解决 ICRA RoboMaster AI 挑战与深度学习的研究

陈明阳 刘 博 茆意风
(美国宾夕法尼亚大学，美国 宾夕法尼亚州 19019)

摘 要 本项目提出了机器人战斗强化学习的模型。通过引入 PyGame 的虚拟环境,在预先定义的环境中开展训练模型,核心模型是神经网络在深度 Q 学习中应用模拟决策过程的功能。除了 DQN 之外,还在训练过程中应用了角色评判方法。通过比较两个模型输出的差异,进行了深入讨论研究和改进。

关键词:强化学习 ;DQN 深度学习 改进
中图分类号:TP18 文献标识码:A 文章编号:2096-4390(2020)29-0104-02

1 概述

本文采用一个通用的强化学习算法，并通过自我发挥和学习，不断优化算法，研究在 AlphaGo 中应用的自我游戏策略和 AlphaZero 的变化。由于 AlphaZero 不会增加训练数据,也不会在 MCTS 期间变换板的位置。因此,使用蒙特卡罗树搜索代替 beta 搜索,采用通过改变其他对称方面来训练非对称情况下的策略,研究这种方法,找到一种通用的自我游戏强化学习方法。

本文主要是将深度强化学习应用于街机学习环境中训练 7 款 Atari 游戏,该方法采用 Q 函数的神经网络训练模型,模型的输入为像素，输出为评估未来回报的价值函数。本文的关键点是 Actor- Critic 算法,它是提出并分析一类基于随机平稳策略的马尔可夫决策过程优化的算法，也是两个时间尺度的算法,其中,Critic 使用具有线性近似结构的时域学习，并且基于 Critic 提供的信息，在近似梯度方向上更新 Actor。通过研究表明,Critic 的特征应该跨越由 Actor 的选择所规定的子空间,提出收敛性和有待解决的问题。

2 虚拟机器人环境——PyGame

2.1 设置虚拟机器人环境——PyGame

PyGame 是一个基于 python 的虚拟格斗游戏环境,在此过程中接收来自键盘和鼠标的输入,一组应用编程接口和预定义的类降低了虚拟环境创建的难度。ICRA 挑战赛的真实环境为 8 米 *5 米的场地,两个机器人的出场地位于左上角和右下角,补充场地位于黄色十字区域。机器人在补给区被修复，当它们站在补给区时，它们的生命值会持续上升,ICRA 的真实现场环境如图 1 所示。

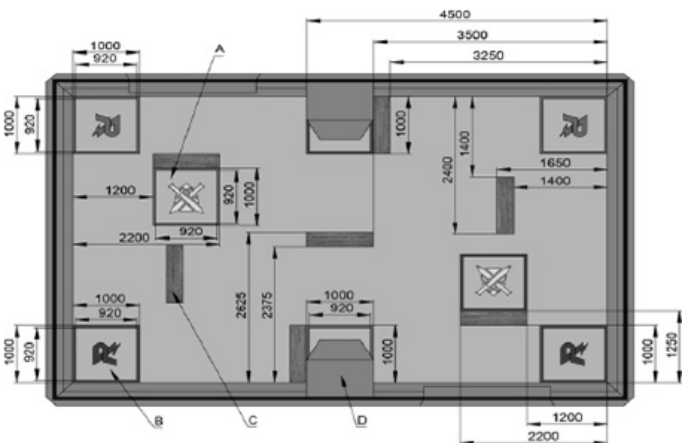

图 1 ICRA 的真实现场环境

在图 1 中,机器人无法通过的障碍物被显示为深灰色矩形。通过研究决定，采用重新创建 ICRA AI 挑战赛的新战斗环境，如图 2 所示,其中,障碍物和补给区域与原来的位置相同。为了增加决策的复杂性,增设弹药重装区,图中的弹药重装区域显示为绿色区域,机器人可以在此区域进行重新装弹,以避免子弹耗尽。

图 2 PyGame 中实现的 ICRA 新环境

奖励规则设置如下：在所有迭代开始时，奖励被初始化为零。如果敌人被击中,那么射手的奖励将增加 10 点,而敌人将减少 20 点。如果敌人被摧毁,奖励会激增到 100 点,如果玩家被摧毁,奖励本身会下降到 200 点。弹药和生命点不会影响奖励,而生存时间将以对数形式加入奖励。

2.2 优化深度学习的算法

实现的深度 Q 学习是基于 Pytorch 的卷积神经网络。网络的输入是模型训练过程中 PyGame 环境的一个截图，输出是给定输入环境下的一个预测动作。该动作包括四个方向的移动(上、下、左、打),两个枪操作方向 (顺时针、逆时针)和射击。该神经网络为三层卷积神经网络，具有不同大小的核和漏项。激活层被分配给非线性 ReLU 层,我们选择时间差异误差作为损失函数,两者具有相同的损失函数。把具有相同的最佳点作为传统的 Q- learning 函数。下面列出了这个损失函数的梯度下降：

$$\varphi_{k+1} = \varphi_k - \eta \nabla_\varphi \left(q_{\varphi_k}(x,u) - \text{target}\left(x'; \varphi_k\right)\right)$$

因此,Q- learning 的目标函数为:

$$Q(s_t,a) \leftarrow Q(s_t,a) + \alpha_{t+1} + \gamma \max_p Q(s_{t+1},p) - Q(s_t,a)$$

本模型的目标是利用神经网络的非线性特性来模拟这个函数，该模型产生 Q- learning 函数的估计，而 TD 误差在 Q- learning 中找到相同的最优值。

作者简介:陈明阳(1996–),男,北京人,研究生,美国宾夕法尼亚大学,主要从事电子信息工程专业的研究。

2.3 Actor- critic 模型设计

实现角色评论设模型设定在 pytorch 中完成，模型的输入是抽象的状态元组，状态元组代表在某一时刻坦克的状态，包括：坦克中的位置、速度、运行状况和子弹数量等。Critic 模型采用价值函数进行估计，其中，选择 Q 值作为估计的目标值。而 Actor 模型是决策制定的，该模型按照 Critic 建议的方向更新政策分配，其中，Critic 函数和 Actor 函数都属于神经网络模拟。

2.4 多人战斗模型设计

上述模型设计是在单人游戏环境中实现的，通过计算机自动控制敌人，电脑玩家可以忽略障碍物的封锁，并且拥有无限数量的弹药。优化后的虚拟环境中可以实现 2 人战斗，两个玩家在后端由两个独立的模型控制，通过重新部署 AlphaGo 战略，试图找出让机器人从零开始学习规则的策略。

3 仿真结果

3.1 DQN 模型结果

通过采用 DQN 模型仿真结果对比，DQN 的训练效果明显优于 Actor- Critic 的训练效果，并且 convolutional neural network 在决策过程中更能有效的找到合适的动作，这是因为图像的复杂性使得模型更容易判断游戏情况，优化后的图像包含许多有用信息，如封锁区域和不同的供电区域位置。然而，有时这种模式会以错误的方式表现，比如向空中射击和在进入近距离战斗前浪费弹药，采用的 DQN 模型奖励功能如图 3 所示。

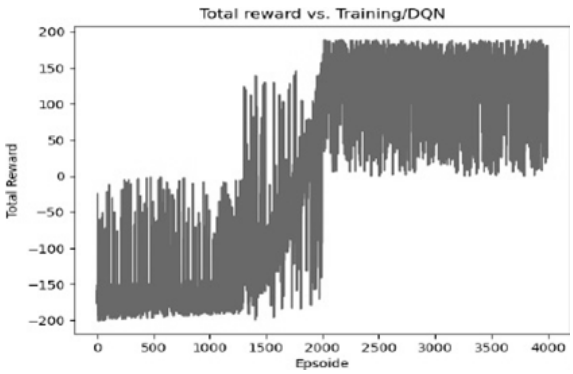

图 3 DQN 模型奖励功能图

3.2 Actor- Critic 结果

由于选择有限维数的状态元组，模型只能感知坦克的当前状态，而无法告诉模型上电区域和障碍物的位置。此外，这种方法的训练难度大于前一种方法，这意味着训练时间较短，可能会导致模型无法收敛到更大的期望回报。从图 4 中可以看出，该模型在提高奖励期望方面并不有效。

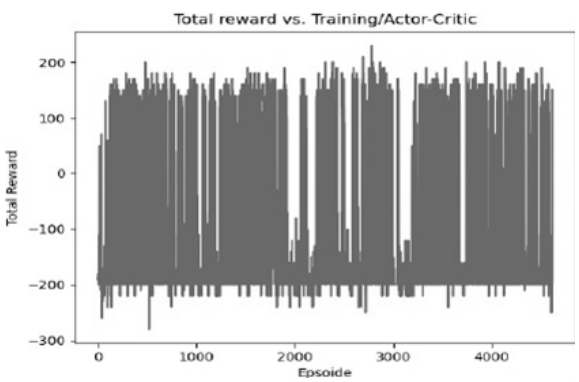

图 4 演员评论模型奖励功能图

3.3 多人战斗模型结果

从奖励情节中可以看到，有时玩家能够找到消灭敌人的策略，而有时两名玩家在空白区域徘徊。这是因为训练时间有限，这导致模型无法探索虚拟环境中的所有可能性，图 5 和图 6 列出了两个玩家的奖励结果。

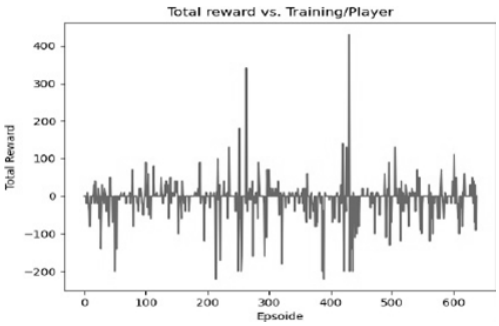

图 5 Player1 奖励功能情节

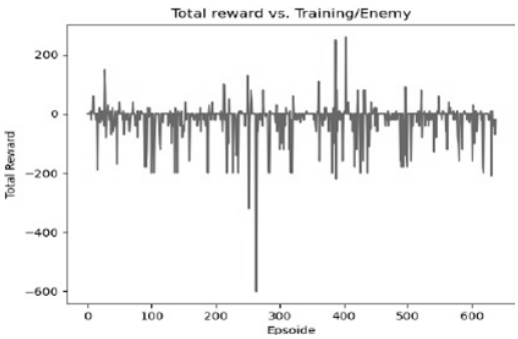

图 6 Player2 奖励功能情节

4 结论

在后续的工作中，需要更多的时间和更先进的设备来完善本模型，pygame 环境每次都需要截屏，这会浪费大量的计算资源，因此以后选择更加简练的环境，以此来提高效率。未来，可以通过调整该模型的神经网络和学习策略来实施进一步优化。

参考文献

[1] Silver, David 等.通用强化学习算法自我发挥?[J].科学,2018: 1140–1144.

[2] Sihem Ouahouah,Tarik Tale 等. 基于无人机的增值服务的高效卸载机制[J].《IEEE 国际商会 2017 特设和传感器网络研讨会》,2017:34–37

[3] S. Jeong, O. Simeone, A. Haimovich, and J. Kang.无人机安装 cloudlet 移动云计算：通信和计算最佳位分配》?[J]. IET Commun., 11: 969–974, 2017.05

[4] G. Scutari, F. Facchinei, L. Lampariello, and P. Song,《arallel and distributed methods for nonconvex optimization》?[P] part I: Theory, arXiv:1410.4754v2, Jan. 2016.

[5] 周炯槃,庞沁华,续大我等. 通信原理. 第 3 版[M].北京 :北京邮电大学出版社. 2008: 226.

[6] 闻欣研. MATLAB 从入门到精通[M].北京 :清华大学出版社. 2017. ISBN:9787302461128.

[7] 田宝玉, 杨洁, 贺志强, 许文俊等. 信息论基础[M].北京 :人民邮电出版社, 2016.

[8] 刘宝玲, 李立华, 张晓莹, 崔琪楣, 邓钢. 通信电子电路[M].北京 :高等教育出版社, 2007 .