成功率为 81%。奖励曲线图如图 9 所示。

3.4　仿真结果对比

在环境 1 中,与原始模型对比,改进模型新目标的训练成功率在 600 个回合之前较原始模型上涨较缓,但在 600 个回合到 870 回合之间上涨超过了原始模型,并在 1 100 个回合之后上涨速度变快并超过了原始模型,达到了 81%,如图 10 所示,改进后的视觉导航模型在环境 1

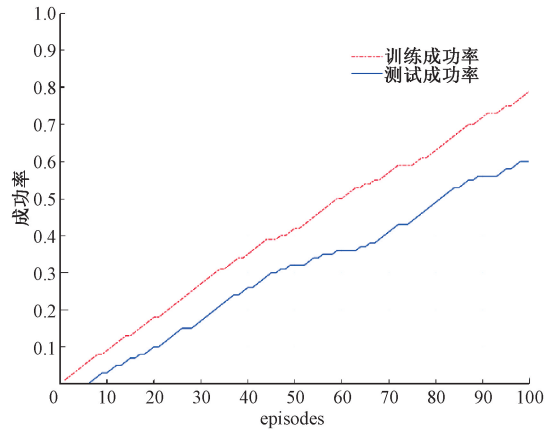


图 8　环境 2 改进模型成功率

Fig. 8　Improve model success rate in Map 2

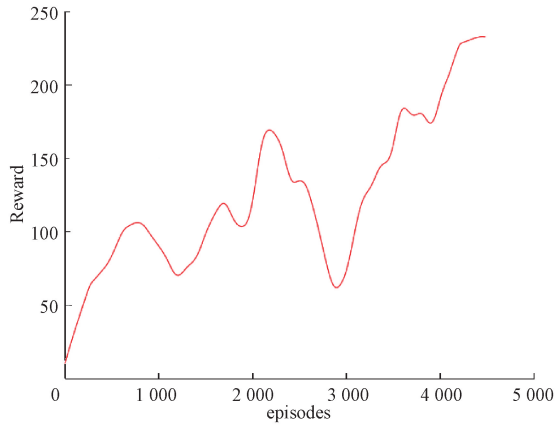


图 9　环境 2 改进模型奖励曲线图

Fig. 9　Improve model reward in Map 2

的训练过程中较优于原始模型。

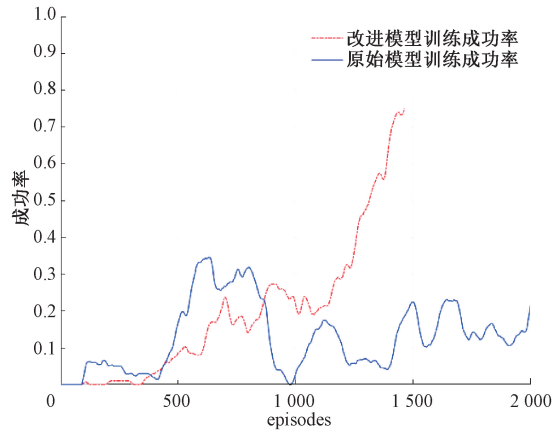


图 10　环境 1 模型训练成功率对比

Fig. 10　Comparison of training success rates in Map 1

改进模型在环境 1 中得的新目标训练成功率在 360

回合之前一直高于原始模型,在350~440回合数之间低于原始模型,而后快速上涨,超过了原始模型,改进后的视觉导航模型在环境1的推理过程中较优于原始模型,如图11所示。

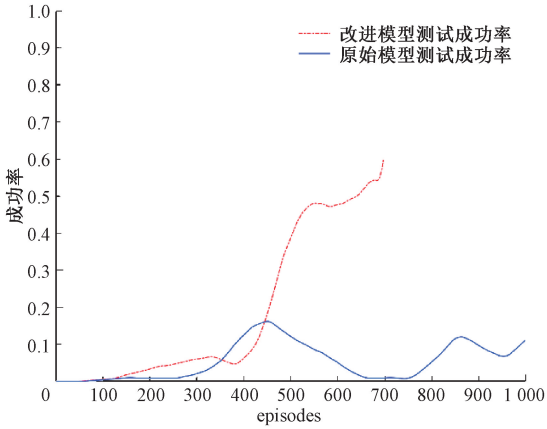


图11　环境1模型推理成功率对比

Fig. 11　Comparison of inference success rates in Map 1

在环境2中,与原始模型对比,改进模型训练目标的训练成功率增长始终高于原始模型,最终成功率达到79%,改进后的视觉导航模型在环境2的训练过程中优于原始模型,如图12所示。

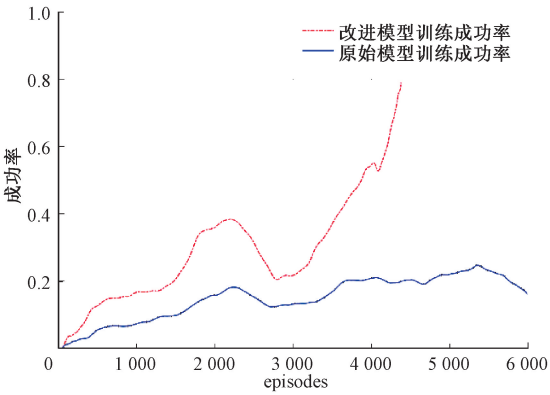


图12　环境2模型训练成功率对比

Fig. 12　Comparison of training success rates in Map 2

在环境2中,改进模型的推理目标训练成功率始终高于原始模型,在1 002回合数时新目标到达成功率就到达60%,改进后的视觉导航模型在环境2的推理过程中优于原始模型,如图13所示。

新提出的网络结构在两个环境中到达新目标的成功率都超过60%,且训练速度与前序算法相比,收敛更快。实际上,最终环境1到达新目标的成功率为93%,环境2到达新目标的成功率为91%,如表1所示。

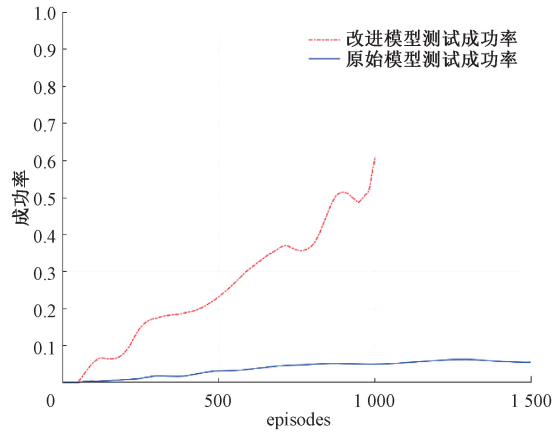


图13　环境2模型推理成功率对比

Fig. 13　Comparison of inference success rates in Map 2

表1　视觉导航模型成功率对比

Table 1　Visual navigation model success rate comparison

	改进旧目标成功率	原始旧目标成功率	改进新目标成功率	原始新目标成功率
环境1	81%	75%	93%	81%
环境2	79%	62%	91%	69%

视觉导航模型推理新目标路径如图14~16所示。

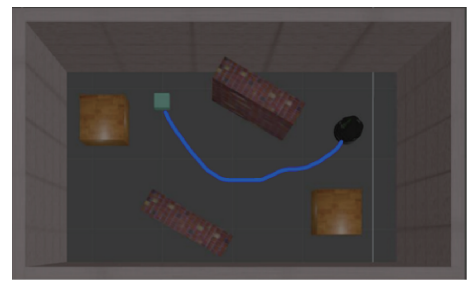


图14　环境1推理目标路径图

Fig. 14　Path to new target in Map 1

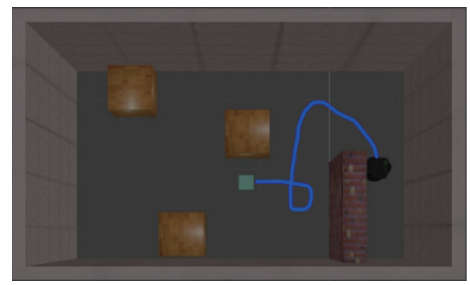


图15　环境2推理目标1路径图

Fig. 15　Path to new target1 in Map 2

4　结　论

本文在已有的深度强化学习视觉导航框架下进行改

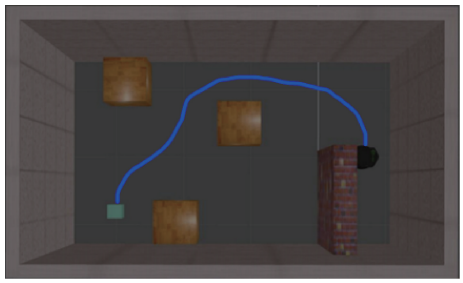


图 16　环境 2 推理目标 2 路径图

Fig. 16　Path to new target 2 in Map 2

进,提出一种新的融合 LSTM 和 PPO 算法的移动机器人视觉导航模型,根据移动机器人第一视角观测到的 RGB-D 图像以及目标点在移动机器人坐标系下的极性坐标,通过改进网络结构及重新设计奖励函数,训练过程能够以较快速度收敛,在 gazebo 仿真环境下,移动机器人到达推理目标的成功率超过 90%,该方法较前序算法有一定提升,在 gazebo 仿真环境中有较好的导航效果。未来的工作将考虑实现更大空间下的移动机器人视觉导航,并应用到实际移动机器人上,并进一步提高模型对于场景的泛化能力。

参考文献

[1] MA L, LIU Y, CHEN J, et al. Learning to navigate in indoor environments: From memorizing to reasoning[J]. ArXiv Preprint,2019, arXiv:1904.06933.

[2] KHAIRUDDIN A R, TALIB M S, HARON H. Review on simultaneous localization and mapping (SLAM)[C]. 2015 IEEE International Conference on Control System, Computing and Engineering (ICCSCE),2015: 85-90.

[3] ZHU K, ZHANG T. Deep reinforcement learning based mobile robot navigation: A review[J]. Tsinghua Science and Technology, 2021, 26(5): 674-691.

[4] CADENA C, CARLONE L, CARRILLO H, et al. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age[J]. IEEE Transactions on Robotics, 2016, 32(6): 1309-1332.

[5] 孙龙龙, 江明, 焦传佳. 基于运动矢量的改进视觉 SLAM 算法[J]. 电子测量与仪器学报, 2020, 34(9): 23-31.
SUN L L, JIANG M, JIAO CH J. Improved visual SLAM algorithm based on the motion vector[J]. Journal of Electronic Measurement and Instrumentation, 2020, 34(9): 23-31.

[6] ZENG F, WANG C, GE S S. A survey on visual navigation for artificial agents with deep reinforcement learning[J]. IEEE Access, 2020, 8: 135426-135442.

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

[8] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[J]. ArXiv Preprint,2015, arXiv:150902971.

[9] MNIH V, BADIA A P, MIRZA M, et al. Asynchronous methods for deep reinforcement learning[C]. International Conference on Machine Learning, 2016: 1928-1937.

[10] SCHULMAN J, LEVINE S, ABBEEL P, et al. Trust region policy optimization[C]. International Conference on Machine Learning, 2015: 1889-1897.

[11] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[J]. ArXiv Preprint,2017, arXiv:170706347.

[12] 袁浩, 刘紫燕, 梁静, 等. 融合 LSTM 的深度强化学习视觉导航[J]. 无线电工程, 2022, 52(1): 161-167.
YUAN H, LIU Z Y, LIANG J, et al. Visual-based navigation algorithm with LSTM and deep reinforcement learning[J]. Radio Engineering, 2022, 52(1): 161-167.

[13] ZHU Y, MOTTAGHI R, KOLVE E, et al. Target-driven visual navigation in indoor scenes using deep reinforcement learning[C]. 2017 IEEE International Conference on Robotics and Automation (ICRA), 2017: 3357-3364.

[14] NWAONUMAH E, SAMANTA B. Deep reinforcement learning for visual navigation of wheeled mobile robots[C]. 2020 SoutheastCon, 2020: 1-8.

[15] DEVO A, MEZZETTI G, COSTANTE G, et al. Towards generalization in target-driven visual navigation by using deep reinforcement learning[J]. IEEE Transactions on Robotics, 2020, 36(5): 1546-1561.

[16] YOKOYAMA K, MORIOKA K. Autonomous mobile robot with simple navigation system based on deep reinforcement learning and a monocular camera[C]. 2020 IEEE/SICE International Symposium on System Integration (SII), 2020: 525-530.

[17] VAN HOUDT G, MOSQUERA C, NÁPOLES G. A review on the long short-term memory model[J]. Artificial Intelligence Review, 2020, 53(8): 5929-5955.

[18] STAUDEMEYER R C, MORRIS E R. Understanding LSTM—a tutorial into long short-term memory recurrent neural networks[J]. ArXiv Preprint, 2019,

arXiv:1909.09586.

[19] WIERING M A, VAN OTTERLO M. Reinforcement learning[J]. Adaptation, Learning, and Optimization, 2012, 12(3): 729.

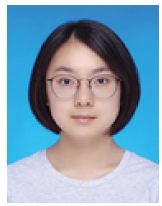
[20] 杨思明, 单征, 丁煜, 等. 深度强化学习研究综述[J]. 计算机工程, 2021, 47(12): 19-29.
YANG S M, SHAN ZH, DING Y, et al. Survey of research on deep reinforcement learning[J]. Computer Engineering, 2021, 47(12): 19-29.

[21] 刘建伟, 高峰, 罗雄麟. 基于值函数和策略梯度的深度强化学习综述[J]. 计算机学报, 2019, 42(6): 1406-1438.
LIU J W, GAO F, LUO X L. Survey of deep reinforcement learning based on value function and policy gradient[J]. Chinese Journal of Computers, 2019, 42(6): 1406-1438.

[22] ARULKUMARAN K, DEISENROTH M P, BRUNDAGE M, et al. Deep reinforcement learning: A brief survey[J]. IEEE Signal Processing Magazine, 2017, 34(6): 26-38.

[23] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[J]. ArXiv Preprint, 2017, arXiv:1707.06347.

作者简介



张仪,2018 年于华中农业大学获得学士学位,现为中国科学院大学硕士研究生,主要研究方向为强化学习。
E-mail:y.zhang@siat.ac.cn

Zhang Yi received her B. Sc. degree from Huazhong Agricultural University in 2018. Now she is a M. Sc. candidate at University of Chinese Academy of Sciences. Her main research interest includes reinforcement learning.

冯伟(通信作者),2001 年于华中科技大学获得学士学位,2006 年于华中科技大学获得博士学位,2016 任香港中文大学研究员,现为中国科学院深圳先进技术研究院研究员,主要研究方向为机器人与数字智造。
E-mail:wei.feng@siat.ac.cn

Feng Wei (Corresponding author) received the B. Sc. degree from the School of Materials Science and Engineering in 2001 and the Ph. D. degree in 2006, both from the Huazhong University of Science and Technology, Wuhan, China. He was appointed as a research fellow with the Chinese University of Hong Kong in 2016. He is now a full professor in Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences. His main research interests include robotics and intelligent system.