量子强化学习的异步训练

Samuel Yen-Chi Chen[1]

[1]富国银行

（日期：2023年1月13日）

arXiv:2301.05096v1 [quant-ph] 12 Jan 2023

摘要

近年来，得益于量子计算（QC）和机器学习（ML）的发展，量子机器学习（QML）受到了广泛关注。强化学习（RL）是可用于解决具有挑战性的序贯决策问题的机器学习范式之一。研究已表明，经典强化学习能够成功完成许多困难任务。构建量子强化学习智能体的一种主流方法依赖于变分量子电路（VQC）。然而，使用变分量子电路训练量子强化学习算法需要大量计算资源，这一问题阻碍了对各种量子强化学习应用的探索。本文通过异步训练量子强化学习智能体来应对这一挑战。具体而言，我们选择对优势演员-评论家变分量子策略进行异步训练。通过数值模拟，我们展示了在所研究的任务中，异步训练量子强化学习智能体的表现可达到或优于模型规模和架构相近的经典智能体。

I. 引言

量子计算（QC）被认为有望在某些经典计算机难以解决的任务中实现计算优势 [1]。尽管潜力巨大，但当前量子计算机缺乏纠错能力，因此难以在这些“含噪中等规模量子”（NISQ）设备上有效实现复杂的量子电路 [2]。要充分利用 NISQ 设备所提供的量子优势，就必须开发专用的量子电路架构。

混合量子-经典计算框架的最新进展 [3] ，该框架同时利用经典计算和量子计算。根据这种方法，预计可从量子处理中受益的某些计算任务将在量子计算机上执行，而其他任务（如梯度计算）则由经典计算机完成。这种混合方法旨在发挥两类计算各自的优势，以应对广泛的任务。利用变分量子电路（VQC）的混合算法已被证明在各种机器学习任务中有效。变分量子电路是具有可调参数的量子电路子类，将其纳入量子机器学习模型后，已在广泛的任务中取得成功 [3, 4]。

强化学习（RL）是机器学习的一个分支，研究序贯决策任务。基于深度神经网络的强化学习在复杂任务中取得了显著成果，达到人类水平 [5] 或超越人类的表现 [6]。然而，量子强化学习仍是一个发展中的领域，存在许多尚未解决的问题和挑战。现有的量子强化学习模型大多基于变分量子电路 [7–11]。尽管这些模型已被证明在各种基准任务中表现良好，但训练它们需要大量计算资源。漫长的训练时间限制了对量子强化学习广泛应用可能性的探索。本文提出一种用于量子强化学习智能体的异步训练框架。我们重点研究多个智能体实例并行运行时，优势演员-评论家量子策略的异步训练。

我们通过数值模拟表明，在所考察的各种基准任务中，量子模型的表现可能优于或接近经典模型。此外，所提出的训练方法具有实际优势：所需训练时间显著更短，从而能够应用于更多量子强化学习场景。

本文的结构如下：在第二节中，我们概述相关

的先前工作，并将我们的方案与这些方法进行比较。在第三节中，我们简要介绍强化学习所需的背景知识。在第四节中，我们介绍变分量子电路（VQCs）的概念，它们是量子强化学习智能体的构建模块。在第五节中，我们介绍所提出的量子 A3C 框架。在第六节中，我们描述实验设置并展示实验结果。最后，在第七节中，我们给出一些总结性评论。

II. 相关工作

催生量子强化学习（QRL）的研究 [12] 可追溯至 [13]。然而，该框架需要量子环境，而大多数现实场景可能并不具备这一条件。采用类 Grover 方法的后续研究包括[14, 15]。量子线性系统求解器也被用于实现量子策略迭代[16]。我们将重点关注近年来基于变分量子电路的量子强化学习在经典环境中的进展。首个基于变分量子电路的量子强化学习 [7], 是深度 Q学习（深度 Q 网络）的量子版本，针对 Frozen-Lake 和 Cognitive-Radio 等测试环境中的离散观测与动作空间。此后，该领域更为复杂的量子深度 Q 网络研究开始考虑 Cart-Pole 等具有连续观测空间的环境 [8, 9]。这一方向的进一步发展包括采用量子长短期记忆网络（QLSTM）等量子循环神经网络作为价值函数近似器 [17] ，以应对部分可观测性或需要更长时间记忆先前步骤等挑战。研究者还开发了量子-经典混合线性求解器等多种方法来求得价值函数[18]。为了提升智能体的收敛性，研究 [19], 也在变分量子电路框架中实现了双重深度 Q 网络（DDQN）等对深度 Q 网络的改进，并将量子强化学习应用于机器人导航任务。近年来，量子强化学习的发展促成了一些旨在直接学习策略函数（记作π）的框架。除了学习价值函数（如 Q函数）之外，这些框架还能够学习给定问题的最优策略。例如，论文[10]介绍了如何使用 REINFORCE 算法实现量子策略梯度强化学习。随后，研究 [11] 探讨了采用变分量子电路实现一种改进的策略梯度算法——近端策略优化，并表明即使参数数量较少，量子模型也能胜过经典对应模型。研究展示了策略梯度可证明的量子优势。

研究 [20]。其他研究（例如 [21], ）探讨了变分量子电路的各种后处理方法对量子策略梯度性能的影响。近年来，人们提出了若干改进的量子策略梯度算法，包括演员-评论家 [22] 和软演员-评论家（SAC） [23, 24]。这些改进旨在进一步提升量子强化学习方法的效率和效果。量子强化学习也已应用于量子控制领域 [25] ，并扩展至多智能体设置 [26–28]。研究 [29] 首次探索了将进化优化用于量子强化学习。在他们的工作中，多个智能体被初始化并行运行，表现最佳的智能体被选为父代，以生成下一代智能体。在研究 [30],中，作者探讨了将深度确定性策略梯度（DDPG）算法等先进的量子策略梯度方法用于连续动作空间中的量子强化学习。

在本研究中，我们在先前关于量子策略梯度的研究基础上 [10, 11, 22] ，引入了一种用于量子策略学习的异步训练方法。以往的方法采用单线程训练，而我们的方法采用异步方式，这可能带来实际优势，例如利用多核 CPU 计算资源缩短训练时间，以及未来使用多个量子处理单元（QPU）的潜力。我们的方法与 [29], 中介绍的进化式量子强化学习方法存在一些相似之处，后者同样利用了并行计算资源。然而，我们的方法有所不同：各个智能体可以直接异步地将其梯度共享给共享的全局梯度，而无需等待所有智能体完成后再计算适应度并创建下一代智能体。这一特性可能会进一步提高训练过程的效率。这些贡献代表了量子强化学习领域的一项新颖进展。

III. 强化学习

强化学习(RL) 是一种机器学习框架，智能体 通过与环境 $\mathcal{E}$ 在离散时间步 [31]进行交互，学习如何实现给定目标。智能体在每个时间步观察状态 s_t t ，然后根据当前策略 π 从动作空间 $\mathcal{A}$ 中选择一个动作 a_t 。策略将特定状态 s_t 映射为选择 $\mathcal{A}$中某个动作的概率。执行动作 a_t后，智能体会从环境中获得一个标量奖励 r_t 以及下一时间步的状态 s_{t+1} 。

对于情节式任务，这一过程会在多个时间步中重复进行，直到智能体到达终止状态或达到允许的最大步数。在训练过程中观察状态s_t 时，智能体旨在最大化期望回报，该期望回报可表示为策略 π下状态 s 的价值函数 $V^{\pi}(s) = \mathbb{E}\left[R_t|s_t = s\right]$，其中$R_t = \sum_{t'=t}^{T} \gamma^{t'-t} r_{t'}$ 是回报，即从时间步 t开始的折扣奖励总和。价值函数还可进一步表示为 $V^{\pi}(s) = \sum_{a\in\mathcal{A}} Q^{\pi}(s,a)\pi(a|s)$，其中动作价值函数或$Q$值函数 $Q^{\pi}(s,a) = \mathbb{E}[R_t|s_t = s, a]$ 是根据策略 π在状态 s 中选择动作 $a \in \mathcal{A}$ 的期望回报。 Q-学习是一种强化学习算法，通过以下公式优化 $Q^{\pi}(s,a)$

$$Q\left(s_t, a_t\right) \leftarrow Q\left(s_t, a_t\right) + \alpha\left[r_t + \gamma \max_a Q\left(s_{t+1}, a\right) - Q\left(s_t, a_t\right)\right]. \tag{1}$$

与基于价值的强化学习技术（例如 Q-学习）不同，后者依赖于学习一个价值函数，并在每个时间步使用它来指导决策；策略梯度方法则着重于直接优化策略函数，记作 $\pi(a|s;\theta)$，其参数化形式由 θ 给出。参数 θ 通过对期望总回报 $\mathbb{E}[R_t]$ 执行梯度上升过程来更新。策略梯度算法的一个著名例子是 REINFORCE 算法，于 [32] 中提出。在标准的 REINFORCE 算法中，参数 θ 沿着方向 $\nabla_\theta \log \pi(a_t|s_t;\theta) R_t$ 更新，这是 $\nabla_\theta \mathbb{E}\left[R_t\right]$ 的无偏估计。然而，这种策略梯度估计通常具有较高的方差，使训练变得困难。为降低该估计的方差，同时保持其无偏性，可以从回报中减去一个称为基线的项。这个基线记作 $b_t(s_t)$，是状态 s_t 的一个学习所得函数。由此得到的更新为 $\nabla_\theta \log \pi(a_t|s_t;\theta)\left(R_t - b_t(s_t)\right)$。在强化学习中，基线 $b_t(s_t)$ 的一种常见选择是价值函数 $V^{\pi}(s_t)$ 的估计。采用这种基线通常能得到方差更低的策略梯度 [31] 估计。量 $R_t - b_t = Q(s_t, a_t) - V(s_t)$ 可以解释为在状态 s_t 下采取动作 a_t 的优势 $A(s_t, a_t)$。直观而言，优势可以理解为动作 a_t 相对于状态 s_t 下平均价值的“好坏程度”。这种方法称为优势演员-评论家（A2C）方法，其中策略 π 是演员，而作为基线的价值函数 V 则是评论家 [31]。

异步优势演员-评论家（A3C）算法 [33] 是 A2C 方法的一种变体，它通过并行化让多个并发演员学习策略。

强化学习智能体的异步训练涉及在环境的多个实例中运行多个智能体，使它们能够在任意时间步遇到不同的状态。状态或观测之间相关性的降低，增强了演员-评论家等同策略强化学习算法的数值稳定性 [33]。此外，异步训练无需维护大型回放记忆，因此降低了内存需求[33]。通过利用一组演员计算出的优势和梯度，A3C 展现出出色的样本效率和稳健的学习性能，因此成为强化学习领域的常见选择。

IV. 变分量子电路

变分量子电路（VQCs），也称为参数化量子电路（PQCs），是一类包含可调参数的量子电路。这些参数可以利用经典机器学习中的各种技术进行优化，包括基于梯度的方法和非基于梯度的方法。图 1 中央部分展示了变分量子电路的一般示意图。

变分量子电路的三个主要组成部分是编码电路、变分电路和量子测量层。编码电路记作 $U(\mathbf{x})$，负责将经典值转换为量子态；变分电路记作 $V(\theta)$，是变分量子电路中可学习的部分。另一方面，量子测量层用于从电路中提取信息。通常会反复执行电路（也称为“测量次数”），以获得每个量子比特的期望值。常见的选择是使用 Pauli-Z 期望值。得到的值不是二进制整数，而是浮点数。此外，其他组件（例如额外的变分量子电路，或深度神经网络等经典组件）也可以处理从电路中获得的值。

变分量子电路可以与其他经典组件协同工作，例如张量网络（TN） [29,34, 35]和深度神经网络（NN），以执行数据预处理（如降维）或后处理（如缩放）。如图1所示，我们将此类变分量子电路称为装饰型变分量子电路。整个模型可以通过基于梯度的方法 [34, 35] 或无梯度方法 [29]进行端到端训练。对于基于梯度的方法，可以将整个模型表示为有向无环图（DAG），然后进行反向传播

应用。此类端到端优化的成功，取决于计算量子梯度的能力，例如参数移位 规则 [36]。基于 VQC 的量子机器学习模型已在分类 [34–38], 、自然语言处理 [39–41]和序列建模 [42, 43]等领域取得成功。

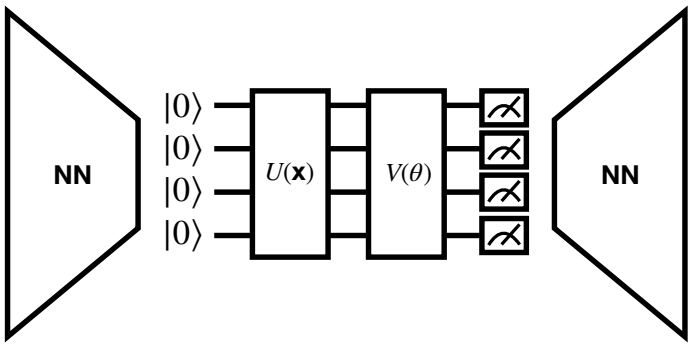


图 1. **混合变分量子电路（VQC）架构**。 混合 VQC 架构包含一个 VQC，以及位于 VQC 前后的经典神经网络（NN）。NN 可用于降低输入数据的维度，并优化 VQC 的输出。

V. 量子 A3C

所提出的量子异步优势演员-评论家（QA3C）框架由两个主要部分组成：一个全局共享内存 以及每个智能体各自的进程专属内存 。全局共享内存保存着装饰型 VQC 策略和价值参数；当某个进程上传自身梯度以更新参数时，这些参数便会随之修改。每个智能体都有自己的进程专属内存，用于保存本地装饰型 VQC 策略和价值参数。在单个进程中，这些局部模型用于在一个回合期间生成动作。满足特定条件后，本地模型参数的梯度会上传至全局共享内存，并相应地修改全局模型参数。更新后的全局模型参数随即下载到刚刚上传自身梯度的本地智能体。图 2 展示了 QA3C 的整体概念。

我们使用图 1 所示的着装变分量子电路构建量子策略 $\pi(a_t \mid s_t;\theta)$ 和价值 $V(s_t;\ \theta_v)$ 函数，其中变分量子电路遵循图中所示的架构

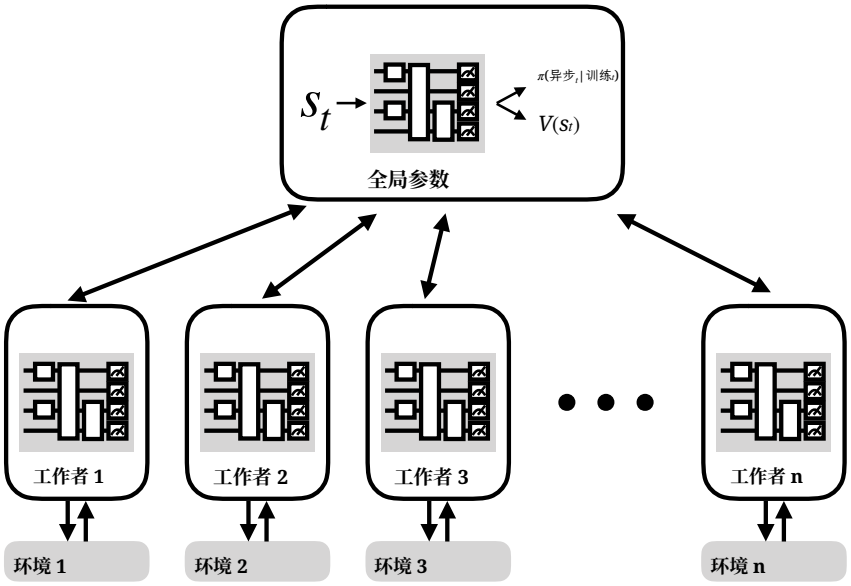


图 2. **量子异步优势演员-评论家（异步优势演员-评论家）学习器**。所提出的量子 A3C 包含全局共享参数和多个并行工作器。每个本地智能体中的动作生成过程使用存储在进程专属内存中的装饰型 VQC 策略和价值函数来执行。在满足特定条件后，本地模型参数的梯度会上传至全局共享内存，并在那里更新全局模型参数。随后，更新后的全局模型参数会立即下载到刚刚上传自身梯度的本地智能体。

图 3。该变分量子电路架构已应用于量子循环神经网络（QRNN） [42], 量子循环强化学习 [17], 量子卷积神经网络 [44], 联邦量子分类 [38] 等研究，并且在某些条件下，其性能优于经典对应模型。此外，我们在变分量子电路前后分别使用经典深度神经网络来降低数据维度，并微调变分量子电路的输出。该混合架构中的神经网络组件由用于维度转换的单层网络构成。具体而言，变分量子电路前的网络是一个线性层，其输入维度等于观测向量的大小，输出维度等于变分量子电路中的量子比特数。变分量子电路后的网络是线性层，其输入维度等于变分量子电路中的量子比特数，输出维度等于动作数（用于演员函数 $\pi(a_t \mid s_t; \theta)$）或 1（用于评论家函数 $V(s_t;\ \theta_v)$）。这些层用于转换

变分量子电路的输出用于演员-评论家算法。每经过 S 步或智能体到达终止状态时，都会更新策略和价值函数。算法的详细内容，例如梯度更新公式，见算法1。

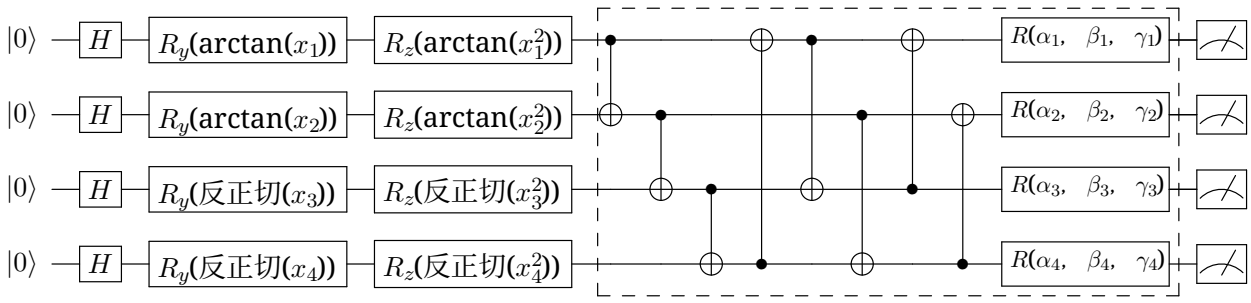


图 3. **用于量子 A3C 的变分量子电路架构**。此处使用的变分量子电路包括用于编码经典值的 R_y 和 R_z x、多个用于纠缠量子比特的受控非门、通用酉旋转 R，以及最终测量。变分量子电路的输出由 Pauli-Z 期望值构成，这些值通过多次运行（测量次数）电路获得。随后，经典神经网络会对这些值进行处理，以供进一步使用。此处以一个 4 量子比特系统为例；不过，可根据所关注的问题扩大或缩小系统规模。本研究中，量子比特数量为 8。

VI. 实验与结果

A. 测试环境

1. Acrobot

OpenAI Gym 中的 Acrobot 环境 [45] 由两个线性连接的连杆组成，其中一端固定。连接两个连杆的关节可通过施加转矩来驱动。目标是从链条向下悬垂的位置开始，尽可能少地用几步将链条的自由端摆过预定高度。该环境中的观测是一个六维向量，包含两个旋转关节角度的正弦值和余弦值，以及它们的角速度。智能体可以选择三种动作之一：对可驱动关节施加 -1、0 或 $+1$ 转矩。若某个动作使自由端达到目标高度，则获得 0 奖励，并结束该回合。任何未能达到目标高度的动作都会获得 -1奖励。奖励阈值为 -100。



图 4。 **OpenAI Gym 中的 Acrobat 环境**。

2. Cart-Pole

Cart-Pole 是一种常用于评估简单强化学习模型的环境，也是 OpenAI Gym [45] 中的标准示例（见图 5）。在该环境中，一个固定接合点将杆连接到一辆小车，小车可在无摩擦轨道上水平移动。摆起初直立，目标是通过左右移动小车，使其尽可能保持在初始位置附近。每个时间步，强化学习智能体都会根据收到的观测，学习采取恰当的动作。该环境中的观测是一个四维向量 s_t，包含小车位置、小车速度、杆角度以及杆顶端的杆速度。每个时间步中，只要杆接近直立，就会获得 $+1$ 奖励。如果杆偏离垂直方向超过 15 度，或小车偏离中心超过 2.4 个单位，回合就会结束。

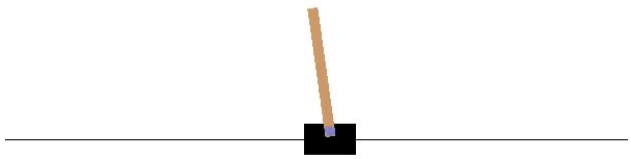


图 5。 **OpenAI Gym 中的倒立摆环境**。

3. MiniGrid-SimpleCrossing

MiniGrid-SimpleCrossing环境 [46] 更加复杂，为强化学习智能体提供了大得多的观测输入。在此场景中，强化学习智能体通过观测接收一个 $7 \times 7 \times 3$ =147 维向量，并必须从提供六种选项的动作空间 $\mathcal{A}$中选择一个动作。需要注意的是，这个 147 维向量是对环境的一种紧凑、高效的表示，而非真实像素。智能体可以从动作空间 $\mathcal{A}$ 中的六个动作 0,$\cdots$,5 中选择。这些动作分别是向左转、向右转、向前移动、拾取物体、放下携带的物体 和切换。在这种情况下，只有前三个动作会产生实际效果。智能体需要学会这一点。在这个环境中，智能体到达目标后会获得 1 的奖励。根据公式 $1 - 0.9 \times$ (步数/允许的最大步数)，从该奖励中扣除惩罚；允许的最大步数定义为 $4 \times n \times n$，而 n 是网格大小 [46]。在本研究中， n 设为 9。这种奖励机制带来了挑战，因为奖励是稀疏的，也就是说，智能体到达目标之前不会获得奖励。如图 6 所示，智能体（以红色三角形表示）需要找到从起点到目标（以绿色表示）的最短路径。我们考虑了该环境中的三种情况：MiniGrid-SimpleCrossingS9N1-v0、MiniGrid- SimpleCrossingS9N2-v0 和 MiniGrid-SimpleCrossingS9N3-v0。其中， N 表示从起点到目标穿越墙壁的有效通道数量。

B. 超参数与模型规模

在所提出的QA3C中，我们使用Adam优化器，学习率为 1×10^{-4}、 $\beta_1 = 0.92$和 $\beta_2 = 0.999$。局部智能体每 $S = 5$ 步使用全局共享内存更新参数。折扣因子 γ 设置为0。9。对于变分量子电路，我们将量子比特数设置为8，并使用两个变分层。因此，每个变分量子电路包含$8 \times 3 \times 2 = 48$ 个量子参数。演员和评论家各自拥有一个变分量子电路，因此量子参数总数为96。在本文所考虑的不同测试环境中，变分量子电路架构保持不变。如第五节所述，在变分量子电路之前和之后均使用单层网络，以转换数据的维度。变分量子电路之前的网络具有基于环境的输入维度，而这些环境

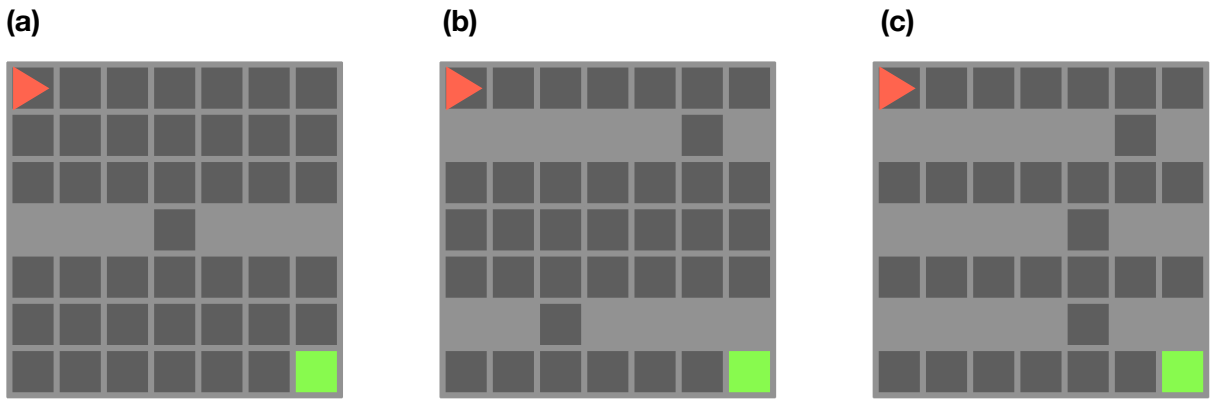


图6。**来自MiniGrid的简单穿越环境**。本文考虑的来自MiniGrid-SimpleCrossing的三个环境。在每个环境中，两侧还各有延伸1个单位的墙（图中未显示）。（a）、（b）和（c）分别表示MiniGrid-SimpleCrossingS9N1-v0、MiniGrid-SimpleCrossingS9N2-v0和MiniGrid-SimpleCrossingS9N3-v0环境的示例。

智能体要解决的问题。对于经典基准，我们考虑与着装变分量子电路模型非常相似的模型。具体而言，我们使经典模型的架构与图1所示的架构相似，同时将 8 量子比特变分量子电路替换为一个输入和输出维度均为 8 的单层结构。这使得该架构与量子模型非常相似，参数数量也非常接近。我们在表 I 中总结了参数数量。我们使用开源的 PennyLane 软件包 [47]

	QA3C			经典
	经典	量子	总计	总计
Acrobot	148	96	244	292
Cart-Pole	107	96	203	251
简单穿越	2431	96	2527	2575

表 I。**参数数量**。 我们提供了所提出的 QA3C 模型的参数数量详情，该模型同时包含量子组件和经典组件。经典基准采用了与量子模型相似的架构进行设计，因此模型规模也相近。

用于构建量子电路模型，并将 PyTorch 作为整体机器学习

框架。本研究使用 80 个 CPU 核心，因此并行智能体的数量也是 80。我们展示了模拟结果，其中对过去 100 个回合的得分取平均值。

C. 结果

1. Acrobot

我们首先评估模型在 Acrobot 环境中的表现。本实验的仿真结果见图 7。总回合数为 100,000。如图所示，量子模型在训练初期逐步提升，而经典模型则难以改进其策略。就平均得分而言，量子模型的表现优于经典模型。此外，量子模型呈现出更稳定的收敛模式，在达到最佳得分后没有明显波动或性能崩溃。这些结果表明，量子模型在该环境中可能更加稳健可靠。

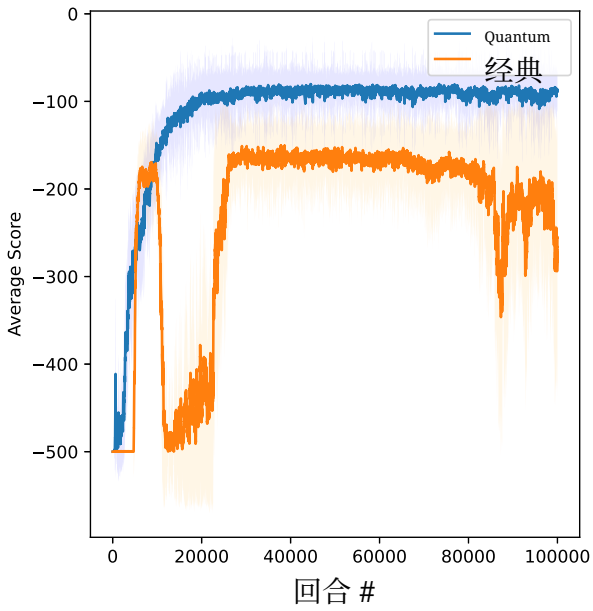


图 7. **结果：Acrobot 环境中的量子 A3C**。

2. Cart-Pole

接下来的实验在 Cart-Pole 环境中进行。总回合数为 100,000。如图 8 所示，量子模型取得的得分显著高于经典模型。尽管经典模型在训练初期学习速度更快，但量子模型最终超越了它，并取得了更高的得分。这些结果表明，量子模型在该环境中可能更加有效。

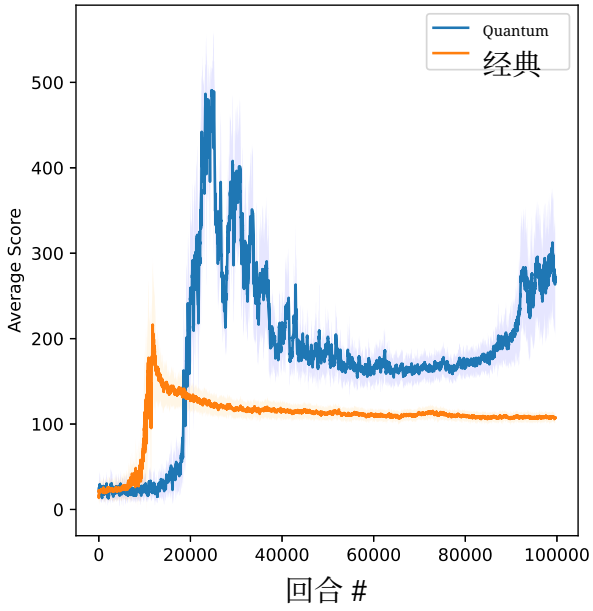


图 8. **结果：倒立摆环境中的量子 A3C**。

3. MiniGrid-SimpleCrossing

最终实验在 MiniGrid-SimpleCrossing环境中进行，共包含 100,000 个回合。如图 9所示，在三个场景 MiniGrid-SimpleCrossingS9N1-v0、MiniGrid-SimpleCrossingS9N2-v0 和 MiniGrid-SimpleCrossingS9N3-v0 中，量子模型在其中两个场景（MiniGrid-SimpleCrossingS9N2-v0 和 MiniGrid-SimpleCrossingS9N3-v0）中优于经典模型，表现出更快的收敛速度和更高的得分。即使在剩余的场景 MiniGrid-SimpleCrossingS9N1-v0 中，两个模型的性能差异

较小。

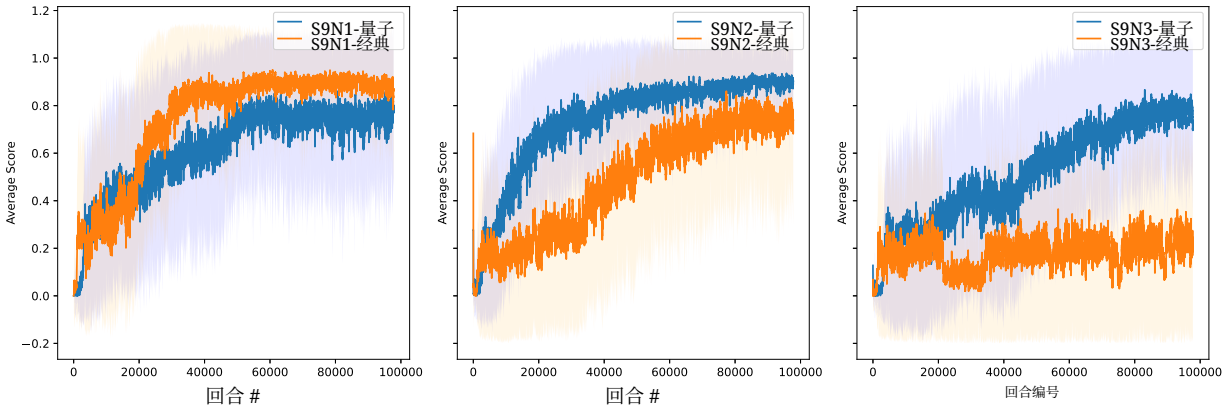


图9。 **结果：MiniGrid-SimpleCrossing环境中的量子 A3C** .

VII. 结论

本研究展示了量子强化学习智能体异步训练框架的有效性。通过数值模拟，我们表明，在所考虑的基准任务中，经过异步训练的优势演员-评论家量子策略，其性能可以优于或媲美架构和规模相近的经典模型。该技术为通过并行化加速量子强化学习智能体的训练提供了一种策略，并可能应用于各种现实场景。

致谢

本文所表达的观点仅代表作者本人，并不代表富国银行的观点。本文仅供参考。本文中的任何内容均不应被视为投资建议。富国银行不作任何明示或暗示的保证，并明确声明不承担与本文相关的任何法律、税务及会计影响。

附录 A：算法

1. 量子异步优势演员-评论家

算法1 量子异步优势演员-评论家学习（每个演员-学习器进程的算法）

定义 全局更新参数 S

假设 全局共享混合变分量子电路策略参数 θ

假设全局共享的混合变分量子电路价值参数 θ_v

假设全局共享的回合计数器 $T = 0$

假设进程专属的混合变分量子电路策略参数 θ'

假设进程专属的混合变分量子电路价值参数 θ'_v

初始化进程专属计数器 $t = 1$

当 $T < T_{max}$ **do**

重置梯度 $d\theta \leftarrow 0$ 和 $d\theta_v \leftarrow 0$

设置 $t_{start} = t$

重置环境并获取状态 s_t

当 s_t 非终止或 $t - t_{start} < t_{max}$ **执行**

根据策略 $\pi(a_t|s_t;\theta')$执行 a_t

接收奖励 r_t 和新状态 s_{t+1}

更新进程专属计数器 $t \leftarrow t + 1$

如果 $t \bmod S = 0$ 或到达终止状态 **则**

设置 $R = \begin{cases} 0 & \text{对于终止 } s_t \\ V(s_t,\ \theta'_v) & \text{对于非终止 } s_t \end{cases}$

对于 $i \in \{t-1, \ldots, t_{\text{start}}\}$，**执行**

$R \leftarrow r_i + \gamma R$

累积关于 θ' 的梯度：$d\theta \leftarrow d\theta + \nabla_{\theta'} \log \pi(a_i \mid s_i;\theta')\,(R - V(s_i;\ \theta'_v))$

累积关于 θ'_v 的梯度：$d\theta_v \leftarrow d\theta_v + \partial\,(R - V(s_i;\theta'_v))^2 / \partial\theta'_v$

结束循环

使用 $d\theta$ 对 θ 执行异步更新，并使用 $d\theta_v$ 对 θ_v 执行异步更新

从全局参数更新进程专属参数：$\theta' \leftarrow \theta$ 和 $\theta'_v \leftarrow \theta_v$

结束条件判断

结束循环

结束循环

[1] M. A. Nielsen 和 I. L. Chuang，《量子计算与量子信息》，2010。[2] J. Preskill，《NISQ 时代及其之后的量子计算》，*Quantum*，第 2 卷，第 79 页，2018。[3] K. Bharti、A. Cervera-Lierta、T. H. Kyaw、T. Haug、S. Alperin-Lea、A. Anand、M. Deg- roote、H. Heimonen、J. S. Kottmann、T. Menke、*et al.*，《含噪中等规模量子算法》，《现代物理评论》，第 94 卷，第 1 期，第 015004 页，2022。[4] M. Cerezo、A. Arrasmith、R. Babbush、S. C. Benjamin、S. Endo、K. Fujii、J. R. McClean、K. Mitarai、X. Yuan、L. Cincio、*et al.*，《变分量子算法》，《自然综述：物理》，第 3 卷，第 9 期，第 625–644 页，2021。[5] V. Mnih、K. Kavukcuoglu、D. Silver、A. A. Rusu、J. Veness、M. G. Bellemare、A. Graves、M. Riedmiller、A. K. Fidjeland、G. Ostrovski、*et al.*，《通过深度强化学习实现人类水平控制》，《自然》，第 518 卷，第 7540 期，第 529–533 页，2015。[6] D. Silver、J. Schrittwieser、K. Simonyan、I. Antonoglou、A. Huang、A. Guez、T. Hubert、L. Baker、M. Lai、A. Bolton、*et al.*，《在没有人类知识的情况下掌握围棋》，《自然》，第 550 卷，第 7676 期，第 354–359 页，2017。[7] S. Y.-C. Chen、C.-H. H. Yang、J. Qi、P.-Y. Chen、X. Ma 和 H.-S. Goan，《用于深度强化学习的变分量子电路》，*IEEE Access*，第 8 卷，第 141007–141024 页，2020。[8] O. Lockwood 和 M. Si，《利用量子变分电路进行强化学习》，载于人工智能与互动数字娱乐 *AAAI* 会议论文集，第 16 卷，第 245–251 页，2020。[9] A. Skolik、S. Jerbi 和 V. Dunjko，《Gym 中的量子智能体：一种用于深度 Q 学习的变分量子算法》，*Quantum*，第 6 卷，第 720 页，2022。[10] S. Jerbi、C. Gyurik、S. Marshall、H. J. Briegel 和 V. Dunjko，《用于强化学习的变分量子策略》，*arXiv* 预印本 *arXiv:2103.05577*，2021。[11] J.-Y. Hsiao、Y. Du、W.-Y. Chiang、M.-H. Hsieh 和 H.-S. Goan，《OpenAI Gym 中的非纠缠量子强化学习智能体》，*arXiv* 预印本 *arXiv:2203.14348*，2022。[12] N. Meyer、C. Ufrecht、M. Periyasamy、D. D. Scherer、A. Plinge 和 C. Mutschler，《量子强化学习综述》，*arXiv* 预印本 *arXiv:2211.03464*，2022。[13] D. Dong、C. Chen、H. Li 和 T.-J. Tarn，《量子强化学习》，*IEEE* 系统、人类与控制论汇刊（*B*辑：控制论）

第 *38* 卷，第 *5* 期，第 *1207–1220* 页，2008。[14] S. Wiedemann、D. Hein、S. Udluft 和 C. Mendl，《通过振幅估计和 Grover 搜索进行量子策略迭代——迈向强化学习的量子优势》，*arXiv* 预印本 *arXiv:2206.04741*，2022。[15] A. Sannia、A. Giordano、N. L. Gullo、C. Mastroianni 和 F. Plastina，《一种加速 Q 学习的混合经典-量子方法》，*arXiv* 预印本 *arXiv:2205.07730*，2022。[16] E. A. Cherrat、I. Kerenidis 和 A. Prakash，《通过策略迭代实现量子强化学习》，*arXiv* 预印本 *arXiv:2203.01889*，2022。[17]S. Y.-C. Chen，《量子深度循环强化学习》，*arXiv* 预印本 *arXiv:2210.14876*，2022。[18]C.-C. CHEN、K. SHIBA、M. SOGABE、K. SAKAMOTO 和 T. SOGABE，《基于混合量子-经典乌拉姆-冯·诺依曼线性求解器的量子动态规划算法》，日本人工智能学会年会论文集，第 JSAI2020 卷，第 2K6ES203–2K6ES203 页，2020。[19] D. Heimann、H. Hohenfeld、F. Wiebe 和 F. Kirchner，《用于机器人导航任务的量子深度强化学习》，*arXiv* 预印本 *arXiv:2202.12180*，2022。[20] S. Jerbi、A. Cornelissen、M. Ozols 和 V. Dunjko，《量子策略梯度算法》，*arXiv* 预印本 *arXiv:2212.09328*，2022。[21] N. Meyer、D. D. Scherer、A. Plinge、C. Mutschler 和 M. J. Hartmann，《采用优化动作解码的量子策略梯度算法》，*arXiv* 预印本 *arXiv:2212.06663*，2022。[22] M. Schenk、E. F. Combarro、M. Grossi、V. Kain、K. S. B. Li、M.-M. Popa 和 S. Vallecorsa，《用于欧洲核子研究中心束流线量子强化学习的混合演员-评论家算法》，*arXiv* 预印本 *arXiv:2209.11044*，2022。[23] Q. Lan，《变分量子软演员-评论家算法》，*arXiv* 预印本 *arXiv:2112.11921*，2021。[24] A. Acuto、P. Barill`a、L. Bozzolo、M. Conterno、M. Pavese 和 A. Policicchio，《用于机械臂控制的变分量子软演员-评论家算法》，*arXiv* 预印本 *arXiv:2212.11681*，2022。[25] A. Sequeira、L. P. Santos 和 L. S. Barbosa，《结合量子控制应用的变分量子策略梯度》，*arXiv* 预印本 *arXiv:2203.10591*，2022。[26]W. J. Yun、Y. Kwak、J. P. Kim、H. Cho、S. Jung、J. Park 和 J. Kim，《通过变分量子电路设计实现量子多智能体强化学习》，*arXiv* 预印本

arXiv:2203.10443，2022年。[27] R. Yan、Y. Wang、Y. Xu 和 J. Dai，“一种用于孤岛微电网分布式频率控制的多智能体量子深度强化学习方法”，*IEEE* 网络系统控制汇刊，第 9 卷第 4 期，第 1622–1632 页，2022年。[28] W. J. Yun、J. Park 和 J. Kim，“量子多智能体元强化学习”，*arXiv* 预印本 *arXiv:2208.11510*，2022年。[29] S. Y.-C. Chen、C.-M. Huang、C.-W. Hsing、H.-S. Goan 和 Y.-J. Kao，“通过进化优化实现变分量子强化学习”，《机器学习：科学与技术》，第 3 卷第 1 期，文章号 015025，2022年。[30] S. Wu、S. Jin、D. Wen 和 X. Wang，“连续动作空间中的量子强化学习”，*arXiv* 预印本 *arXiv:2012.10711*，2020年。[31] R. S. Sutton 和 A. G. Barto，《强化学习导论》。麻省理工学院出版社，2018年。[32] R. J. Williams，“用于连接主义强化学习的简单统计梯度跟随算法”，《机器学习》，第 8 卷第 3–4 期，第 229–256 页，1992年。[33]V. Mnih、A. P. Badia、M. Mirza、A. Graves、T. Lillicrap、T. Harley、D. Silver 和 K. Kavukcuoglu，“用于深度强化学习的异步方法”，载于国际机器学习会议，第 1928–1937 页，PMLR，2016。[34] S. Y.-C. Chen、C.-M. Huang、C.-W. Hsing 和 Y.-J. Kao，“一种端到端可训练的经典-量子混合分类器”，《机器学习：科学与技术》，第 2 卷第 4 期，文章号 045021，2021年。[35] J. Qi、C.-H. H. Yang 和 P.-Y. Chen，“Qtn-vqc：一种量子神经网络端到端学习框架”，*arXiv* 预印本 *arXiv:2110.03861*，2021年。[36] K. Mitarai、M. Negoro、M. Kitagawa 和 K. Fujii，“量子电路学习”，《物理评论*A*》，第 98 卷第 3 期，文章号 032309，2018年。[37] M. Chehimi 和 W. Saad，“基于量子数据的量子联邦学习”，载于*ICASSP 2022-2022 IEEE* 国际声学、语音与信号处理会议 *(ICASSP)*，第 8617–8621 页，IEEE，2022年。[38] S. Y.-C. Chen 和 S. Yoo，“联邦量子机器学习”，《熵》，第 23 卷第 4 期，文章号 460，2021年。[39] C.-H. H. Yang、J. Qi、S. Y.-C. Chen、P.-Y. Chen、S. M. Siniscalchi、X. Ma 和 C.-H. Lee，“利用量子卷积神经网络实现特征提取去中心化，用于自动

语音识别”，载于*ICASSP 2021-2021 IEEE* 国际声学、语音与信号处理会议 *(ICASSP)*，第 6523–6527 页，IEEE，2021年。[40] C.-H. H. Yang、J. Qi、S. Y.-C. Chen、Y. Tsao 和 P.-Y. Chen，“当 BERT 遇上量子时序卷积学习：异构计算中的文本分类”，载于*ICASSP 2022-2022 IEEE* 国际声学、语音与信号处理会议 *(ICASSP)*，第 8602–8606 页，IEEE，2022年。[41] R. Di Sipio、J.-H. Huang、S. Y.-C. Chen、S. Mangini 和 M. Worring，“量子自然语言处理的黎明”，载于*ICASSP 2022-2022 IEEE* 国际声学、语音与信号处理会议 *(ICASSP)*，第 8612–8616 页，IEEE，2022年。[42]S. Y.-C. Chen、S. Yoo 和 Y.-L. L. Fang，“量子长短期记忆网络”，载于 *ICASSP 2022-2022 IEEE* 国际声学、语音与信号处理会议 *(ICASSP)*，第 8622–8626 页，IEEE，2022年。[43] S. Y.-C. Chen、D. Fry、A. Deshmukh、V. Rastunkov 和 C. Stefanski，“通过量子循环神经网络实现储备池计算”，*arXiv* 预印本 *arXiv:2211.02612*，2022年。[44] S. Y.-C. Chen、T.-C. Wei、C. Zhang、H. Yu 和 S. Yoo，“用于高能物理数据分析的量子卷积神经网络”，*Physical Review Research*，第 4 卷第 1 期，文章号 013231，2022年。[45] G. Brockman、V. Cheung、L. Pettersson、J. Schneider、J. Schulman、J. Tang 和 W. Zaremba，“OpenAI Gym”，*arXiv* 预印本 *arXiv:1606.01540*，2016。[46] M. Chevalier-Boisvert、L. Willems 和 S. Pal，“面向 OpenAI Gym 的极简网格世界环境”。
https://github.com/maximecb/gym-minigrid，2018。[47] V. Bergholm、J. Izaac、M. Schuld、C. Gogolin、C. Blank、K. McKiernan 和 N. Killoran，“PennyLane：量子-经典混合计算的自动微分”，*arXiv* 预印本 *arXiv:1811.04968*，2018。