智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

场景图谱驱动目标搜索的多智能体强化学习

陆升阳, 赵怀林, 刘华平

引用本文:
陆升阳,赵怀林,刘华平. 场景图谱驱动目标搜索的多智能体强化学习[J]. 智能系统学报, 2023, 18(1): 207–215.
LU Shengyang,ZHAO Huailin,LIU Huaping. Multi–agent reinforcement learning for scene graph–driven target search[J]. *CAAI Transactions on Intelligent Systems*, 2023, 18(1): 207–215.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202111034

DOI: 10.11992/tis.202111034
网络出版地址: https://kns.cnki.net/kcms/detail/23.1538.TP.20220928.1957.002.html

场景图谱驱动目标搜索的多智能体强化学习

陆升阳[1], 赵怀林[1], 刘华平[2]
（1. 上海应用技术大学 电气与电子工程学院, 上海 201418; 2. 清华大学 计算机科学与技术系, 北京 100084）

摘　要: 针对强化学习在视觉语义导航任务中准确率低, 导航效率不高, 容错率太差, 且部分只适用于单智能体等问题, 提出一种基于场景先验的多智能体目标搜索算法。该算法利用强化学习, 将单智能体系统拓展到多智能体系统上将场景图谱作为先验知识辅助智能体团队进行视觉探索, 利用集中式训练分布式探索的多智能体强化学习的方法以大幅度提升智能体团队的准确率和工作效率。通过在 AI2THOR 中进行训练测试, 并与其他算法进行对比证明此方法无论在目标搜索的准确率还是效率上都优先于其他算法。

关键词: 多智能体; 强化学习; 视觉语义导航; 场景图谱; 先验知识; 分布式探索; 集中式训练; 目标搜索
中图分类号: TP391　　**文献标志码:** A　　**文章编号:** 1673−4785(2023)01−0207−09

中文引用格式: 陆升阳, 赵怀林, 刘华平. 场景图谱驱动目标搜索的多智能体强化学习 [J]. 智能系统学报, 2023, 18(1): 207–215.
英文引用格式: LU Shengyang, ZHAO Huailin, LIU Huaping. Multi-agent reinforcement learning for scene graph-driven target search[J]. CAAI transactions on intelligent systems, 2023, 18(1): 207–215.

Multi-agent reinforcement learning for scene graph-driven target search

LU Shengyang[1], ZHAO Huailin[1], LIU Huaping[2]
(1. School of electrical and Electronic Engineering, Shanghai Institute of Technology, Shanghai 201418, China; 2. Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

Abstract: To solve the problems of reinforcement learning in the visual semantic navigation task, such as low accuracy, low navigation efficiency, poor fault tolerance rate, and the suitability of only some problems for a single agent, we propose a multi-agent target search algorithm based on scene prior. This algorithm extends the single-agent system to a multi-agent system through reinforcement learning. It mainly includes two aspects: first, a scene atlas is used as prior knowledge to assist the agent team in visual exploration; second, the multi-agent reinforcement learning method of centralized training and distributed exploration is used to greatly improve the accuracy and work efficiency of the agent team. Training tests in AI2THOR and comparison with other algorithms prove that this method is superior to other algorithms in target search accuracy and efficiency.

Keywords: multi-agent; reinforcement learning; visual semantic navigation; scene graph; prior knowledge; distributed exploration; centralized training; target search

在环境中高效准确地导航能力是智能行为的基础, 多年来一直是机器人研究的重点。视觉语义导航作为智能体视觉导航极为重要任务之一, 要求机器人在未知环境中, 根据给定的目标, 利用观测到的视觉信息导航到目标位置。这在人工智能领域的研究中都有着极为深远的意义和前景, 如战场搜索, 医疗服务, 现场救灾等。与此同时, 视觉语义导航的发展对很多其他相关工作也有一定程度上的促进, 如视觉语言导航[1]、具身问题回答[2]、视觉对话导航[3] 等。

传统的导航方法[4] 通常利用环境地图进行导航, 并将导航任务分成绘制地图、定位和路径规划 3 个步骤。然而地图在不可见环境中是无法使用的。最近, 强化学习 (reinforcement learning, RL) 常用于视觉导航中。文献 [5] 开发了 3D 迷宫

收稿日期: 2021−11−17.　**网络出版日期:** 2022−09−30.
基金项目: 国家自然科学基金项目（U1613212）.
通信作者: 刘华平. E-mail: hpliu@tsinghua.edu.cn.

环境中的导航方法，并引入了深度预测和环路闭合分类任务，以提高导航性能。不仅如此，很多研究利用强化学习作为动作决策模块，进行控制机器人在未知环境中进行导航，如 A3C[6]、PPO[7]、DDPG[8]、DQN[9] 等。此外，文献 [10] 引入了贝叶斯关系记忆，以探索房间之间空间布局为目的，并非以用最少的步骤引导智能体导航到指定的目标处为目的。为了提高智能体在未知环境中的导航能力，文献 [11] 提出了一种自适应元学习的方法。此外，注意力机制[12]、记忆力机制[13] 也被用于基于强化学习的视觉导航中。虽然大量的方法为提升智能体导航的稳定性和准确率做出了一定的贡献，然而智能体导航仍存在着智能体搜索效率低下，准确率难以提升的问题。与此同时，上述的方法只适用于单个智能体，当智能体搜索的目标有多个时，单个智能体只能连续地搜索整个场景，这使得智能体的容错性较差，失误的概率大大增加。

多智能体合作解决搜索目标的方法无疑成为解决上述问题的最优方式之一。多智能体之间的协作也具有重要意义。近年来，人们提出了多智能体决策模型来解决多个智能体的具身任务，如捉迷藏游戏[14]、雷神之锤游戏[15] 等。但是一个错误的协作策略反而会降低智能体工作的效率，因此智能体之间的协作策略[16-17] 是多智能体系统研究的重点。许多工作将多智能体强化学习的重心放在通信机制[18-20] 上，文献 [21] 将通信机制应用到一个多智能体的导航任务上，该任务中所有的智能体通信找到同一个目标。不仅如此，文献 [22] 提出单独推断通信，让智能体学会去和谁通信，什么时候去通信，这大大减少了通信的开销。然而在探索的过程中通信虽然对智能体之间的协作有着一定的帮助，但是在实际探索过程中仍会存在通信中断，带宽过小等问题。因此文献 [23] 提出了一种集中式训练、分布式探索的方法，将智能体之间的通信在集中式的训练中完成。

此外，当智能体在探索环境的时候，场景先验知识可以提取各个物体之间的空间关系，可以有效地帮助智能体以更少的步骤向目标物体移动。在视觉语义导航任务中，场景先验知识被广泛地应用于构建空间和语义关系[24-27]。

为了研究多智能体在复杂任务中的协作，本文提出了一种多智能体视觉语义导航的框架，在该框架中，多智能体利用第一人称视觉观测以及物体标签，与其他智能体进行协作寻找一个或多个目标。利用集中化训练分布式搜索的方式，减少智能体在探索过程中的通信，同时提高了探索的策略。

本文结合了多智能体强加学习以及场景图谱的优点，来解决未知环境中目标搜索的问题。主要贡献如下：

1) 将单智能体目标搜索扩展到了多智能上，并使用场景图谱作为语义先验提高了目标搜索的效率。

2) 提出了利用语义地图进行多智能体集中式训练，分布式执行的网络框架来解决目标搜索问题。

3) 本文将此框架应用到 AI2THOR 中，并将其与单智能体目标搜索对比，实验证明本文提出的框架具有一定的先进性。

1　问题描述

本次实验的目标是给定智能体团队一个目标标签，智能体团队可以根据视觉信息，一起协同探索整个场景，一直到找到目标。

在未知场景s中，智能体团队的视觉信息被记作O，$O=\{O_t^1,O_t^2,\cdots,O_t^k\}$，其中$O_t^k$为第$k$个智能体$A^k$在时间点$t$的视觉信息。在初始时，智能体团队中的每一个智能体被初始化在房间内的随机位置，随机姿态，记作$P_0^k=\{x_0^k,z_0^k,\theta_0^k,\phi_0^k\}$，其中$(x_0^k,z_0^k)$代表智能体在初始时刻被初始化在房间内的随机位置，$\theta_0^k$为初始旋转角，$\phi_0^k$为初始俯仰角。初始时智能体团队接收到目标的标签$T$作为输入。在每一个时间点，智能体团队从视觉传感器接收到视觉信息O_t以其里程计接收到位姿信息P_t。其中视觉信息O_t由第一人称的 RGB 以及深度信息组成。

在每一个时间点t，第k个智能体A^k从视觉传感器中接收到第一人称视觉信息O_t^k，从里程计中获取位姿信息P_t^k。利用视觉信息以及位姿信息来预测出对应的语义地图m_t^k。每一个智能体学习其对应的导航策略π^k来决策对应的预期导航点L^k。其中$L_t^k=\pi^k(m_{t-1}^k,T,L_{t-1}^k)$为到达预期目标点，智能体通过路径规划的方式来获取下一个动作A_t^k，$A_t^k\in\{\mathrm{M,U,D,L,R}\}$，其中M为前进，U为向上看，D为向下看，L为左转，R为右转。当每一个智能体中执行动作A_t^k后则会更新其视觉输入为O_{t+1}^k以及里程计的输入P_{t+1}^k。如此往复直到找到所有的目标。

2　总体框架

本文主要提出了一种在交互式环境中基于多智能体强化学习的目标搜索模型，该模型主要由 4 个部分组成，如图 1 所示，分别为语义映射模

块, 对象关系特征提取模块, 特征融合模块以及动作决策模块。语义决策模块将第一人称视觉映射成自上而下的语义地图特征。对象关系特征利用场景先验知识提取模块将第一人称的 RGB 信息中的对象关系。特征融合模块则是将当前的对象特征关系以及语义映射向量和先前状态进行特征融合。而动作决策模块则是决策下一个预期目标并生成智能体需要执行的下一个动作。

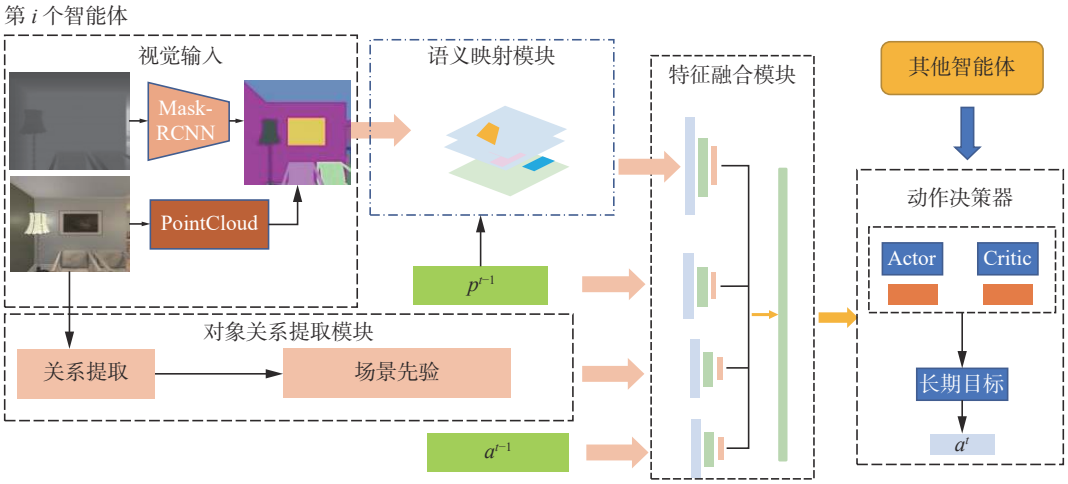


图 1　总体框架

Fig. 1　Overall framework

2.1　语义映射生成

本文使用类似于文献 [28] 的方法来生成语义映射地图。语义映射模块首先使用预训练的 Mask-RCNN[29] 来获取 RGB 图像中的语义分割结果, 然后根据深度信息以及智能体的位姿将每一个像素映射到三维空间中得到点云信息。然而点云信息由于其计算参数量要求过大所以将其映射到二维语义地图上。将地图上的点云在高度上求和得到自上而下的二维语义地图。为了减少智能体行走重复的轨迹以及生成重复的预期目标点, 因此将智能体的行走轨迹以及生成的预期目标点都考虑在内, 在二维语义地图的基础上加入两个图层。第一层为智能体的行走轨迹, 第二个图层为预期目标点的位置。所以生成的二维语义地图的大小为$(C+2)\times L\times W$, C为二维语义地图的通道, 即物体类别的数量, $L\times W$分别为语义地图的长和宽。具体语义地图生成过程如图 2 所示。

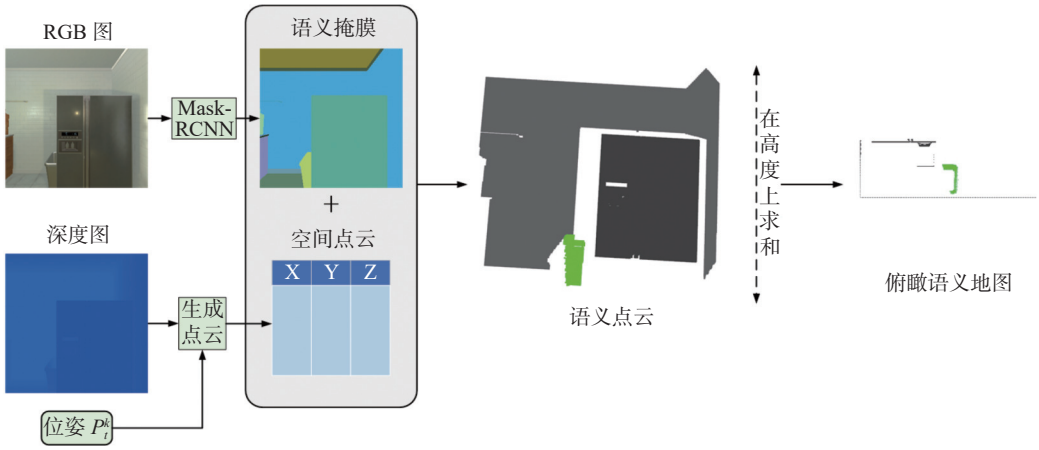


图 2　空间语义地图生成过程

Fig. 2　Spatial semantic map generation process

在导航的过程中, 可以得到每一步智能体的位姿P_t^k以及视觉信息O_t^k, 为了实验的准确率, 本文使用 AI2THOR[30] 中提供的 Ground Truth 语义分割生成语义结果生成空间语义映射$\boldsymbol{m}_t^k$, 空间语义映射的大小为$(C+2)\times L\times W$, 其中第 1 层和第 2 层分别为智能体走过的区域以及预期目标生成的位置。在初始时, 语义映射被初始化为一个全 0 矩阵, 记作$\boldsymbol{m}_0^k=[0]^{(C+2)\times L\times W}$, 随着整个场景不断

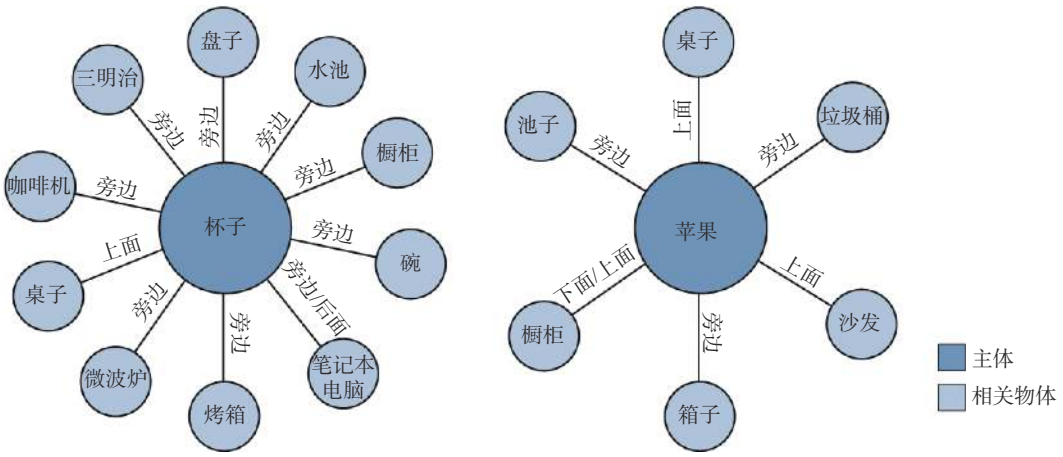
被探索语义映射被不断更新。

2.2　对象关系特征提取

采用类似于文献 [31] 的方法，将语义知识纳入强化学习框架，并使用图卷积网络 (graph convolutional networks, GCNs)[32] 整合之前的知识，在智能体接收到环境信息时动态的更新保存它们。

2.2.1　先验知识的设定

本文将场景先验知识以无向图的形式表现，场景图谱$G=\{V,E\}$，V中的节点表示不同的物体种类，E表示两个类别物体之间特殊的位置关系。如图 3 所示。



(a) 杯子场景图谱　　(b) 苹果场景图谱

图 3　场景图谱示意图

Fig. 3　Schematic diagram of scene map

本文使用视觉基因组数据集[32] 作为来源，根据在本次实验中使用的 AI2THOR 中出现的所有对象的类别来构建知识图。每一个类别都表示为图中的一个节点。在视觉基因组数据集中对象关系出现频率大于 3 时才会在两个节点中使用变来链接。

2.2.2　图卷卷积网络设计

GCNs 作为图神经网络对图结构的扩展，其目标是学习给定图形$G=\{V,E\}$的函数表示。将所有的节点归纳为特征矩阵$\boldsymbol{F}=[F_1\ F_2\ \cdots\ F_{|V|}]$。图结构用二进制邻接矩阵$\boldsymbol{A}$表示，将矩阵$\boldsymbol{A}$[31] 进行标准化得到矩阵$\hat{\boldsymbol{A}}$。GCNs 输出的每一个节点表示为$\boldsymbol{Z}=[z_1\ z_2\ \cdots\ z_{|V|}]$。因此可以得到：

$$\boldsymbol{H}^{(l+1)}=f(\hat{\boldsymbol{A}}\boldsymbol{H}^{(l)}\boldsymbol{W}^{(l)}) \tag{1}$$

式中：$\boldsymbol{H}^{(0)}=\boldsymbol{X}$，$\boldsymbol{H}^{(L)}=\boldsymbol{Z}$，其中$\boldsymbol{W}^{(l)}$为第$l$层的参数，而$L$为 GCNs 的总层数。

如图 4 所示，使用 3 层的 GCN，输入为 RGB 图像，利用 ResNet34 对图像进行预处理，得出根据当前图像特征向量为一个 1 000 维的向量，对于不同的节点，将当前的图像特征向量映射为 512 维的特征向量，然后将所有类别的名称用单词嵌入分别映射成为 512 维特征向量，再将两个特征向量拼接，为每个图节点形成 1 024 维的联合表示。图卷积神经网络的输入为临接矩阵$\boldsymbol{A}$和节点特征向量，前两层的输出为 1 024 维的潜在特征，最后一层的输出为每个节点输出的值，得到特征向量$|\boldsymbol{V}|$。该特征向量为当前场景和环境上下文的语义编码信息。最后将这个特征向量映射到 512 维的特征向量$\boldsymbol{f}_{K,t}^{k}$。将 512 维的特征向量作为输入特征协助智能体进行视觉导航。

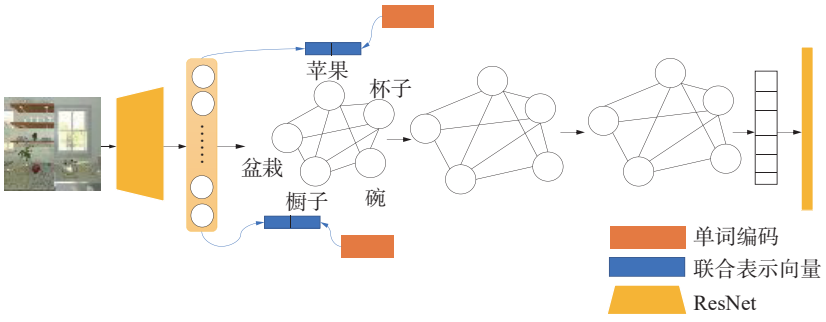


图 4　对象关系特征提取网络

Fig. 4　Object relationship feature extraction network

2.3 特征融合

主要使用上述提到的空间语义映射、对象关系特征以及当前的位置以及上一个时刻的动作作为输入, 利用特征融合模块生成智能体当前状态特征。在时间点t上, 我们使用 CNN 将智能体的语义映射结果进行编码, 生成智能体语义特征结果$\boldsymbol{f}_t^k$。利用全连接层将当前的位姿编码成为位姿特征向量$\boldsymbol{f}_{L,t}^k$, 再利用全连接层将上一个时刻的动作编码生成上一时刻的动作特征信息$f_{a,t}^k$。最后, 我们将语义特征向量$\boldsymbol{f}_t^k$, 位姿特征向量$\boldsymbol{f}_{L,t}^k$, 动作特征信息$f_{a,t}^k$以及空间位置特征关系$f_{K,t}^k$进行拼接得到融合特征向量$\boldsymbol{f}_{C,t}^k$。

2.4 动作决策模块

该模块主要使用多智能体强化学习的方法来解决多智能体视觉探索问题。长期目标是让智能体团队寻找给定标签的物体, 因此智能体团队必须根据视觉导航到目标物体所在区域。

2.4.1 多智能体强化学习模型

此部分主要设计了一个分布式探索, 集中式训练的多智能体强化学习演员–评论家模型。使用 2.3 节所叙述的特征向量$\boldsymbol{f}_{C,t}^k$作为模型的输入, 根据演员网络输出智能体能到达的短期目标位置。短期目标的格式$\{x,z,\phi\}$。其中x、z分别为短期目标的位置, $\phi\in\{-1,0,1,2\}$表示智能体的俯仰角, 当$\phi=-1$时, 智能体的俯仰角为$-30°$, 以此类推。短期目标会引导智能体去探索整个场景, 当智能体的视觉传感器接收到相关语义目标的视觉信息时, 语义映射m_t对应层将会变得非 0, 对象关系特征向量也会给出对应的空间对象特征向量。演员网络则会根据这些融合特征引导智能体去探索更多的场景, 直到寻找到目标。

2.4.2 训练策略

本节主要介绍了一种集中式多智能体训练的方式。设定目标为$\{t_1,t_2,\cdots,t_n\}$, 智能体团队每一次采用 M 指令环境都会给出−0.01 的奖励, 每一次智能体接收到与目标相同语义类别但是不同实例的信息时都会给出 $1/n$ 的奖励。因此每个智能体的奖励可以定义为

$$R^k=\sum_{i=1}^{N}\frac{1}{n}-0.01b \tag{2}$$

式中: N代表智能体看到目标标签类别但是不同实例的目标的次数; b代表智能体采用通过前进指令的次数; n 代表在这个环境内目标标签的实例数量。因此, 智能体团队的总奖励则可定义为

$$R=\sum_{i=1}^{k}R_i \tag{3}$$

式中: k表示智能体的数量。根据多智能体强学习模型, 演员网络生成每r个步骤的短期目标 (r设置成 15), 智能体如果在r步骤内已经导航到短期目标, 那么智能体则会随机选择{ L,U,D,R}中的步骤熟悉整个场景, 如果r步骤内依然没到达短期目标, 则重新生成短期目标, 进行导航。

本文采用类似于文献 [33] 的方法, 对每个智能体分别设计了一个演员和一个评论家网络。对于演员网络而言, 每个智能体仅观测到局部信息, 并且利用局部信息进行导航获得对应奖励。演员网络训练的目标就是根据局部观测的结果期望可以获得奖励最大化, 学习的是一个从部分可见的融合向量到动作$a_i^{(k)}$的映射。因此演员网络的损失可以定义为

$$L(\theta)=\left[\frac{1}{Bk}\sum_{i=1}^{B}\sum_{m=1}^{k}\min(\gamma_{\theta i}^{(m)}A_i^{(m)},\right.$$
$$\left.\mathrm{clip}(\gamma_{\theta i}^{(m)},1-\epsilon,1+\epsilon)A_i^{m}\right)\Big]+\sigma\frac{1}{Bk}\sum_{i=1}^{B}\sum_{m=1}^{k}S[\pi_\theta(f_{c,t}^{(m)})]$$

式中: $\gamma_{\theta i}=\dfrac{\pi_\theta(a_i^{(k)}|f_{ct}^{(k)})}{\pi_{\mathrm{old}}(a_i^{(k)}|f_{ct}^{(k)})}$; 优势函数$A_i^{(m)}$是采用图自编码 (graph auto-encoders ,GAE) 方法计算的; S表示策略的熵; σ是一个超参数; ε 为贪婪搜索的一个设定参数, 操作 c1ip 将$\gamma_{\theta i}^{(m)}$的值限制到区间$(1-\varepsilon,1+\varepsilon)$内。

同时对于每一个演员网络都引入一个评论家网络, 评论家网络主要为了拟合出一个价值函数$V_\phi(f)$, 该价值函数只在中心化训练的过程中使用。评论家网络的输入不仅为当前智能体的状态, 动作信息更加包含了其他智能体的信息。在中心化训练的过程中, 每一个智能体都维护了一个大小为 500 000 字节的缓冲内存用来存储经验池。在每一个演员网络的运行过程中通过输入的融合特征决策出短期目标并获得对应的奖励, 将其存放到经验池中。评论家网络的输入为所有智能体融合特征向量的拼接, 通过训练学习出一个从融合特征向量$\boldsymbol{f}_{C,t}$到奖励的映射。评论家网络损失函数可定义为

$$L(\boldsymbol{f})=\frac{1}{Bk}\sum_{i=1}^{B}\sum_{m=1}^{k}(\max[(V_\phi(\boldsymbol{f}_{C,t}^{m})-\hat{R}_i)]^2)$$

式中: B为batch size的大小, k为智能体的数量, $\hat{R}_i$为折扣奖励, 可定义为

$$\hat{R}_i=R_i+rR_{i+1}+r^2R_{i+2}+\cdots+r^{B-i}R_B$$

其中，r为折扣系数，在本文中设置为 0.995。评论家网络针对当前状态计算生成期望奖励，并代替真实奖励值形成策略梯度，用于演员网络的参数更新。

3　仿真实验

3.1　数据集设置

本次实验在 AI2THOR[29] 仿真环境中训练和测试模型。AI2THOR 仿真环境包含了 4 类房间，一共 120 个。对于每一类房间选取其中两个房间，并随机初始化它们的物品摆放，共生成 800 个模拟场景作为训练集，200 个模拟场景作为测试集。

主要使用 IQuAD v1 数据集[33] 作为基本数据集，该数据集主要包括了存在性问题，计数问题以及空间关系问题。然而该数据集更多考虑了智能体与虚拟环境之间的交互，因此将 IQuAD v1 数据集进行重构。将存在性问题中在仿真环境内仅出现一次的问题整理出来，并提取其目标的标签作为单目标寻找数据集。将计数问题中的目标在仿真环境中出现两次、三次的问题进行整理并提取出目标标签，作为实验中单目标多个目标实例的数据集。

智能体团队在一开始时随机初始化在房间内的任意一个位置，智能体的视觉输入有 RGB 图像以及深度图组成，为$k\times4\times300\times300$像素的矩阵。智能体的位姿矩阵为$\{x_t^k,z_t^k,\theta_t^k,\varphi_t^k\}$，大小为$k\times4$。

3.2　训练细节

在数据集内训练了超过 3×10^6 次，对于每一个仿真场景来说，对于单目标搜索问题而言，当智能体团队在环境中获得的总奖励超过 0 并且其中任意一个智能体距离目标物体小于 0.25 m 时则代表智能体团队完成探索，本次探索结束，更换下一个场景。而对于单目标多次出现的数据集而言，当智能体团队获得的总奖励超过 0.8 则代表本次实验成功。不仅如此，当智能体团队在这个场景中导航了超过 1 000 步时，则表示智能体团队已经探索完整个场景，并且对场景足够熟悉，此次导航失败。

3.3　评价函数

本文主要使用成功率 R 以及平均路径长度 P 两种评价函数来验证算法的效果。成功率被定义为

$$R=\frac{1}{X}\sum_{i=1}^{X}S_i$$

式中：X代表测试实验的总次数，S_i代表第i次实验是否成功，若该次实验成功，则$S_i=1$，反之则为 0。该评价函数验证了模型的准确率。

平均路径长度则被定义为

$$P=\frac{1}{X_{\text{succ}}}\sum_{i=1}^{X_{\text{succ}}}\left(\frac{1}{k}\sum_{j=1}^{k}l_j\right)$$

式中：X_{succ}代表成功的次数；l_j代表第j个智能体导航的步骤。$\frac{1}{k}\sum_{j=1}^{k}l_j$代表在该次实验中智能体导航的平均步骤。整个公式表示智能体导航成功的平均步骤数量。该式子衡量整个实验的时效性，即P越小代表智能体团队寻找到目标物体的速度越快，整体的时效性也越好，整个模型的性能也就越好。

3.4　定量实验分析

为了评估本文提出的模型中语义映射的有效性，我们删除了语义映射模块，将 RGBD 信息作为输入形成了 RGBD+RL。为评估本文算法中场景先验知识的效果，将对象关系特征提取模块删除形成无场景图谱算法。同时为体现多智能体集中式学习分布式探索的算法，将本文算法与随机行走进行对比。分别在单目标单个实例的场景以及单目标多个实例的场景进行训练和测试。将本文的算法与上述算法进行对比，实验结果见表 1 和表 2。

表 1　单目标单次存在搜索结果

Table 1　Single-target single-time existence search results

算法	N=2		N=3		N=4		N=5	
	R/%	P/步	R/%	P/步	R/%	P/步	R/%	P/步
本文算法	**52.89**	**150.25**	**57.96**	**106.27**	**60.43**	**80.56**	**76.51**	**66.57**
无场景图谱算法	43.57	165.57	52.65	115.68	52.75	98.69	69.62	70.57
随机行走	10.53	956.65	16.25	1 000	27.59	976.94	37.65	978.86
RGB+RL	22.93	326.34	24.53	278.85	30.24	224.43	29.93	215.55

表 2　单目标多次存在搜索结果

Table 2　Single-target multi-times existence search results

算法	*N*=2		*N*=3		*N*=4		*N*=5	
	R/%	*P*/步	*R*/%	*P*/步	*R*/%	*P*/步	*R*/%	*P*/步
本文算法	**24.2**	**150.25**	**25.64**	**106.27**	**40.59**	**80.56**	**41.32**	**66.57**
无场景图谱算法	20.15	161.28	21.56	115.35	25.64	100.31	30.12	96.56
随机行走	0	1 000	5.31	998.96	5.12	999.83	7.53	983.62
RGB+RL	12.22	450.33	15.33	420.57	25.42	310.43	23.35	300.32

分析表 1 和表 2 得到以下结论:

1)在本次实验中, 使用 AI2THOR 提供的 ground-truth 语义分割结果代替预训练的 mask-RCNN[29] 结果。实验结果证明, 本文提出的中心化训练, 分布式执行的框架在目标搜索方面有着明显优势。

2)由于目标物体较小, 而且在仿真环境中目标可能在“垃圾桶”或者“盒子”里, 不容易被发现, 所以实验结果中的成功率无法达到很高。

3)随着智能数量的增加, 多智能体目标搜索的准确率和实效性有着很大的提升, 但是当智能体团队的数量从 4 变成 5 时效果提升的并不是很多。这是由于整体场景并不是很大, 4 个智能体已经足够有效地探索整个场景。

4)使用对象关系特征提取模块在能有效的提升整体模型的准确率和实效性。

5)对于表 2 而言, 目标物体可能存在 2 个或者 3 个, 设置的是对于这一类探索, 必须把所有的目标全部找到才算探索成功, 因此实验的成功率并不是特别高

3.5　定性实验分析

为了更好地展示算法实现的整个过程, 本节展示了在 AI2THOR 中 3 个智能体对于单目标多实例的探索全过程, 分别如图 5 所示。

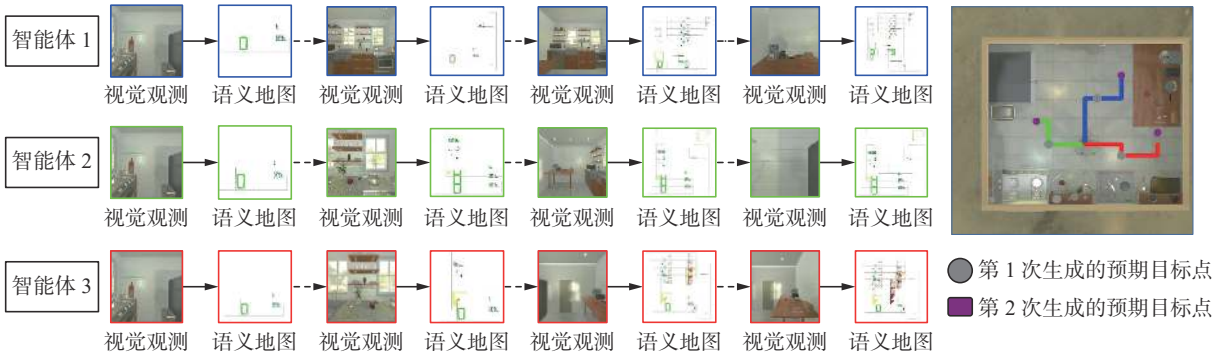


图 5　两个智能体去寻找两个勺子示意图

Fig. 5　Two agents look for two spoon diagrams

在实验设置中, 3 个智能体组成智能体团队互相协作去寻找目标——勺子, 最终成功找到目标。在实验的设计过程中, 针对单目标单个实例搜索问题, 智能体团队会协作探索环境, 当其中一个智能体的视觉传感器中检测到了目标物体, 并距离它 0.25 m 以内时, 则会发出完成的指令, 则智能体团队立刻停止探索。而对于单目标多个实例存在的问题而言, 智能体找到目标并导航到目标周围, 当距离目标小于 0.25 m 并且目标在使用范围内则会获得对应奖励。而智能体团队则是一直导航一直到获得的团队总奖励超过 0.8 的时候才会停止, 因为当奖励超过 0.8 则意味着智能体团队检测到了所有的目标物体。

如图 5 所示, 由于 AI2THOR[30] 中的房间不大, 所以本次实验设置中 3 个智能体初始化在房间任意位置的同一位姿。初始时智能体只有视觉以及位姿信息, 语义地图为全 0 的矩阵, 但是由于有视觉以及位姿信息的输入, 语义地图逐渐被填充。根据语义地图以及场景先验知识, 智能体不断地去探索场景。当智能体看到第一个物体时, 即轨迹图中蓝色的智能转向看到桌子上面的一个勺子, 智能体获得了 0.5 的奖励。由于该房间内有两个勺子, 所以并没有结束导航, 而是继续向前走去探索环境。这个时候演员网络决策出新的预期目标, 智能体就各自想各自的新目标导航。当红色的智能体导航到预期目标的过程中又看到

了第二个勺子，智能体团队又获得了 0.5 的奖励，因此立刻结束导航，发出结束的指令，智能体团队中所有的智能体都结束导航。在探索的过程中绿色的智能体一直没有找到任何目标，但是其依然在探索环境。

4　结束语

为解决单智能体在利用视觉导航解决目标探索问题中出现准确率低下，容错率较低，资源利用不合理等问题，本文设计了一种基于场景先验的多智能体强化学习的框架，用于视觉导航进行目标搜索的上述问题，并在 AI2THOR 中测试了该方法。经上述仿真实验结果可得：本文的集中式训练分布式探索的方法适用于解决在室内环境中基于视觉导航的目标搜索问题。智能体利用场景先验获取目标之间的空间关系相较于无场景先验而言无论在准确率还是在导航效率上都有一定的提升。

参考文献：

[1] ANDERSON P, WU Qi, TENEY D, et al. Vision-and-language navigation: interpreting visually-grounded navigation instructions in real environments[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 3674−3683.

[2] DAS A, DATTA S, GKIOXARI G, et al. Embodied question answering[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 1−10.

[3] THOMASON J, MURRAY M, CAKMAK M, et al. Vision-and-dialog navigation[C]//Proceedings of the Conference on Robot Learning. Cambridge MA: JMLR, 2020, 100: 394−406.

[4] STURM J, ENGELHARD N, ENDRES F, et al. A benchmark for the evaluation of RGB-D SLAM systems[C]// 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. Algarve: IEEE, 2012: 573−580.

[5] MIROWSKI P, PASCANU R, VIOLA F, et al. Learning to navigate in complex environments[EB/OL]. (2016−11−11)[2021−11−17].https://arxiv.org/abs/1611.03673.

[6] BABAEIZADEH M, FROSIO I, TYREE S, et al. GA3C: GPU-based A3C for deep reinforcement learning[C]//30th Conference on Neural Information Processing Systems. Barcelona, 2016: 1−6.

[7] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[EB/OL]. (2017−07−20)[2021−11−17].https://arxiv.org/abs/1707.06347.

[8] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning[EB/OL]. (2015−09−09) [2021−11−17].https://arxiv.org/abs/1509.02971.

[9] ANSCHEL O, BARAM N, SHIMKIN N. Averaged-DQN: variance reduction and stabilization for deep reinforcement learning[C]//International Conference on Machine Learning. Cambridge MA: JMLR, 2017: 176−185.

[10] WU Yi, WU Yuxin, TAMAR A, et al. Bayesian relational memory for semantic visual navigation[C]//2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 2769−2779.

[11] WORTSMAN M, EHSANI K, RASTEGARI M, et al. Learning to learn how to learn: self-adaptive visual navigation using meta-learning[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 6743−6752.

[12] 黄晓辉, 杨凯铭, 凌嘉壕. 基于共享注意力的多智能体强化学习订单派送 [J/OL]. 计算机应用, 2022: 1−7. (2022−07−26).https://kns.cnki.net/kcms/detail/51.1307.TP.20220726.1030.002.html.
HUANG Xiaohui, YANG Kaiming, LING Jiahao. Order dispatch by multi-agent reinforcement learning based on shared attention[J/OL]. Journal of computer applications, 2022: 1−7. (2022−07−26).https://kns.cnki.net/kcms/detail/51.1307.TP.20220726.1030.002.html.

[13] DU Heming, YU Xin, ZHENG Liang. Learning object relation graph and tentative policy for visual navigation[M]//Computer Vision – ECCV 2020. Cham: Springer International Publishing, 2020: 19−34.

[14] CHEN Boyuan, SONG Shuran, LIPSON H, et al. Visual hide and seek[EB/OL]. (2019−10−15) [2021−11−17].https://arxiv.org/abs/1910.07882.

[15] JADERBERG M, CZARNECKI W M, DUNNING I, et al. Human-level performance in 3D multiplayer games with population-based reinforcement learning[J]. Science, 2019, 364(6443): 859–865.

[16] 张文旭, 马磊, 贺荟霖, 等. 强化学习的地-空异构多智能体协作覆盖研究 [J]. 智能系统学报, 2018, 13(2): 202–207.
ZHANG Wenxu, MA Lei, HE Huilin, et al. Air-ground heterogeneous coordination for multi-agent coverage based on reinforced learning[J]. CAAI transactions on intelligent systems, 2018, 13(2): 202–207.

[17] 连传强, 徐昕, 吴军, 等. 面向资源分配问题的 Q-CF 多智能体强化学习 [J]. 智能系统学报, 2011, 6(2): 95–100.
LIAN Chuanqiang, XU Xin, WU Jun, et al. Q-CF multi-Agent reinforcement learning for resource allocation problems[J]. CAAI transactions on intelligent systems, 2011, 6(2): 95–100.

[18] 韩兆荣,钱宇华,刘郭庆.自注意力与强化学习耦合的多智能体通信 [J/OL].小型微型计算机系统:1−8.

(2022−05−13) [2022−07−31].DOI:10.20009/j.cnki.21-1106/TP.2021-0802.
HAN Zhaorong, QIAN Yuhua, LIU Guoqing. Multi-agent communication coupled with self-attention and reinforcement learning[J/OL]. Journal of Chinese Mini-Micro Computer Systems. 1−8. (2022−05−13) [2022−07−31]. DOI:10.20009/j.cnki.21-1106/TP.2021-0802.

[19] 方维维, 王云鹏, 张昊, 等. 基于多智能体深度强化学习的车联网通信资源分配优化 [J]. 北京交通大学学报, 2022, 46(2): 64–72.
FANG Weiwei, WANG Yunpeng, ZHANG Hao, et al. Optimized communication resource allocation in vehicular networks based on multi-agent deep reinforcement learning[J]. Journal of Beijing Jiaotong university, 2022, 46(2): 64–72.

[20] KIM D, MOON S, HOSTALLERO D, et al. Learning to schedule communication in multi-agent reinforcement learning[EB/OL]. (2019−02−05) [2022−07−31].https://arxiv.org/abs/1902.01554.

[21] DAS A, GERVET T, ROMOFF J, et al. Tarmac: Targeted multi-agent communication[C]//International Conference on Machine Learning. Cambridge MA: JMLR, 2019: 1538−1546.

[22] DING ZILUO, HUANG TIEJUN, LU ZONGQING. Learning individually inferred communication for multi-agent cooperation[EB/OL]. (2020−06−11) [2022−07−31].https://arxiv.org/abs/2006.06455.

[23] FOERSTER J, FARQUHAR G, AFOURAS T, et al. Counterfactual multi-agent policy gradients[C]// Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. New Orleans: AAAI Press, 2018, 32(1): 2974−2982.

[24] 陈新元, 谢晟祎, 陈庆强, 等. 结合卷积特征提取和路径语义的知识推理 [J]. 智能系统学报, 2021, 16(4): 729–738.
CHEN Xinyuan, XIE Shengyi, CHEN Qingqiang, et al. Knowledge-based inference on convolutional feature extraction and path semantics[J]. CAAI transactions on intelligent systems, 2021, 16(4): 729–738.

[25] YANG WEI, WANG XIAOLONG, FARHADI A, et al. Visual semantic navigation using scene priors[EB/OL]. (2018−10−15) [2022−07−31].https://arxiv.org/abs/1810.06543.

[26] 闫超, 相晓嘉, 徐昕, 等. 多智能体深度强化学习及其可扩展性与可迁移性研究综述 [J/OL]. 控制与决策, 2022: 1−20. (2022−06−14).https://kns.cnki.net/kcms/detail/21.1124.TP.20220613.1041.023.html.
YAN Chao, XIANG Xiaojia, XU Xin, et al. A survey on the scalability and transferability of multi-agent deep reinforcement learning[J/OL]. Control and decision, 2022: 1−20. (2022−06−14).https://kns.cnki.net/kcms/detail/21.1124.TP.20220613.1041.023.html.

[27] QIU YIDING, PAL A, CHRISTENSEN H I. Learning hierarchical relationships for object-goal navigation [EB/OL]. (2020−03−15) [2022−07−31].https://arxiv.org/abs/2003.06749.

[28] CHAPLOT D S, GANDHI D P, GUPTA A, et al. Object goal navigation using goal-oriented semantic exploration[J]. Advances in Neural Information Processing Systems, 2020: 33.

[29] HE Kaiming, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[C]//2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 2980−2988.

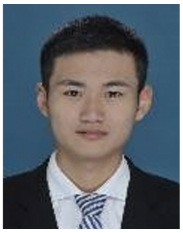
[30] KOLVE E, MOTTAGHI R, HAN W, et al. AI2-THOR: an interactive 3D environment for visual AI[EB/OL]. (2017−12−14) [2021−11−17].https://arxiv.org/abs/1712.05474.

[31] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[EB/OL]. (2016−09−09) [2021−11−17].https://arxiv.org/abs/1609.02907.

[32] GORDON D, KEMBHAVI A, RASTEGARI M, et al. IQA: visual question answering in interactive environments[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 4089−4098.

[33] YU CHAO, VELU A, VINITSKY E, et al. The surprising effectiveness of PPO in cooperative, multi-agent games[EB/OL]. (2021−03−02) [2021−11−17].https://arxiv.org/abs/2103.01955.

作者简介:



陆升阳, 硕士研究生, 主要研究方向为多智能体系统, 多智能体强化学习。

赵怀林, 教授, 博士, 主要研究方向为机器人学、多智能体系统和人工智能。

刘华平, 副教授, 博士生导师, 博士, 中国人工智能学会理事、中国人工智能学会认知系统与信息处理专业委员会秘书长, 主要研究方向为机器人感知、学习与控制、多模态信息融合。获吴文俊人工智能科技进步奖二等奖, 主持国家自然科学基金重点项目 2 项。