**DQN实战——以东方非想天则格斗AI为例**

田永铭 钱禹昂 袁理 蒋李杰 陆迅

**[摘 要]** 当下深度强化学习非常火热，而DQN算法作为其中举足轻重的算法，具有非常高的研究价值。仔细检索，可以知道将DQN算法应用到很多小游戏上都有很好的成果，而将DQN算法应用到格斗AI 这样较为复杂的实时策略同步博弈情况下的研究成果非常少。因此，本文将以利用DQN算法实现东方非想天则格斗AI为例，探讨将DQN算法应用到格斗AI上的困难与挑战、解决方案和措施。这将为后来制作格斗AI的研究者提供良好的思路和启发。

**[关 键 词]** 深度强化学习；DQN算法；格斗AI

**目录**

**引言2**

研究背景2

研究意义2

**基础知识2**

游戏环境2

实验设置3

强化学习算法3

**实现过程4**

第一代——防御型4

第二代——局部最优型4

第三代——全面优秀型5

**综合与讨论6**

reward函数的设计6

全面优化7

反思与不足8

**总结9**

**第一部分 引言**

一、研究背景

强化学习是一种用于解决智能体与环境交互的决策问题的学习范式。早期的基于表格的强化学习算法，在处理简单的二维表格世界问题时表现良好，但在处理复杂问题时表现不佳。[1]因此，我们在学术平台上检索利用DQN玩游戏的文章，发现的大多是那种玩简单的小游戏的文章；而以关键词“DQN + 格斗AI”进行检索，所得结果甚少。这是因为，深度强化学习面临着普遍的问题，比如实时性、复杂信息决策、稀疏奖励、随机性和状态动作空间大等问题，而这些问题将在我们的研究中凸显出来。

二、研究意义

回合制的游戏AI，例如AlphaGo、麻将AI（Suphx）等等，已经取得了举世瞩目的成就。而实时性AI，例如王者荣耀AI（绝悟），也有能和职业玩家抗衡的水平，然而其实现依赖于大量数据和强大计算资源。[2]我们研究的东方非想天则AI，具备实时策略同步博弈的特点，同时我们仅靠个人电脑完成实验。在有限的条件下，我们聚焦于博弈方法的设计及策略目标问题的优化等等，实现了一个很强的格斗AI，该研究将为后续研究提供良好的思路和启发。

**第二部分 基础知识**

一、游戏环境

东方非想天则是一款出色的格斗游戏。其游戏背景设定在一个充满东方幻想元素的世界。画面和角色设计融合了传统的东方元素，包括古老的神话、武侠风格以及奇幻的设定，为玩家呈现了一场别具一格的东方风格的格斗。玩家需要操控特定的角色，与对手进行格斗，使用复杂的连招和绚丽的技能，战胜对方。游戏中的基本动作如下表所示（括号内为按键）：

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 进攻型： | 普通攻击（Z） | 远程轻攻击（X） | 远程重攻击（C） | / |
| 移动型： | 上（↑） | 下（↓） | 左（←） | 右（→） |
| 防御型： | 站防（向后） | 蹲防（同时向下向后） | / | / |
| 技能型： | 切换技能符卡（S） | 释放技能（D） | / | / |

玩家通过战斗，最终使得对方血量先减少至0的一方获胜。

二、实验设置

我们的实验平台是个人电脑，我们获取了东方非想天则游戏到本地，利用vscode环境跑实验。我们通过侵入游戏内存地址获取游戏参数，为了简化状态空间，我们忽略了对格斗双方影响相同的增益，如游戏中的设定“天气”等等，最终将输入简化为14个，即双方的编号、血量、蓝量、技能符卡量、位置横坐标、位置纵坐标、动作状态，其中动作状态包含各个基本动作以及简单组合。我们调用DQN算法进行训练，以胜率为基本指标，以双方血量变化和胜率为基本奖励（reward），尝试了使用不同的神经网络参数进行训练，取得了不同阶段的成果。

三、强化学习算法

DQN 法是一种基于深度学习的强化学习算法，最初由 DeepMind 团队提出。其来源可以追溯1992年由Watkins 提出的 Q-Learning 算法，Q-Learning 是一种基于动态规划的强化学习算法，该算法通过迭代更新状态-动作对的 Q 值来学习最优策略。[1] DQN算法在强化学习方法的基础上增加了深度卷积神经网络，实现了从高维数据到动作空间的端到端映射[3]，在游戏AI领域表现出良好性能。

**第三部分 实现过程**

一、第一代——防御型

我们首先研发了第一代东方非想天则格斗AI——防御型。在多代的训练过程中，我们注意到该AI在面对掉血时受到了过量的负反馈，导致其采用一种持续防御的策略来减少即时的负反馈，这完全并不是理想的行为表现。

为了解决这一问题，我们采取了一系列调整和改进措施。其中之一是调整奖励机制，特别是增大了对于成功打掉对方血量所给予的奖励比例。通过这一改变，我们的目标是激励AI更积极主动地攻击对手，而不仅仅是被动地采取防御措施。

通过提高对成功攻击的奖励，我们期望AI能够更好地理解进攻性策略的重要性，并在实际对战中更灵活地应对对手的行为。这一调整在后续的实验和训练中得到验证，使得第一代防御型AI的行为更加多样化和适应性更强。这个解决方案的思路是通过优化奖励机制，引导AI在训练中更好地理解游戏目标，并更主动地追求成功攻击。这种调整不仅提高了AI的学习效率，也使其在实际游戏中展现更为出色的表现。

二、第二代——局部最优型

我们然后研发了第二代东方非想天则格斗AI——局部最优型。然而，在训练到后期时，我们观察到该AI开始过度依赖某一特定招式，因为该招式能够获得相对较大的正向奖励。这导致AI逐渐陷入一种局部最优的状态，表现出缺乏多样性和灵活性的问题。

为了解决这一问题，我们采取了一系列创新性的解决方案。其中之一是引入对于重复使用相同招式的负奖励机制。通过为频繁使用的招式施加负向奖励，我们的目标是鼓励AI更加多样地运用其技能，避免过度专注于某一特定招式而忽略其他潜在有效的策略。此外，我们调整了奖励函数的权重分配和超参数大小，确保正向奖励不仅仅依赖于某一招式的效果，而是更全面地考虑多个因素，也使得AI能够探索更多的状态。这有助于平衡AI在决策中的权衡，使其更灵活地适应不同的战斗场景。通过这些改进措施，我们的目标是使第二代局部最优型AI在实际对战中表现得更为多样化、灵活且具有挑战性。

三、第三代——全面优秀型

在制作过程中，我们迈向了第三代《东方非想天则格斗AI》——全面优秀型。通过前两代AI的经验和教训，我们致力于创造一种更加全面、灵活且优秀的格斗人工智能。在第三代的设计中，我们对训练框架进行了深刻的改进，以使AI更好地理解游戏规则、对手行为，并制定更具战略性的决策。以下是第三代的一些关键改进：

（1）优化奖励机制： 我们进一步细化奖励机制，考虑到更多的游戏因素，包括板边位置、技能连击、防御敌人进攻、反击成功等等。通过精心设计的奖励机制，我们旨在激发AI更广泛的游戏策略，并在不同场景中作出更为智能的决策。

（2）全局学习： 针对第二代中局部最优的问题，我们引入了全局学习的概念。通过更广泛的训练数据和更强大的神经网络结构，第三代AI能够更好地理解整个游戏的动态，避免陷入单一策略的困境。

（3）动态调整超参数： 我们实现了动态调整超参数的算法，使得AI能够在不同阶段自适应地调整学习速度和决策权重。这种动态性的调整有助于适应游戏的发展和对手的变化，使AI更具应变能力。

（4）加快训练速度：我们的游戏运行一场平均需要1分钟左右，这是比较庞大的时间，因此很难在短时间内训练多代。我们考虑了加速游戏进程、加快训练一代中止条件（仅剩百分之30血量时就判定胜负）等等方案，取得不错效果。

和已有的研究论文提到的一样，我们的实验过程中也出现了一些有趣的现象：人工智能越来越倾向于选择更稳健和多变的格斗方法。[4]我们的AI自己学到了知识，它会在敌人距离自己远的时候倾向于释放远程技能，也会主动尝试并且记住一些连招等等。

通过这些创新性的改进，我们的第三代AI在全面性、灵活性和战斗智能上取得更大的突破，最终胜率已经非常可观，可以比较轻松地战胜该游戏本身自带的最高难度AI。

**第四部分 综合与讨论**

一、reward函数的设计

根据对东方非想天则实时动作状态而做出奖励的方式，容易出现奖励值稀疏的问题，这是因为动作的奖励值不是孤立存在的而是不断累加的，而奖励值设定的目标也不相同。为解决奖励值的稀疏问题，最有效的方法是对DQN 算法模型 的奖惩函数进行优化。[5]

我们的最终reward考虑了如下情况：

（1）敌人血量减少（权重较高）和自己血量减少（权重较低）

（2）自己重复出招给惩罚

（3）防御掉敌人攻击给奖励

（4）打出combo对敌方造成大量伤害给出奖励

以及一些细微的方面。这样的reward函数已经比较符合我们的需求。

二、全面优化

（1）图象辅助格斗

我们虽然能够直接侵入内存获取游戏参数，但是例如对手发射的子弹这样的变化状态多的参数并不能轻易获取。而进行图象捕捉和传入进行决策面临着效率低的问题。于是，我们着手优化图象处理的流程，采用了高效的图象处理算法，以在捕捉和分析游戏画面时降低计算负担，引入了并行计算的策略，通过同时处理多个图象帧，提高了图象处理的速度。而我们的图象仅仅起到辅助训练的作用，并非主导。

（2）改变训练对手

我们的AI在已经能打败电脑自带AI后，还进行了更多的训练。例如，将对手换为我们训练过的其他AI模型进行模型互相训练。同时，我们还尝试联机和真实玩家对战，也取得了很好的效果，进一步增强了我们的模型。

（3）细化训练

对格斗AI进行训练的时候，我们一开始的训练过于笼统。后来，我们尝试了分人物训练的方法，包括对自身人物和对手人物的独立训练，以更好地模拟实际对战情境。我们期望AI能够更加深入地理解每个角色的特点、技能和战术风格。这有助于提高AI在不同人物间的切换和适应能力，使其能够更灵活地应对多样化的对手。这类似于一种特征学习，针对每个人物，我们引入了更细致的特征学习。通过对每个人物的动作、技能和反应等方面进行深入分析，我们能够更精确地捕捉到人物个性化的特征。这种人物特征学习不仅有助于提高AI在模拟对手行为时的真实感，还增强了AI对于游戏中不同人物间差异的理解。

三、反思与不足

（1）方法较为单一

制作格斗AI，还有很好的其他想法。例如，Giovanna[6]利用基因编程制作了强大的格斗AI。就DQN算法而言，唐振韬[2]提到构建合理高效的前向模型显得尤为重要。在保证实时性的前提下, 进一步提升前向模型的系统辨识度是当前的研究重点之一。这些良好的方法，限于设备和能力，我们没有实现。

（2）模型的泛化能力不够

尽管我们的AI在东方非想天则游戏中取得了显著的成就，但我们也意识到其在泛化能力方面存在一些限制。换句话说，尽管在特定游戏环境下表现出色，但我们的格斗AI并不具备通用性，无法轻松适应其他格斗游戏或不同的虚拟格斗场景。

为了解决这一挑战，我们拟在后续研究中采取以下策略：

跨游戏数据集引入： 通过引入跨游戏的数据集，包括不同格斗游戏的信息。通过让AI学习并适应不同游戏规则和战术风格，我们试图提升其在不同游戏环境中的泛化能力。这种方法旨在使格斗AI更具通用性，能够在不同游戏中表现出相似的高水平格斗技能。

模型结构调整： 我们对AI的模型结构进行了调整，以更好地适应不同格斗游戏的特征。通过引入更灵活的模型设计，我们希望提高AI对于不同游戏中规则和战术变化的适应能力，从而增强其泛化性。

**第五部分 总结**

通过对DQN算法在东方非想天则格斗AI上的应用，我们实现了我们制作的三代该格斗AI的发展，不断优化奖励函数和训练框架。在讨论中，我们强调了奖励函数的多因素考虑、全面的优化策略，以及对模型泛化能力的反思。未来，我们计划引入跨游戏数据集和调整模型结构，以提高格斗AI的通用性。这一研究为格斗AI领域的研究，特别是在有限资源和能力情况下的个人探索和研究，提供了思路和启示。

参考文献

1. *<复杂环境下的DQN（Dee...-Network）算法研究\_史鸿远.pdf>.*

2. *<实时格斗游戏的智能决策方法\_唐振韬.pdf>.*

3. *<基于DQN的多智能体深度强化学习运动规划方法\_史殿习.pdf>.*

4. Liang, H., J. Li, and P.K.R. Maddikunta, *A Study on the Agent in Fighting Games Based on Deep Reinforcement Learning.* Mobile Information Systems, 2022. **2022**: p. 1-8.

5. *<基于D-DQN强化学习算法的双足机器人智能控制研究\_李丽霞.pdf>.*

6. Martinez-Arellano, G., R. Cant, and D. Woods, *Creating AI Characters for Fighting Games Using Genetic Programming.* IEEE Transactions on Computational Intelligence and AI in Games, 2017. **9**(4): p. 423-434.