基于深度强化学习的 Super Mario Bros 游戏智能训练

车景平，王　强，吉　凡
（周口师范学院 网络工程学院，河南 周口 466001）

摘　要：近年来，深度强化学习在复杂决策和控制任务中得到了广泛应用，并在游戏 AI 领域展现了卓越性能。基于双重深度 Q 网络的方法，提出一种通过智能体与 Super Mario Bros 环境的持续交互、逐步学习并优化游戏策略。首先，利用 gym－super－mario－bros 框架构建训练环境，并通过帧跳、灰度转换和图像缩放等技术提升训练效率。其次，智能体采用 DDQN 架构结合卷积神经网络进行特征提取，并通过经验回放和目标网络减少 Q 值波动。最后，通过衰减的 epsilon－greedy 策略平衡探索与利用。实验结果表明，该方法能有效提升智能体表现。

关键词：深度强化学习；DDQN；Super Mario Bros 游戏训练
中图分类号：TP391.4　　**文献标志码**：A　　**文章编号**：1671－9476(2024)05－0060－05
DOI：10.13450 / j.cnki.jzknu.2025.02.011

近年来，强化学习（Reinforcement Learning，RL）在游戏 AI 中展现了强大的学习和适应能力，尤其在处理高维状态空间和复杂决策问题上取得了显著进展。深度强化学习（Deep Reinforcement Learning，DRL）结合深度神经网络的特征提取能力，使智能体能够在高维环境中高效学习策略。Mnih 等人[1]提出的深度 Q 网络（Deep Q－Network，DQN）通过引入经验回放和目标网络，成功解决了传统 Q－learning 的不稳定性问题，并在 Atari 游戏中取得了突破性成果，推动了游戏 AI 的研究发展。随着强化学习算法不断创新，如 MuZero[2]结合蒙特卡洛树搜索（MCTS）与深度学习模型，使智能体能够在无需预先了解规则的情况下推理出最优策略。国内研究者[3－4]也在动作类游戏中采用对抗性学习和生成对抗网络（GANs），提升了智能体在复杂环境中的策略优化能力。这些研究证明了深度强化学习在开发智能游戏 AI 方面的重要性。

作为经典动作游戏，Super Mario Bros 以其复杂的动态环境和多样化动作需求，成为测试强化学习算法的理想平台。游戏设计要求智能体在有限时间和生命内完成多关卡挑战，克服敌人和动态障碍物，实时决策和策略优化成为关键挑战。面对这些难题，智能体需要在高维状态空间中快速适应并找到最优策略，如何高效训练智能体成为研究热点。

本文提出了一种基于双重深度 Q 网络（Double Deep Q－Network，DDQN）的训练方法，结合卷积神经网络进行特征提取，能够有效提升智能体的环境感知和策略学习能力。DDQN 通过将动作选择与 Q 值更新分离，解决了传统 Q－learning 的过度估计问题，显著提高了算法的稳定性。基于 gym－super－mario－bros 框架构建的训练环境，通过帧跳、灰度转换和图像缩放等预处理技术，进一步提升了训练效率。实验结果表明，该方法在复杂关卡中表现出色，验证了其在动作游戏 AI 中的

收稿日期：2024－06－20；**修回日期**：2024－08－10
基金项目：河南省高等学校重点科研项目“基于生成式 AI 模型的智能建筑设计研究”（25B520026）；2024 年河南省本科高校大学生创新创业训练计划项目“胸部 CT 实性肺病变体积的形态分割与部分体积的分析”（202410478011）；2024 年周口市科技计划项目普通科技攻关“基于扩散模型的文生图研究”（96）；2024 年周口市科技计划项目普通科技攻关“基于身份寻求自监督表示学习的智能寻人系统”（98）
作者简介：车景平（1991－），女，河南长葛人，助教，硕士，主要研究方向为弱目标检测和计算机视觉。

应用潜力，并为后续研究提供了新的参考方向。

1　Mario 游戏中的 AI 训练模型

游戏设计要求智能体在 3 条生命内通过 32 个关卡，需在有限时间内做出快速决策，克服敌人和移动平台等障碍。这种环境为 AI 模型的训练提供了丰富的挑战和验证场景。

在 AI 训练过程中，Super Mario Bros 通过设计奖励机制来驱动智能体策略的优化。奖励函数通常包含三部分：速度奖励(v)、时间惩罚(c)和死亡惩罚(d)。智能体通过向右移动获得速度奖励，停滞时会受到时间惩罚，死亡则会导致大幅扣分。该设计引导智能体尽可能快地完成关卡，并避免失误，从而促进了策略的有效形成。此外，游戏环境的灵活性极大地支持 AI 模型的训练与测试。即可以选择完整关卡挑战，也可针对单一关卡进行训练，以提高模型的学习效率和效果。随机选择关卡的功能提升了智能体的泛化能力，使其能够适应多样化的游戏环境。Super Mario Bros 还支持多种 ROM 模式和自定义关卡配置，提供调整和优化 AI 模型的灵活工具。

在实际应用中，通常结合深度强化学习模型(如 DQN、DDQN)与卷积神经网络，以处理游戏中的高维图像输入，并通过经验回放等技术提高模型的训练效果。通过这种结合方式，AI 智能体能够逐步学习并优化游戏策略，在复杂的动态环境中做出实时决策，展现出较高的游戏表现。

2　算法设计

2.1　环境构建

为了在 Super Mario Bros 游戏中实现深度强化学习的智能训练，本文采用 gym－super－mario－bros 库，该库是一个为 OpenAI Gym 环境设计的 Mario 游戏接口，它通过 NES－Py 模拟器实现，并支持多种不同的游戏模式和 ROM 格式。这个环境为智能体提供了一个复杂的、高维的动态环境，使得智能体可以通过与游戏的不断交互来学习最优策略。

(1)游戏环境初始化

首先使用 gym_super_mario_bros.make() 方法初始化 Mario 游戏环境，可随机选择关卡，具体的选择可以根据以下格式进行：

SuperMarioBros－＜world＞－＜stage＞－v＜version＞。其中，＜world＞ 表示游戏的世界(范围是 1 到 8)；＜stage＞ 表示该世界中的关卡(范围是 1 到 4)；＜version＞ 表示使用的 ROM 模式，0：标准 ROM，1：下采样 ROM，2：像素化 ROM，3：矩形化 ROM。

如图 1 所示，SuperMarioBros－1－1－v0 表示我们使用第 1 世界第 1 关的标准 ROM 模式。

图 1　第 1 世界第 1 关(SuperMarioBros－1－1－v0)标准 ROM 模式的游戏界面

(2)动作空间设定

为了简化强化学习任务并加快训练效率，通过 JoypadSpace 限制 Mario 游戏中的动作空间。原始游戏的动作空间较大，智能体可以执行多种组合动作。本文将动作空间限制为 4 种基本动作：向左移动、向左跳跃、向右移动、向右跳跃。

通过这种限制动作空间的方式，智能体可以专注于最基本且有效的动作策略，大大减少了动作探索的复杂性，进而加速了训练过程。这种配置确保智能体能够高效地学习如何通过关卡，避免无意义的复杂动作，同时保持了动作多样性以适应不同的关卡需求。此外，环境中仅包含游戏过程中产生的可奖励帧，不包括过场动画或加载画面，进一步提升了训练的效率。

(3)环境互动

在强化学习过程中，智能体与环境的每次互动都涉及以下几个步骤：环境根据智能体的动作返回新的状态，提供即时的奖励，并指示游戏是否结束。智能体通过不断与环境的交互，从而逐步优化其策略。在实际训练中，智能体会在每个 episode(完整的一局游戏)中重复上述交互过程，直到成功通过关卡或失败为止。

2.2　DDQN 算法架构

本文基于双重深度 Q 网络设计了一种强化学习智能体，用于在 Super Mario Bros 游戏中执行智能化训练任务。通过引入在线网络和目标网络解决了传统 Q－learning 中 Q 值过度估计的问题，有效提升了智能体在复杂环境中的学习效果。以下为该算法架构的详细设计。

DDQN 采用卷积神经网络进行特征提取，输入为经过预处理的游戏状态图像，输出为智能体在当前状态下的动作选择值(Q 值)。CNN 的结构如表 1 所示。

表 1　CNN 网络结构

层级	输入大小	输出大小	卷积核大小	步幅	激活函数
卷积层 1 (Conv2D 1)	4 通道	32 个特征图	8x8	4	ReLU
卷积层 2 (Conv2D 2)	32 通道	64 个特征图	4x4	2	ReLU
卷积层 3 (Conv2D 3)	64 通道	64 个特征图	3x3	1	ReLU
展平层 (Flatten)	64 个特征图 (3136 维度)	3136			
全连接层 1 (Dense 1)	3136 维度	512 单元			ReLU
全连接层 2 (Dense 2)	512 维度	动作空间维度			

如图 2 所示，该网络由在线网络和目标网络组成。在线网络负责动作选择和训练，目标网络用于稳定训练，防止过度估计。目标网络每隔固定步数从在线网络复制参数，其余时间参数保持不变。

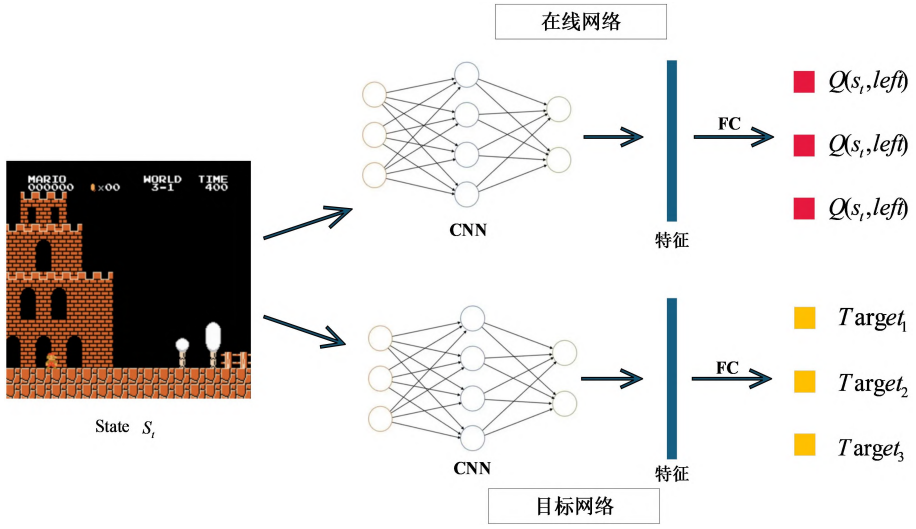

图 2　DDQN 网络结构图

2.3　智能体的核心功能模块

为了实现智能体的高效训练，设计了以下功能模块。

(1)动作选择(act)

智能体在每个时间步基于当前状态选择一个动作。采用探索与利用的结合策略(epsilon－greedy)，初始阶段智能体有较高的探索率(exploration rate)，即随机选择动作，以积累经验；随着训练的深入，探索率逐渐下降，智能体更多地利用已有经验执行最优动作[5]。

(2)经验回放(experience replay)

智能体通过与环境交互生成状态——动作——奖励——下一状态的数据对，并将其存储在经验池中。为了提高学习效率和稳定性，智能体在每次更新网络参数时，随机从记忆池中采样小批量数据进行训练。这种经验回放机制能够有效避免智能体陷入局部最优，同时提高数据的利用率[6]。

(3)Q 值更新

在训练过程中，智能体通过计算当前状态下的预测 Q 值(来自在线网络)和实际的目标 Q 值(来自目标网络)来更新在线网络的参数。目标 Q 值基于当前奖励以及下一状态的最大 Q 值计算得到，具体公式如下：

$$Q_{\text{target}}(s_t,a_t)=r_{t+1}+\gamma\max_{a_{t+1}}Q_{\text{target}}(s_{t+1},a_{t+1})$$

其中，S_t 表示当前时刻 t 的状态，a_t 是在状态 S_t 时智能体采取的动作，r_t 是执行动作 a_t 后的及时奖励，γ 为折扣因子，$\max_{a_{t+1}}Q_{\text{target}}(s_{t+1},a_{t+1})$ 为目标网络计算的下一状态的最大 Q 值。在线网络参数通过平滑 L1 损失进行更新[7]。

(4)网络同步

为了进一步提升训练的稳定性，DDQN 在每训练一定步数后，将在线网络的参数复制到目标网络中，这一同步过程能够有效减少策略更新的波动，避免过度拟合当前策略。

3　算法性能测试

3.1　训练流程

本文实验在配备 Intel Core i9 处理器、1TB 内存、NVIDIA GeForce GTX 4070 显卡的计算机上，使用 Python 3.8 和 PyTorch 2.0.1 进行编程和实验。实验采用 DDQN 算法的智能体进行游戏的训练，训练过程由智能体与环境的不断交互构成，旨在通过优化策略逐步提升智能体的游戏表现。整个训练过程如下。

1. 初始化：智能体初始化 Q 值、随机初始化网络权重，环境重置，并设置初始探索率为 1.0。

2. 动作选择：根据当前状态，智能体通过 epsilon－greedy 策略选择动作，即随机探索或利用最优策略。

3. 状态更新与奖励获取：执行动作后，环境返回下一状态、即时奖励和是否结束的信息。

4. 记忆存储：将状态、动作、奖励、下一状态和完成标志存储到经验池中。

5. 经验回放与网络更新：随机从经验池中抽取样本，计算 Q 值，更新在线网络的参数，并定期将其同步至目标网络。

6. 权重保存：每隔固定的 episode 保存网络权重，确保后续训练或评估的连续性。

7. 循环与结束：当游戏结束时，环境重置并开始新的训练回合，直到满足终止条件。

3.2　实验结果分析

实验结果以每 10 个 episode 的累计奖励为衡量指标，分析智能体在训练过程中的表现。图 3 和图 4 展示了智能体在不同训练阶段的累计奖励变化趋势。横轴表示每隔 10 个 episode 记录一次表现，单位为 10 个 episode。纵轴表示每 10 个 episode 的累计奖励，数值越高表明智能体表现越好。

如图 3 所示，智能体在前 500 个 episode 内的累计奖励变化。初始阶段智能体的累计奖励波动较大，奖励值在 300 到 800 之间不稳定波动。这种不稳定性与探索率较高、智能体进行大量随机探索有关。由于智能体在早期主要通过随机选择动作积累经验，因此奖励值呈现出明显的波动趋势。然而，随着训练的进行，智能体逐渐开始探索到有效的游戏策略，奖励值整体趋于上升，但在初期依然存在较大的波动。

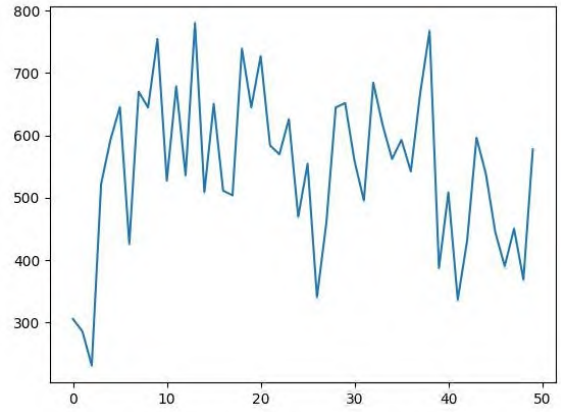

图 3　智能体训练前 500 个 episode 累计奖励波动图

如图 4 所示，智能体在 10 000 个 episode 内的累计奖励变化。随着训练的深入，智能体的奖励值整体呈上升趋势，这表明智能体逐渐学会了更有效的策略。在前 2 000 个 episode 中，智能体的表现波动较大，累计奖励值较低，主要由于探索阶段智能体仍然在尝试不同的策略。

从约 2 000 个 episode 开始，智能体的累计奖励呈现出较为平稳的上升趋势，说明此时智能体已经逐渐从探索阶段过渡到利用阶段，开始依赖学到的策略进行动作选择。之后，随着 episode 数的增加，累计奖励值继续增长，并逐步稳定在 2 000 至 2 500 之间，展现了智能体在游戏中的高效表现。

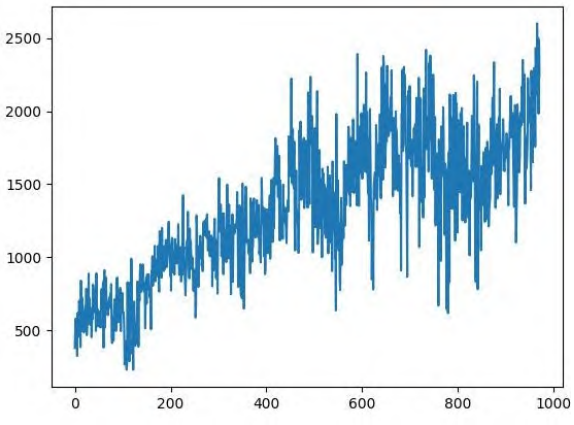

图 4　智能体训练前 10000 个 episode 累计奖励波动图

4　结语

本文设计一种基于 DDQN 算法的智能体训练方法，并成功应用于 Super Mario Bros 游戏的智能训练。通过设计合理的神经网络结构、经验回放机制以及探索与利用策略的结合，智能体在复杂动态的游戏环境中逐步优化其策略，实现了较高的游戏

表现。实验结果表明,智能体在早期阶段表现波动较大,随着训练的深入,智能体的表现逐渐稳定并获得了更高的奖励,充分验证了DDQN算法在复杂决策环境中的有效性。本文不仅展示了DDQN算法在动作游戏中的应用潜力,也为强化学习技术在游戏AI领域的进一步研究提供了参考。未来的研究可以在智能体的动作选择、策略优化以及模型的泛化能力方面进行更深入的探索,以应对更复杂的游戏场景和更高难度的任务。

参考文献:

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

[2]Schrittwieser J, Antonoglou I, Hubert T, et al. Mastering Atari, Go, chess and shogi by planning with a learned model[J]. Nature, 2021, 588(7839): 604-609.

[3]田佩, 臧兆祥, 张震, 等. RTS游戏中基于强化学习的行动参数配置优化[J]. 计算机仿真, 2023, 40(8): 355-359.

[4]项宇, 秦进, 袁琳琳. 一种基于条件生成对抗网络的强化学习数据增强方法[J]. 计算机与数字工程, 2024, 52(6): 1739-56.1745

[5]况立群, 冯利, 韩燮, 等. 基于双深度Q网络的智能决策系统研究[J]. 计算机技术与发展, 2022, 32(2): 137-142.

[6]罗国攀, 张国良, 李德胜, 等. 基于深度强化学习的移动机器人路径规划优化[J]. 组合机床与自动化加工技术, 2023, 45(4): 36-39.

[7]Van Hasselt H, Guez A, Silver D. Deep Reinforcement Learning with Double Q-learning[C]. Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, 2016: 2094-2100.

(责任编辑 闫建军)

Intelligent training in Super Mario Bros game based on deep reinforcement learning

CHE Jingping, WANG Qiang, JI Fan

(School of Network Engineering, Zhoukou Normal University, Zhoukou 466001, China)

Abstract: In recent years, deep reinforcement learning has been widely applied to complex decision-making and control tasks, demonstrating exceptional performance in the field of game AI. This paper proposes a method based on Double Deep Q-Networks , where the agent continuously interacts with the Super Mario Bros environment to progressively learn and optimize game strategies. First, the gym-super-mario-bros framework is used to build the training environment, and preprocessing techniques such as frame skipping, grayscale conversion, and image resizing are employed to improve training efficiency. Next, the agent adopts a DDQN architecture combined with a convolutional neural network for feature extraction, using experience replay and a target network to reduce Q-value fluctuations. Finally, an epsilon-greedy strategy with gradual decay is used to balance exploration and exploitation. Experimental results show that this method effectively improves the agent's performance, demonstrating the potential of deep reinforcement learning in real-time decision-making and providing a reference for further research in game AI.

Key words: deep reinforcement learning; DDQN; Super Mario Bros game training