一种用于自主学习的虚拟仿真环境

钟方威，王亦洲
（北京大学数字媒体所，北京 100871）

摘要：赋予智能体通过与环境交互自主学习的能力是实现下一代人工智能的关键。本文，我们介绍了一种基于虚幻 4 的虚拟仿真环境，用于训练和测试自主智能体。该环境具有高逼真、可交互、灵活通用的特点，使得智能体能够在其中自由探索，自主学习场景感知、常识推理、决策控制等多项能力。为了验证该环境的可用性，我们用实验演示了如何在虚拟环境中构建自主智能，即利用强化学习方法训练端到端的神经网络实现基于视觉感知的目标搜索和目标追踪任务。

关键词：自主学习；虚拟环境；
中图分类号：O422　　**文献标识码：A**

Virtual Environments for Autonomous Learning

ZHONG FANGWEI, WANG YIZHOU

(Peking University, Beijing,100871,China)

Abstract: The key to realizing the next generation of artificial intelligence is how to make agents to autonomous learning from the interaction with the environment. In this paper, we introduce a collection of UE4 (Unreal Engine 4)-based virtual environments for training and testing the autonomous agents. The environment is high-fidelity, interactable, and flexible to meet the requirements of enabling agents to freely explore and autonomously learn the ability of scene perception, common sense, and decision making. To validate the usability of the environments, we developed agents via reinforcement learning to perform object searching and active object tracking.

Key words： autonomous learning; virtual environment;

1 引言[1]

自主学习（Autonomous Learning）是机器人最终实现通用人工智能所需具备的关键能力之一[1]。具体地，自主学习智能体（机器人）需要通过主动地在环境中探索以获取经验，进而从中总结常识学习技能。近年来，随着深度学习的发展，以强化学习[2]为代表的自主学习方法在诸多序列决策问题上取得了令人瞩目的成就，如下围棋[3]、玩电子游戏[4]等。实现自主学习的关键要素之一是智能体与环境的主动交互。通过主动交互获取环境的反馈后，智能体可以通过自监督或试错迭代的方式进行学习，而不再依赖大规模人工标注的数据集[5]。

图 1 基于虚幻 4 的仿真场景示例

基金项目：中国*****项目 3132016XNG1625
作者简介：钟方威（1993-），男（畲族），浙江温州人，北京大学博士研究生，zfw@pku.edu.cn。

然而，构建实物系统在真实场景中进行自主学习是危险且代价高昂的[6]。例如，机器人在真实环境中学习视觉导航需要面对物理碰撞、机械损耗、硬件故障等诸多风险，并且收集过程相当漫长。因此，我们认为有必要构建高逼真的三维仿真环境用于训练自主机器人和相关算法验证。

本文我们将介绍一种基于虚幻4（Unreal Engine 4，UE4）的高逼真虚拟环境用于自主学习。虚幻4是一种开源的游戏引擎，集成了高精度的物理引擎和高质量的渲染引擎，图1为仿真场景示例。为兼顾通用性和灵活性，我们将环境设计成三层结构，如图2所示。底层由基于虚幻4的仿真场景构成，包含了丰富多样的场景实例，如不同的室内室外场景。为实现外部程序与仿真场景的通信，我们基于UnrealCV[8]提供了通用的通信交互接口。最后，根据具体任务要求定义智能体-环境交互接口，并配置相关环境要素（奖赏函数、动作状态空间等）。为保证环境的通用性，我们交互接口设计规范，与OpenAI Gym[9]环境完全兼容。

我们通过实验演示了如何利用我们的环境实现智能体的自主学习。具体的，我们设计了两个基于视觉的机器人任务：物体搜索和目标追踪。结果表明，只要调用通用的强化学习算法，如A3C，智能体就能在我们的虚拟环境中自主地优化端到端（图像→动作）的策略网络，最终完成指定任务。

2 预备知识

2.1 强化学习

强化学习[2]是一种常用的自主学习方法，可用于解决时间序列决策问题。典型的强化学习基础概念包括了智能体(Agent)、环境(Environment)、状态(State)、动作(Action)和奖赏(Reward)五大要素。在一般的强化学习框架下，智能体需要从与环境的交互中学习一个策略函数(Policy)，使得每一时刻采取的动作从环境获得的累计奖赏最大。策略函数实现状态到动作的映射。在学习的过程中，智能体需要通过与环境进行交互试错，利用环境反馈的奖赏，更新策略函数。

2.2 三维仿真

三维仿真是指利用计算机技术生成的一个逼真的，具有多种感知功能的虚拟环境[7]。随着人们对三维仿真的真实感和精确度的更高要求，仿真软件需要提供物理模拟，碰撞检测，光照渲染等系统实现高逼真的仿真模拟。现阶段，支持实时高逼真渲染的三维仿真软件主要有 Unity 和 Unreral Engine 4（以下简称 UE4）两款工具。Unity 是一款由 Unity Technologies 所研发的跨平台 2D / 3D 游戏引擎。Unity 提供一整套软件解决方案，可用于创作、运营和变现实时互动的 2D 和 3D 内容。但是，因为其只提供了开发工具，不公开程序源代码，使得大大限制了其可扩展性。不同的是，由 Epic Games 开发的 UE4 完全开源，且功能强大。值得一提的是，UE4 提供的高级动态光照处理功能，使其能够在保证实时渲染的同时提供高逼真视觉效果。

相比于真实环境，三维仿真环境具有以下五个特点：a）可以便捷获取精准的环境真值（物体三维坐标到像素级图像分割），利用真值可以设计奖赏函数用于强化学习；b）支持智能体与环境的交互试错，且不用担心任何物理损耗；c）环境要素都是数值可控的，可以根据需求人为设计改动环境；d）可复制，方便研究社区间共享使用，可以为研究者提供具有可重复性的科学实验平台；e）可并行，同时运行多个环境供智能体探索，显著提高训练效率。

3 构建环境

我们在环境构建过程中对其进行分层设计，在充分利用高质量三维仿真环境素材的同时，保证了上层应用的通用性和灵活性。具体的，我们分三步构建我们的虚拟环境：1）收集制作基于 UE4 的仿真环境，2）基于 UnrealCV[8]设计通用的通信控制接口，实现外部程序对仿真环境的连接，3）基于 OpenAI Gym[9]定义任务，设计标准交互接口，实现智能体与环境的交互。图 2 为本虚拟环境的设计框架。

2.1 基于 UE4 的仿真场景

为了尽可能逼真地模拟真实场景，我们采用 UE4（Unreal Engine 4）构建高逼真的三维场景。UE4 是一款开源的可实现高逼真三维仿真的游戏引擎，集成了高精度物理引擎和高质量图形渲染引擎。此外，专业的设计师在虚幻商场中提供了丰富的环境资源素材供广大开发者共享使用，这可以节省我们构建仿真场景的工作量。因为不同环境素材针对的应用场景不同，使得不同项目之间差异较大。因此，我们需要对其进行“二次加工”，使之满足自主学

习的需求。具体的改进包括了：a）加入可交互的智能体（人），b）赋予物体一定的物理属性，c）加入 UnrealCV 插件实现与外部程序连接通信。

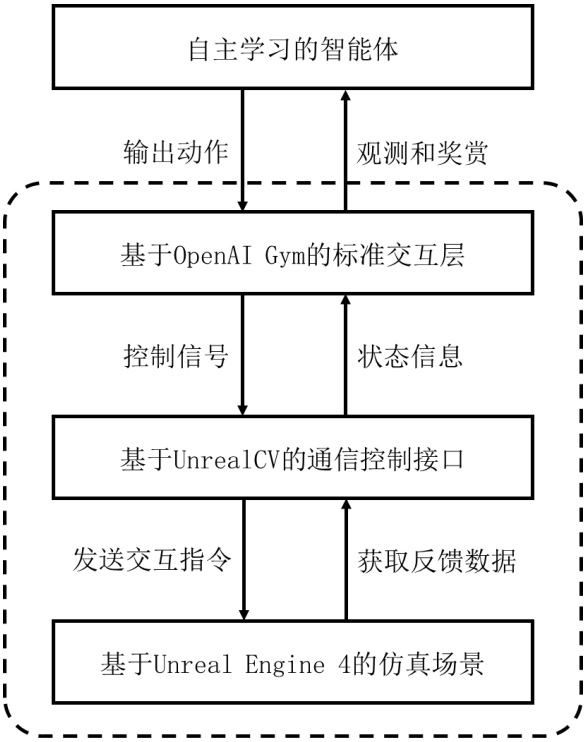

图 2 可交互的虚拟环境设计框架

2.2 基于 UnrealCV 的通信控制接口

我们使用 UnrealCV 实现 UE4 场景与外部程序（Python）的信息交互。具体的，外部程序可通过 UnrealCV 与 UE4 建立连接，发送指定命令，获取相应场景信息或控制相应物体，如获取特定视角的图像，将某个物体移动到特定位置等。由于 UnrealCV 是通过发送字符串方式传递指令，实现功能比较基础，往往需要对数据进行繁琐的预处理和后处理，因此我们根据常见机器人任务需求定义了更加简单通用的交互函数，以下我们列举了几个常用的交互函数：

1. moveto(i, d, theta) % 将 i 相机沿其前方 theta 度方向移动 d cm，返回 True/False 代表移动过程中是否发生碰撞。
2. get_observation(i, type) % 获取 i 相机当前时刻的观测图像，type 指定图像类型，包括了彩色图、深度图、语义分割图等。
3. get_pose(obj) % 获取物体 obj 的位姿(x, y, z, roll, yaw, pitch)。
4. set_pose(obj, (x, y, z, roll, yaw, pitch)) % 将物体 obj 以(roll, yaw, pitch)的朝向放置于坐标(x, y, z)处。

2.3 基于 OpenAI Gym 的标准交互接口

在通用交互函数的基础上，我们按照 OpenAI Gym 的标准接口范式，设计智能体与环境的交互接口。具体的，主要包括了设计接口函数（如 reset() 和 step()）和定义任务相关的参数（环境名称、观测空间、动作空间、奖赏函数等）。这样的设计，可以保证环境的灵活扩展性，当用户需要设计新的自主任务或者加入新的仿真场景时，只需要对相应的函数和参数进行简单修改即可实现。研究者只需要与使用标准 Gym 一样调用接口，就可以实现自主智能体与高逼真虚拟环境的交互，进而能够专注于算法研究。具体使用例子及说明如下：

```
env = gym.make("UnrealTrack-UrbanCity-DiscreteC
olor-v0") # 创建一个环境实例，环境的命名规则为 {任务}-{场景}-{动作空间}{观测空间}-{版本}
agent = Agent(env.action_space, env.observation
_space) # 根据动作和观测空间实例化一个智能体
for _ in range(1000):
  observation = env.reset() # 重置环境，返回最新的观测
  while True:
      action = agnet.policy(observation) # 智能体根据观测，输出要执行的动作
      observation, reward, done, info = env.step
(action) # 在环境中执行动作，并返回更新后的观测，奖赏，中止状态，以及其他额外信息。
  if done: # 如果真，中止当前 episode，跳出循环，重置环境
    break
env.close() # 关闭环境
```

3 实验验证

本节，我们将进一步通过实验演示如何利用本虚拟环境训练训练自主机器人完成任务。

3.1 任务定义

（1）物体搜索：智能体需要在场景中快速搜寻到指定物体，并移动到一个合适的视角进行拍照。只

有当目标物体在拍照获得画面中的比例达到一定大小时，才计做一次成功的搜索。在搜索过程中，如发生碰撞、移动超过 200 步、误触发“拍照”按钮，都将计做一次失败的搜索。因此，要完成这个任务，智能体既要理解场景的语义信息和三维结构，还要做到合理的路径规划，以实现最短路径搜索和躲避障碍物。在每次重置环境时，相机将被随机放置环境中任意区域。

（2）目标追踪：智能体需要根据视觉观测主动控制相机移动，以跟随目标物体并保证其以适当大小出现在画面中心。只有当相机持续跟踪目标超过 500 步时，计做一次成功的跟踪。在跟踪过程中，如果发生碰撞或者目标从画面中消失，都将计做一次失败的跟踪。要实现精准的相机控制，需要对目标进行识别和定位，并能对其运动轨迹进行合理预测。每次重置环境时，相机将被放置于环境中的任意位置，目标对应的将放置于在相机正前方 3 米处。

3.2 实验设置

我们采用 A3C[10]算法，一种通用的强化学习算法用于智能体的训练。如图 x 所示，智能体的神经网络先后采用了卷积神经网络（Convolutional Neural Network, CNN[11]）对图像信息进行编码，门控循环神经网络（Gated Recurrent Unit，GRU[12]）对时序观测进行编码，最后通过演员-批评家网络（Actor-Critic Network）分别输出动作和价值函数实现强化学习。智能体的输入为 84x84x3 的图像，输出为离散的动作选择。训练过程中，启动了 4 个并行环境用于智能体交互试错。智能体的学习率为 0.0001。

图 3 物体搜索实验所要寻找的四种物体，从左到右依次用红色包围框标出了电视机、沙发、床、冰箱。

图 4 目标追踪实验所用到的三个场景，从左到右分别为公寓、大客厅、城市街道，画面中心的人物为所要跟踪的目标。

3.3 实验结果

（1）物体搜索：

我们选择了一个环境（公寓），尝试让机器人在环境中通过自主学习，实现寻找指定目标。我们分别指定了电视机、沙发、床、冰箱这四种常见物体。从学习曲线中可以看到，智能体在与环境交互 10 至 20 万次后，其表现基本趋于收敛。另外，我们注意到，智能体学习搜索床和冰箱的速度较快，主要是因为物体外形较大，视觉特征明显，容易在初期随机探索时发现。但是后期上升缓慢，学习效率较低。因此，如何提升智能体与环境交互学习的效率是未来一个十分重要的研究方向。从表 2 的测试结果上看，智能体搜索四种物体的成功率均达到了 90%以上。寻找床的成功率最高（达到了 97%），并且对应的平均搜索步长最短（25.13 步）。

表 1 智能体在房间中搜索不同目标物体的测试性能。

目标物体	成功率	平均奖赏	平均步数
电视机	95%	7.7	30.0
沙发	94%	4.6	32.2
床	97%	6.4	25.1
冰箱	90%	7.9	32.3

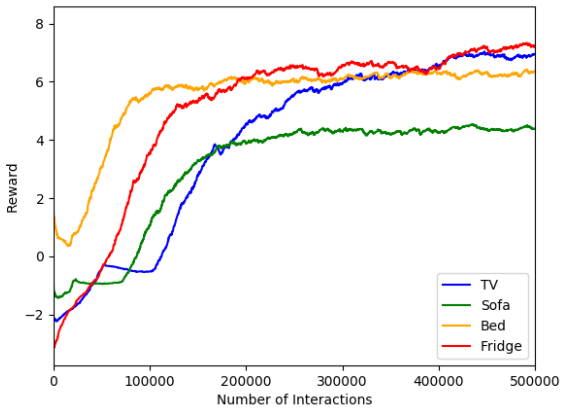

图 5 智能体在房间中学习搜索不同物体的学习曲线。

（2）目标追踪：

我们选择了三种不同的场景（公寓、大客厅、城市街道），尝试让机器人在环境中通过自主学习，实现主动跟踪移动的目标（人）。从学习曲线上可以看到，智能体在不同环境的收敛速度有一定差异，但最终都能达到一个比较稳定的跟踪效果。相比于目标搜索任务，实现目标追踪所需的时间总体偏长，只要是因为目标的运动轨迹多样复杂带来的新的挑战。表 2 的测试结果表明，利用强化学习，智能体能够在不同环境中自主学习主动跟踪目标，且达到了 94%以上的成功率。对应的平均奖赏和平均跟踪步长也达到了较高的水平。

表 2 智能体在不同场景中跟踪目标的测试性能。

场景	成功率	平均奖赏	平均步数
公寓	98%	426.1	493.7

大客厅	95%	387.0	485.4
城市街区	94%	376.5	485.1

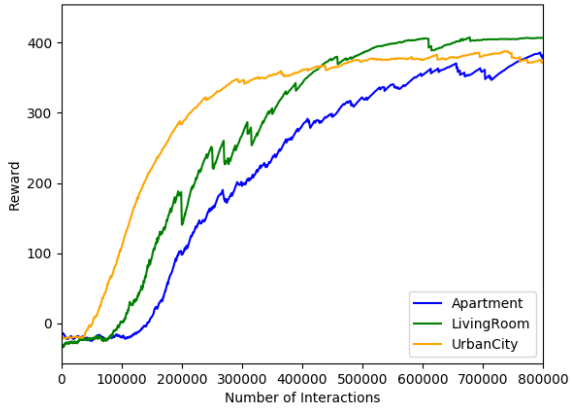

图 6 智能体在不同环境中学习主动目标追踪的学习曲线。

4 总结与展望

本文介绍了一种高逼真且可交互的虚拟环境用于自主机器人的训练和测试。本文详细介绍了该环境的设计思路和使用方法，并通过实验演示了如何使用本环境训练基于强化学习算法的智能体实现物体搜索和目标追踪。

未来，开发者将进一步丰富场景内容，扩展自主任务，使之成为一种通用的智能测试平台，促进自主学习算法学习效率提升和应用场景扩展。另外，探索如何实现智能体从虚拟世界到真实应用场景的快速迁移也是一个有意义的研究方向。

参考文献

[1] 郑南宁. 人工智能新时代[J]. 智能科学与技术学报, 2019, 1(1): 1-3.

[2] Sutton R S, Barto A G. Reinforcement learning: An introduction[M]. MIT press, 2018.

[3] Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of go without human knowledge[J]. Nature, 2017, 550(7676): 354-359.

[4] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.

[5] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]//2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009: 248-255.

[6] Dulac-Arnold G, Mankowitz D, Hester T. Challenges of real-world reinforcement learning[J]. arXiv preprint arXiv:1904.12901, 2019.

[7] 夏蕾. 三维技术的分类及应用[J]. 硅谷, 2014, 7(15): 121-122.

[8] Qiu W, Zhong F, Zhang Y, et al. Unrealcv: Virtual worlds for computer vision[C], Proceedings of the 25th ACM international conference on Multimedia. 2017: 1221-1224.

[9] Brockman G, Cheung V, Pettersson L, et al. Openai gym[J]. arXiv preprint arXiv:1606.01540, 2016.

[10] Mnih V, Badia A P, Mirza M, et al. Asynchronous methods for deep reinforcement learning[C]. International conference on machine learning. 2016: 1928-1937.

[11] LeCun Y, Bengio Y. Convolutional networks for images, speech, and time series[J]. The handbook of brain theory and neural networks, 1995, 3361(10): 1995.

[12] Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv preprint arXiv:1406.1078, 2014.

（篇幅所限，略去后面）