Journal of System Simulation

Volume 36 | Issue 1 Article 3

1-20-2024

Strategy Optimization Method of Multi-dimension Projection Based on Deep Reinforcement Learning

Jing An
Joint Logistics College, PLA National Defense University, Beijing 100858, China; Graduate School, PLA National Defense University, Beijing 100091, China; Joint Operations College, PLA National Defense University, Beijing 100091, China, anj21_2000@sina.com

Guangya Si
Joint Operations College, PLA National Defense University, Beijing 100091, China, sgy863@sina.com

Lei Zhang
Joint Logistics College, PLA National Defense University, Beijing 100858, China; Graduate School, PLA National Defense University, Beijing 100091, China; Joint Operations College, PLA National Defense University, Beijing 100091, China

Strategy Optimization Method of Multi-dimension Projection Based on Deep Reinforcement Learning

Abstract

Abstract: Based on the perfect performance of deep reinforcement learning (DRL) in strategy optimization, this paper proposes a strategy optimization method of action taking the multi-dimension projection action as the main research object. The method combines the simulation experiment method with the DRL method. After analyzing the current situation of strategy optimization research, the deep learning framework is selected according to the research problems, and a DRL multi-dimension projection strategy model based on the asynchronous advantage actor-critic (A3C) algorithm is constructed. Through simulation experiments, the interactive learning between the DRL model and the simulation of "out of the loop" is realized, and the optimized multi-dimension projection strategy is obtained. Finally, the effectiveness of the cooperative optimization strategy between the DRL framework and the simulation experiment is verified.

Keywords

deep reinforcement learning (DRL), simulation, strategy optimization, multi-dimension projection, asynchronous advantage actor-critic (A3C) algorithm

Recommended Citation

An Jing, Si Guangya, Zhang Lei. Strategy Optimization Method of Multi-dimension Projection Based on Deep Reinforcement Learning[J]. Journal of System Simulation, 2024, 36(1): 39-49.

第 36 卷第 1 期 系统仿真学报© Vol. 36 No. 1
2024 年 1 月 Journal of System Simulation Jan. 2024

基于深度强化学习的立体投送策略优化方法研究

安靖[1,2,3]，司光亚[3*]，张雷[1,2,3]
(1. 国防大学 联合勤务学院，北京 100858；2. 国防大学 研究生院，北京 100091；3. 国防大学 联合作战学院，北京 100091)

摘要：基于深度强化学习算法在策略优化问题中的良好表现，以立体投送作战行动为主要研究对象，提出了一种深度强化学习框架与仿真推演实验协同的作战行动策略优化方法。在分析策略优化研究现状的基础上，根据研究问题对深度学习框架进行了分析比较，构建了基于A3C算法的深度强化学习立体投送策略模型，并通过仿真推演和分布式计算，实现深度强化学习模型与“人不在回路”仿真推演的交互学习，获得优化后的立体投送策略，验证了深度强化学习框架与仿真推演实验协同优化策略的有效性。

关键词：深度强化学习；仿真推演；策略优化；立体投送；A3C 算法
中图分类号：TP391.9 文献标志码：A 文章编号：1004-731X(2024)01-0039-11
DOI: 10.16182/j.issn1004731x.joss.22-0886
引用格式: 安靖, 司光亚, 张雷. 基于深度强化学习的立体投送策略优化方法研究[J]. 系统仿真学报, 2024, 36(1): 39-49.
Reference format: An Jing, Si Guangya, Zhang Lei. Strategy Optimization Method of Multi-dimension Projection Based on Deep Reinforcement Learning[J]. Journal of System Simulation, 2024, 36(1): 39-49.

Strategy Optimization Method of Multi-dimension Projection Based on Deep Reinforcement Learning

An Jing[1,2,3], *Si Guangya*[3*], *Zhang Lei*[1,2,3]
(1. Joint Logistics College, PLA National Defense University, Beijing 100858, China; 2. Graduate School, PLA National Defense University, Beijing 100091, China;3. Joint Operations College, PLA National Defense University, Beijing 100091, China)

Abstract: Based on the perfect performance of deep reinforcement learning (DRL) in strategy optimization, this paper proposes a strategy optimization method of action taking the multi-dimension projection action as the main research object. The method *combines the simulation experiment method with the DRL method.* After analyzing the current situation of strategy optimization research, the deep learning framework is selected according to the research problems, and *a DRL multi-dimension projection strategy model based on the asynchronous advantage actor-critic (A3C) algorithm is constructed.* Through simulation experiments, *the interactive learning between the DRL model and the simulation of "out of the loop" is realized, and the optimized multi-dimension projection strategy is obtained.* Finally, the effectiveness of the cooperative optimization strategy between the DRL framework and the simulation experiment is verified.

Keywords: deep reinforcement learning (DRL); simulation; strategy optimization; multi-dimension projection; asynchronous advantage actor-critic (A3C) algorithm

收稿日期：2022-08-02 修回日期：2022-09-27
第一作者：安靖(1981-)，女，副教授，博士，研究方向为军事运筹学、战争设计系统工程。E-mail：anj21_2000@sina.com
通讯作者：司光亚(1967-)，男，正高级工程师，博导，博士，研究方向为战争模拟系统、军事运筹学。E-mail：sgy863@sina.com

0 引言

战争复杂系统的不确定性、对抗性、涌现性日趋显著，给不完全信息下作战策略的分析与优化带来了前所未有的技术挑战。主要体现在：①决策行为与效益之间难以形成有效映射，无法采用传统的理论分析和解析计算；②宏观决策的效益需要较长时间才能体现，收益反馈的滞后影响策略探寻的全局最优[1-5]。

由于作战行动决策过程可以被视为马尔可夫决策过程(Markov decision process, MDP)，即当前阶段的作战行动会影响下一阶段的战场态势，而下一阶段的战场态势又将影响行动策略。于是现阶段不少研究人员开始利用适合处理马尔可夫决策过程的深度强化学习算法对作战行动的策略优化问题进行研究。例如姚桐等[6]对深度强化学习框架应用在作战任务规划上的可能性进行了探讨。吴昭欣等[7]研究了基于深度强化学习技术的智能仿真平台。于博文等[8]研究了分层强化学习在联合作战仿真作战决策中的应用。石鼎等[9]研究了海战环境下利用强化学习技术协作作战的方法。上述研究有的是从某一具体作战装备出发，仅研究具体装备运用，未能较好考虑全局性和复杂性；有的是概念算法上的研究，未能有效结合典型作战场景，缺乏军事现实意义。

为解决上述问题，本文针对立体投送作战行动，提出一种深度强化学习框架与仿真推演实验协同优化策略方法。该方法采用博弈策略的评价机制和学习型策略演进机制，构建了基于A3C (asynchronous advantage actor-critic)算法的深度强化学习框架，并依托分布式计算环境进行训练，通过深度强化学习模型与“人不在回路”仿真推演的分布式交互，实现作战策略的自主进化与能力提升，最终获得立体投送的决策模型，得出全局近似最优的作战策略。

1 深度强化学习框架的选择

强化学习(reinforcement learning, RL)是机器学习中的一个重要分支，用于描述和解决智能体(agent)在与环境(environment)的交互过程中，通过学习策略达成回报最大化或实现特定目标的问题。按给定条件，强化学习可分为有模型的强化学习(model-based RL)和无模型的强化学习(model-free RL)；按求解强化学习问题所使用的算法，可分为同策学习(on policy)和异策学习(off policy)两类；按照优化的中心可以分为基于策略方法(policy-based)和基于价值方法(value-based)两类，此外还有结合两者优点的评价–执行(critic-actor)方法；按照更新频率可以分为单步更新(temporal-difference update)和回合更新(episodic update)。另外，深度学习模型可以应用于强化学习，用来拟合价值函数，这样就形成了深度强化学习(deep reinforcement learning, DQN)。上述方法各具特点，需要根据不同的研究问题进行选择。下文将结合立体投送策略优化这一实际作战问题和各种强化学习方法的特点，构建符合该问题研究需求的深度强化学习框架。

1.1 有模型强化学习与无模型强化学习的选择

模型是指强化学习中的环境，通常认为有模型的强化学习中，转移函数P是已知的，而无模型的强化学习中，转移函数P是未知的。通常来说，有模型方法的学习效率会更高，但是在实际应用中，较少能够获得精确简洁的拟合模型，因此无模型方法在实际应用中更为普遍。

在本文研究的过程中，尽管可以把整个立体投送仿真过程视为一个模型，但要获取这个拟合模型的难度较高，且计算速度要远远慢于强化学习的学习速度。因此，本文选择无模型强化学习方法。

1.2 同策学习与异策学习的选择

同策学习(on-policy)即边决策边学习，生成的样本策略和学习策略相同，学习者和决策者相同。典型的算法是SARAS(state-action-reward-state-

DOI: 10.16182/j.issn1004731x.joss.22-0886

action)算法、Epsilon-Greedy 算法等。该方法的优点是思路直接，学习速度快，劣势是不一定能找到全局最优策略。

异策学习(off-policy)是指通过之前的历史策略进行学习，学习者和决策者不需要相同。典型算法为Q-learning 算法。该方法意在探索，优势是更为强大和通用，劣势是曲折、收敛慢。

考虑到异策学习可以一边优化策略，一边探索新策略收集数据，同时可以开启多个模拟器驱动并行学习任务，满足立体投送策略优化中仿真速度较慢，需分布并行运算的需求，因此后续将采用异策学习方法。

1.3 回合更新与单步更新的选择

一个回合通常包含很多个时间步，在一个完整的回合结束后，根据奖励函数获得的奖励总和来进行策略修正称之为回合更新方法。这种方法逻辑简单，缺点在于由于回合结束前无法进行修正，这意味着即使知道了目前不是最优的行动，也不得不等到回合结束后才能修正。

相反，每次行动后根据预测值进行修正的方法称之为单步更新方法，如时序差分法(temporal difference learning)。该方法不需要等到回合全部结束，根据当前状态和策略的估计值即可展开修正。由于该方法是根据当前的估计值，而不一定是真正的奖励值进行修正，所以修正的精度有所不足。

本文后续将采用模拟仿真和深度强化学习框架进行分布式交互学习，因此，选择介于两者之间的，能够兼顾实际奖励值和一定多步预测能力的multi-step learning方法构建强化学习框架。

1.4 基于价值强化学习方法与基于策略强化学习方法的选择

在强化学习中，价值函数是根据对未来奖励值的预测来衡量状态良好程度的函数，有两种常用的表示方式。

一种是状态值函数 $V^{\pi}(s)$，表示在初始状态 s 时采取策略 π 所获得的期望总价值，表示为

$$V^{\pi}(s)=E_{\pi}\left[R_t|s_t=s\right]=E_{\pi}\left[\sum_{k=0}^{\infty}\gamma^k r_{t+k+1}|s_t=s\right] \tag{1}$$

另一种是动作值函数 $Q^{\pi}(s,a)$，表示在状态 s 下采取动作 a，后续动作遵从策略 π 所获得的期望总价值，表示为

$$Q^{\pi}(s,a)=E_{\pi}\left[R_t|s_t=s,a_t=a\right]=E_{\pi}\left[\sum_{k=0}^{\infty}\gamma^k r_{t+k+1}|s_t=s,a_t=a\right] \tag{2}$$

策略 π 本质是一个概率分布，反应了在某一状态下选择动作的概率：

$$\pi_{\theta}(s,a)=P(a|s,\theta) \tag{3}$$

式中：θ 为策略函数中的参数。

在无模型的强化学习中，通常很难直接获得精确的价值函数，因此需要用一个价值近似函数(value function approximation)来代替。

基于价值的强化学习旨在学习价值近似函数，令其符合实际奖励，以便生成最佳策略；基于策略的强化学习旨在使用参数化函数直接学习策略函数，不需要计算价值函数。

两种方法各有优缺点，基于价值的方法必须选择使价值函数最大的动作，如果动作空间非常高维或连续，成本就会很高，而基于策略的方法是通过直接调整策略的参数来运行的，不需要进行最大化计算。因此基于策略的方法更稳定，收敛性更好，且既可以学习确定性策略，也可以学习随机策略。但是与基于价值的方法相比，基于策略的方法速度更慢，方差更大，且通常收敛于局部最优而不是全局最优。

为了结合两者的优势，目前比较流行的做法是使用执行–评价算法(actor-critic algorithm)，这是一种结合策略梯度和时序差分学习的强化学习方法。其中 actor 是指策略函数 $\pi_{\theta}(s,a)$，即学习一个策略来得到尽量高的奖励。critic 是指价值函数

$V^{\pi}(s)$，对当前策略的价值函数进行估计。

借助于价值函数，actor-critic 算法可以单步或者多步更新参数，不需要等到整个回合结束才进行更新。在每步更新中，分别进行策略函数和价值函数的学习，学习开始时 actor 随机执行策略，critic 随机评价。通过不断的学习，critic 的评价越来越准，actor 执行的策略也越来越好。

在本文的后续研究中，将采用 actor-critic 算法构建强化学习框架。

1.5 深度强化学习

由于状态空间(state spaces)和动作空间(action spaces)的爆炸，传统的强化学习方法会受限于维数灾难(curse of dimensionality)。因此可以利用神经网络的万能逼近特性来拟合价值近似函数和策略函数，如果采用 LSTM(long short term memory)或者 CNN(convolutional neural network)等深度神经网络，则可以获得深度强化学习框架，较为典型的算法有 DDPG，PPO，A2C，A3C 等。

深度确定策略梯度(deep deterministic policy gradient, DDPG)是利用神经网络输出一个确定性的动作，可用于连续动作空间的算法。同时也是一种单步更新的策略网络，通过 off-policy 的方式来训练一个确定性策略。因为策略是确定的，在开始阶段，它可能无法尝试足够多的 action 来找到有用的学习信号。为了让 DDPG 的策略更好地探索，训练的时候需要给 action 加入噪音。DDPG 的原作者推荐使用时间相关的 OU noise，但是使用这种有很多超参数的方法去探索环境，训练慢且不太稳定[10]。

近端策略搜索(proximal policy optimization, PPO)算法是一种 off-policy 的策略梯度算法，此类算法通常对步长敏感，难以选择合适的步长，在训练过程中新旧策略的的变化差异如果过大则不利于学习。PPO 提出了新的目标函数可以再增加一个训练步骤实现小批量的更新，解决了策略梯度算法中步长难以确定的问题。其优点是训练稳定，调参简单，鲁棒性强[11]。

优势动作评价(advantage actor-critic, A2C)算法，利用优势函数代替 critic 网络中的原始回报，可以作为衡量选取动作和所有动作平均期望好坏的指标。式(4)表示优势函数 $A^{\pi}(s,a)$ 是状态函数和状态–动作函数之差，如果该函数大于零，说明该动作比平均动作好，反之则说明比平均动作差。

$$A^{\pi}(s,a)=Q^{\pi}(s,a)-V^{\pi}(s) \tag{4}$$

A2C 算法采用同步更新的方式进行学习，学习效率较低。为了提高学习效率，又提出了异步优势动作评价(asynchronous advantage actor-critic, A3C)算法，asynchronous 指开启多个 actor 在环境中探索，并异步更新。所谓异步的方法是指数据并非同时产生，可以利用多个线程，每个线程相当于一个智能体在随机探索，多个智能体共同探索，并行计算策略梯度，维持一个总的更新量[12]。

由于 A3C 算法的学习效率更高，因此在本文的后续研究中，将采用 A3C 算法作为强化学习框架。

2 基于深度强化学习的立体投送策略模型构建

2.1 作战行动构想

立体投送作战行动的概念图如图 1 所示。

(1) 作战目的

作战行动中，红方的作战目的是尽快利用火力压制蓝方，并综合利用平面、空中、超越等多种投送方式将兵力投送至蓝方陆地。

整个过程中追求战损尽可能小，投送成功率尽可能高，同时上陆兵力能够在一定时间内完成固守任务。

(2) 红方兵力

红方兵力可大致分为作战力量、支援力量以及投送力量三大类。

1) 作战力量：主要负责直接毁伤对方有生力量，空基对空力量主要指空优战斗机，主要负责对空任务。空基对地/海力量包括武装直升机、歼

DOI: 10.16182/j.issn1004731x.joss.22-0886

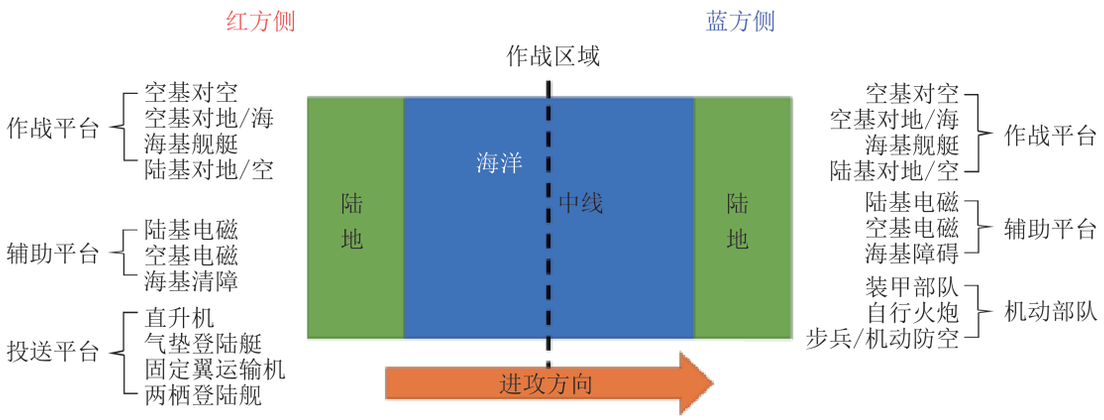


图1 立体投送作战行动概念图

Fig. 1 Concept map of multi-dimension projection

轰机、轰炸机等，主要负责对地对海轰炸任务。海基力量包括护卫舰、驱逐舰等，兼顾防空和对岸炮击任务。陆基对空/对地力量包括巡航导弹、远程火力、防空导弹阵地等，主要负责对地重点打击和近岸防空。

2) 支援力量：主要包括陆基电磁，空基电磁和海基清障三大类，陆基电磁主要为固定雷达站，空基电磁包括电子侦察机、预警机、电子战干扰机等，海基清障主要为扫雷舰和工兵舰。

3) 投送力量：主要包括多种可执行登陆作战的装备平台，如两栖登陆舰、气垫艇、运输直升机、固定翼运输机等，可以用于投送上陆力量。

(3) 蓝方兵力

蓝方兵力与红方大致相同，但由于其主要任务为拒止和反登陆作战，因此作战力量中，一是没有舰艇部队，但是在重点港口处部署了大量的海基障碍，如水雷和其他登陆障碍物。二是没有投送力量，但是在陆地上部署了大量机动部队，如步兵、自行火炮、坦克、机动防空阵地等，主要负责对抗红方上陆部队。

2.2 状态空间

上述行动构想涉及因素较多，如果为每一个相关因素设计一套状态参数进行表示，则会陷入维度爆炸问题中。因此需要对相关因素进行重要度排序，忽略部分次要因素。

本文结合定性分析和预实验定量分析的方法，对涉及的各种因素进行了敏感性分析，同时为便于后续深度学习框架的搭建和学习，对状态参数进行离散化设置，最终筛选出了本文研究的状态空间。

本文设想的作战想定中关注 69 个作战单元、3 支投送编队(除运输直升机)，2 支舰载运输直升机编队，1 支岸基运输直升机编队。

由于本文主要聚焦投送方式、投送时序，投送能力等投送策略的研究，因此对作战单元的状态参数不做过多关注，仅关注其数量，忽略其位置、速度等属性，仅考虑存活情况，即仅考虑在某一时间步内，某型装备的剩余数量。例如在时间步 10 中，红方的 A 型武装直升机剩余 13 架，则该变量的取值为 13，显然这是一个离散变量。

对于本文关心的投送力量的状态参数，则考虑更为充分一些，但是投送单元也包含有大量的状态参数，以运输直升机为例，可供设置的参数包括起飞位置、目标位置、突防线路(沿途路径点)、飞行高度、飞行速度、电磁干扰、电磁静默、起飞时间等。为了判断哪些状态参数对仿真结果产生的影响更大，首先设计了一组仿真预实验对其进行了分析。

预实验首先设计基准想定 1 个，然后在想定中改变起飞位置、目标位置和突防路线，设计了对比想定 2 组，同时尽量保证每条突防路线上的

蓝方防空力量基本一致。每个想定运行 50 次，统计最终的红方平均投送损失情况。

在基准想定的基础上，分别改变飞行高度、飞行速度等变量，每次改变变量后设置了 1~2 组对比想定。同样运行对比想定 50 次，统计每个想定最终的红方平均投送损失情况，如图 2 所示。

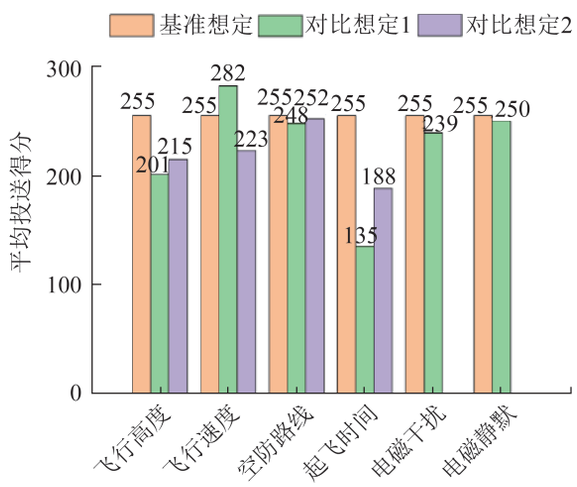


图 2　针对运输直升机投送平台的仿真预实验结果

Fig. 2　Simulation experiment results for transport helicopter

可以看出，飞行速度、飞行高度和起飞时间这 3 个变量对仿真结果的影响较大，即敏感性高。而突防路线、电磁干扰、电磁静默则影响较小。所以最终选择飞行速度、飞行高度和起飞时间作为运输直升机的状态空间。

针对其他投送平台，如固定翼运输机、两栖登陆舰等，同样执行类似的筛选过程。

经过上述定性分析和定量分析后，决定考虑 3 种作战力量在每一个时间步内的剩余数量、速度、高度(仅直升机)和出航状态 4 个变量为状态参数，见表 1，均为离散参数。剩余数量取当前时间步内的数量；速度取值分为[蠕行速度，巡航速度，突击速度，全速]4 种；高度特指运输直升机的离地高度，取值分为[100 m，200 m，400 m，1 000 m] 4 种；出航状态表示该投送平台是否出动，用 2 个离散变量[0，1]来表示，0 表示尚未出动，1 表示已在投送途中。

表 1　每一个时间步内的状态空间

Table 1　State space in each time step

序号	作战力量	状态参数
1	作战单元(除投送编队外)69 个	剩余数量
2	投送编队(除运输直升机外)3 支	剩余数量，当前速度，出航状态
3	运输直升机编队 3 支	剩余数量，当前速度，出航状态，当前高度

2.3　动作决策空间

由于本文聚焦投送策略优化，因此忽略作战单元的作战策略。对于投送力量，在每一个时间步内，均存在两种状态，即已行动和未行动。

为保证动作空间参数的离散化，做如下假设：一是所有投送力量均只能执行一次投送任务，不存在返航再次执行投送任务的情况；二是不存在投送力量编组的拆分问题。

基于上述假设，在已行动和未行动两种状态的转换过程中，可以选择执行的动作空间包括 3 种：①上一时间步未行动，该时间步决定是否行动；②上一时间步未行动，本时间步决定行动，决定行动的同时决定出动的编队规模、速度、高度等；③上一时间已经行动，本时间步继续行动，可以调整该编队的速度、高度。形成的动作决策空间如表 2 所示。

表 2　每一个时间步内的动作决策空间

Table 2　Action space in each time step

序号	作战力量	动作空间
1	未行动的投送力量	[0,1] (0 表示维持等待，1 表示出航)
2	未行动的投送力量，但已决定在本时间步内行动	规模数量，速度，高度(直升机)
3	已行动的投送力量	速度，高度(直升机)

2.4　奖励函数

奖励函数是从状态得到奖励的回归公式。合理的设计是令强化学习取得良好效果的关键因素，如果奖励函数的设计不够合理，则强化学习框架

DOI: 10.16182/j.issn1004731x.joss.22-0886

可能会寻找到一些无意义的策略。

本文重点关注投送策略，即上陆兵力规模需求、编成结构、编队队形、空海比例、投送方式、投送时机等。因此，本文奖励函数的设计思路主要聚焦红方作战目的的达成，将兵力到位率，战斗损耗，固守时间等多个变量整合成一个标量，在仿真想定推演的每一个时间步内进行计算。另外，也要对额外投入的投送力量进行惩罚。

基于上述分析，给出如下奖励值的定义。

定义1：将任一时刻t的奖励值记为R_t，则

$$R_t=R_{t-1}+(R_1+R_2+R_3+R_4+R_5\,|s=s_t)$$

式中：

R_{t-1}为上一个时刻的奖励值。

R_1为红方战损，即某一仿真时间步内红方损失的兵力，该值为负数，由各种具体的装备分值累加得到。例如在某一时间步内，红方损失战斗机3架，运输直升机2架。战斗机分值6分，运输直升机5分，则本时间步内的$R_1=-6\times3-5\times2=-28$分。

R_2为红方投送兵力上陆的一次性奖励，为正数，在被投送部队登陆成功的时间步内给出，如果投送平台在投送过程中被击毁，则无法获得该奖励。例如在某一时间步内，3架运输直升机成功完成步兵投送任务，运输直升机投送成功奖励为5分，则本时间步内的$R_2=5\times3=15$分。

R_3为红方投送兵力存活的持续性奖励，为正数，在投送成功之后的每一个时间步内检查上陆兵力是否存活，如存活，则在此时间步内给予奖励。例如在某一时间步内，蓝方阵地上存在40个步兵小队和10个装甲单位，每个步兵小队0.2分的存活奖励，装甲单位1分，则$R_3=0.2\times40+1\times10=18$分。

R_4为红方完成占领任务的奖励，在最后一个仿真时间步结束后，检查登陆阵地，如果阵地内存在红方单位而不存在蓝方单位，视为占领成功，此时给予正数奖励。共10个关键阵地，完成占领每个阵地给予10分奖励。

R_5为红方的投送平台数量奖励，是一个负数，该值与当前时间步内已行动的投送力量呈负相关关系，即投入投送的力量越多，该值越小，意味着罚分更多。这是因为，虽然投送平台越多，上陆兵力也会越多，但这与所希望的，用最少运力达成作战目的不吻合。因此奖励函数设置中对平台数量策略予以负反馈，才能令最终学习到的结果符合预期要求。以运输直升机为例，其罚分函数为$-1.5^{(x-40)}$，这是一个指数函数，式中的40为期望的出动上限，当运输直升机的出动数量超过40架后，罚分就会迅速升高。

按照上述定义，策略优化的目的是令最后一个时间步结束时，奖励值最高。本文采用固定仿真时间的方式来触发仿真结束，总仿真时间为6 h，每5 min作为一个时间步，总计72个时间步。在每个时间步的开始阶段仿真环境与深度强化学习环境进行交互，在每一个时间步结束时输出当前奖励值，并在最后一个时间步结束后输出最终的奖励值。

2.5 价值函数和策略函数

基于第一节的分析，本文采用无模型off-policy，multi-step Learning的A3C方法构建深度学习框架，A3C方法是actor-critic的一种，需要利用两组神经网络critic网络和actor网络来评估价值函数和策略函数，两组神经网络都需要进行离线训练。如图3所示。

对于critic网络，输入层为上一个时间步内的动作指令(如果是第一个时间步，则上一个时间步的动作指令视为0)，以及当前时间步的状态参数，经过归一化处理后与一个双曲正切激活函数连接，再连接到一个宽度为256的长短时记忆神经网络(LSTM)中。LSTM神经网络由于其擅长处理时序问题，因此非常适用于作战序列类的MDP问题。LSTM的输出一方面作为它自身的输入，另外与一个多层感知机连接(multi-layer perceptron, MLP)，利用多层感知机擅长拟合非线性关系的特

点，该多层感知机具有两个隐藏层，每层有256个神经元，每一层内采用ELUs (exponential linear units)作为激活函数以映射非线性关系，最后用一个线性层来输出当前行动的价值，即Q值。ELUs函数为

$$f(x)=\begin{cases}a(e^x-1), & x\leqslant 0\\ x, & x>0\end{cases} \tag{5}$$

式中：α为超参数，可以通过调节这一参数调整收敛速度。

对于actor网络，其输入层和critic网络相同，通过一个线性层连接到具有256的神经元的隐藏层中，该线性层的输出进行归一化处理后利用双曲正切函数的有界性进行激活，然后送入3层MLP里，该MLP的结构与critic网络相同，最后通过一个线性层连接到输出层。输出层包括每个动作决策的平均值以及对应高斯分布的标准差，实际反馈到模型中执行的时候，只需要执行动作决策的平均值即可。

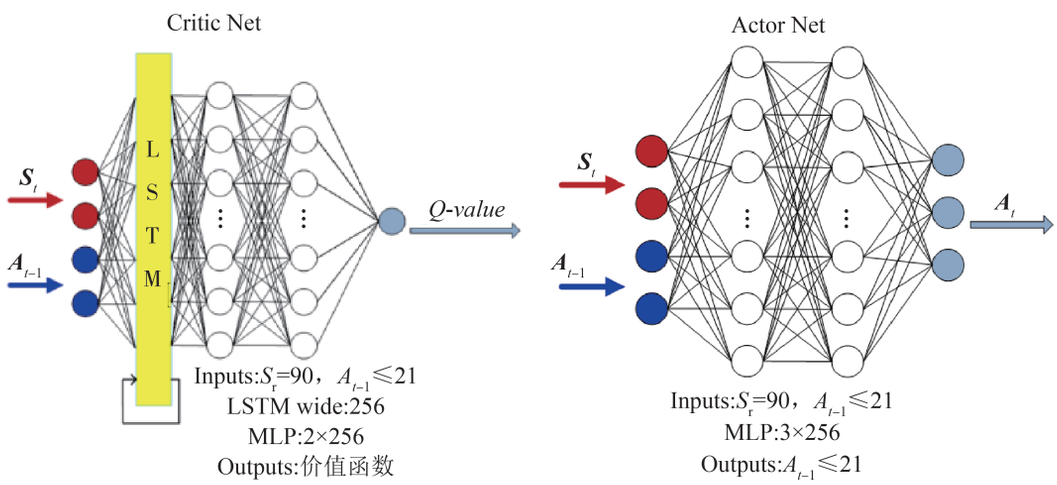


图3 Critical-actor网络结构

Fig. 3 Critical-actor network structure

3 模型训练与实验结果分析

模型训练的仿真实验环境如图4所示，主要包括3部分：

(1) 可视化的仿真推演环境。主要依托CMO仿真推演平台，采用“人不在回路”的模式，仿真实体按照预设好的作战条令和想定运行脚本实施推演。

(2) 深度强化学习环境。基于PyTorch框架实现，用于智能体决策的学习和实施。

(3) 进行交互的数据接口。仿真实验环境和深度强化学习环境，在每一个时间步中通过该数据接口，以离线方式进行交互。

3.1 交互学习的模型训练流程

由于仿真模拟的运行速度要远慢于强化学习框架的学习速度，为了缓解训练数据相对较少的问题，采用A3C算法进行异步分布式训练，用以提高训练速度。每一个运行于张量进程中的学习模型实例对应10个critic-actor网络，每一个critic-actor网络都驱动了一个独立的仿真模型，总共并行运行了100个仿真模型，每个仿真模型均使用30倍模拟加速，运行完整的6 h模拟作战时间大约需要0.2 h左右的现实时间。

训练过程是由仿真模型和强化学习框架相互交互实现的[13-14]，在每一个时间步上，强化学习框架给出的指令加载至仿真模型上，通过仿真模型给出计算结果。在训练的初始阶段，动作指令集按照对角高斯分布的方式加载至仿真模型上，这是为了在学习阶段尽可能多的探索较好的指令序列。

DOI: 10.16182/j.issn1004731x.joss.22-0886

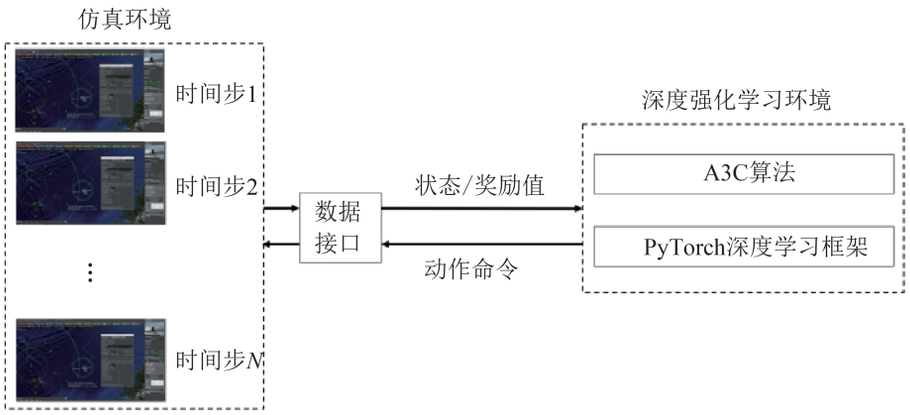


图4 仿真实验环境
Fig. 4 Simulation experiment environment

学习过程在72个时间步上重复进行，学习率设为0.001，折扣率设为0.95，batch size设为256。

3.2 训练结果分析

按照上述方法进行交互学习训练，学习次数和所获奖励值的变化趋势如图5所示。

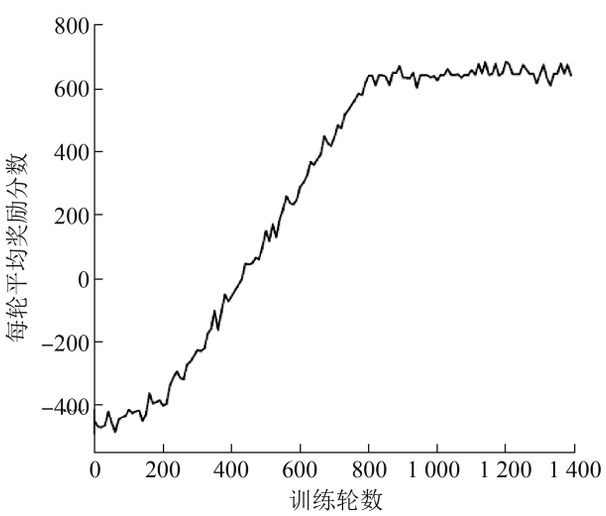


图5 平均奖励值随训练轮数的变化
Fig. 5 Average reward value varied with number of training rounds

由图5可以看出，在800轮训练(运行仿真模型80 000个，现实时间约40 h)后，基于深度强化学习的立体投送策略模型逐渐收敛，平均奖励收敛至635分左右。分析可知，在初始学习阶段，即前200轮训练时，由于此时按照预定的初始化策略去探索策略空间，采取的策略较为低效[15]，例如投送平台可能会直接冲入敌方防空火力网中被击毁，因此获取的奖励分数也较低。但很快模型就会寻找到较为高效的策略，平均奖励也逐渐走高，至800轮训练后逐渐收敛。

3.3 优化策略分析

利用最终优化后的策略再次运行仿真推演，获得每个时间步内累计奖励值的曲线如图6所示，对应的仿真系统截图如表3所示。

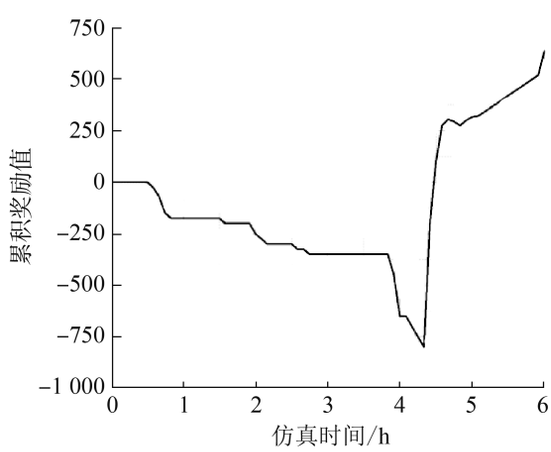
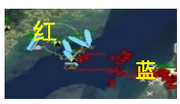


图6 每个时间步内累计奖励值变化
Fig. 6 Cumulative reward value in each time step

可以看出，在优化策略中，仿真开始的前3个小时内，为避免投送平台遭蓝方优势火力的打击，红方未出动投送平台，而是由作战力量执行火力打击任务，由于红与蓝争夺制空权所带来的战损较大，红方的奖励值均为负数。

表3 仿真推演过程示例
Table 3 Simulation process

仿真时间/h	行动	仿真系统截图
0	战斗力量升空	
2.0	两栖输送编队航渡	
4.0	空中输送编队起飞	
4.5	海上编队卸载 空中编队空机降	
5.0	上岛力量集结	

在4 h左右，红方的奖励值继续减少，这是由于为确保投送任务顺利执行，红方深入蓝方阵地执行对地打击，带来了大量战损。

在2 h和4 h左右，红方海空投送平台分别开始行动，并于4.5 h左右几乎同时抵达蓝方阵地，这一轮成功的登陆带来了投送成功奖励R_2值的急速增长。

随后登陆成功的红部队与蓝地面部队进行了小规模交火，在红方优势空军的配合下很快清理了蓝军地面部队，5 h后的奖励值由于存活奖励R_3的存在而稳步上升，并且在最后一个时间步后拿到了大量成功占领的奖励R_4。

整体投送数量受到R_5负反馈的约束，最终经过强化学习后选择的出动规模与人工预期值接近(具体数值略)。

3.4 算法比较分析

采用PPO方法和A2C方法训练模型，与本文构建的A3C方法进行对比实验，分别记录训练轮次和平均奖励值的关系，得到结果如图7所示。

由于PPO方法和A2C方法采用同步更新的策略，因此比异步更新的A3C方法更新耗时更长，且最终获得的优化结果也略低一些。PPO方法的收敛速度与A2C差别不大，同时最终收敛以后获得的奖励值波动也更大，这与PPO算法中新策略和旧策略的比例这一超参数有关。如果新策略所占比例较大，则收敛会加快，但由于其倾向于选择收益更大，同时风险也较大的新策略，所以最终收敛值的波动也较大。如果下调这一比例，则会令其收敛速度进一步变慢。

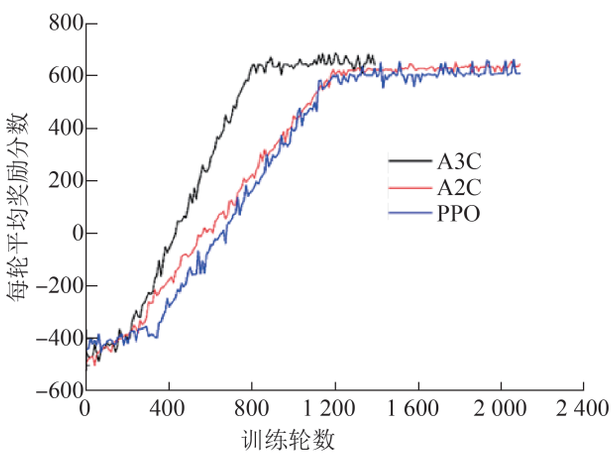


图7 算法比较
Fig. 7 Algorithm comparison

由此可见，A3C方法在收敛速度和最终优化结果等方面，均具有一定的优势。

4 结论

本文以立体投送作战行动为主要研究对象，提出了一种深度强化学习框架与仿真推演实验协同策略优化方法。该方法中，构建了基于A3C算法的深度强化学习框架，设计了合理的奖励函数，利用LSTM和MLP神经网络构建了Critic网络用以计算价值函数，利用多层MLP神经网络构建生成行动指令的Actor网络，通过深度强化学习模型与“人不在回路”仿真推演的交互，最终获得优化后的立体投送策略。优化后的投送策略内容合理，符合预期，显著提高了最终奖励得分，验证了深度强化学习框架与仿真模型协同优化决策的有效性。

该领域目前仍处于起步探索阶段，有以下问

DOI: 10.16182/j.issn1004731x.joss.22-0886

题需要做进一步的研究：

(1) 深度神经网络的可解释性。

(2) 为了简化计算过程，所设计的想定、动作空间、状态空间等均经过了大幅度的简化，与真实战场相比，复杂度仍较低。

(3) 所获模型尽管在某一具体想定仿真模型中表现优异，但不一定适用于同类但不完全相同的想定仿真模型，即所获作战策略的泛化性不足。未来拟采用迁移学习、元学习等方法予以改进。

(4) 无模型的强化学习框架尽管不需要提前训练有标签的数据进行学习，但训练效率低，可解释性较差，而且无法充分利用专家经验。未来拟在有模型强化学习领域以及与决策树等方法结合的领域进行进一步探索。

参考文献：

[1] 杨峰, 李群, 王维平, 等. 基于仿真的探索性评估方法论[J]. 系统仿真学报, 2003(11): 1561-1564.
Yang Feng, Li Qun, Wang Weiping, et al. Simulation Based Exploratory Evaluation Methodology[J]. Journal of System Simulation, 2003(11): 1561-1564.

[2] 杨镜宇, 司光亚, 胡晓峰. 信息化战争体系对抗探索性仿真分析方法研究[J]. 系统仿真学报, 2005(6): 1469-1472, 1496.
Yang Jingyu, Si Guangya, Hu Xiaofeng. Study on Simulation-based Exploratory Analysis Method of Information Warfare System of System (SoS) Encounter [J]. Journal of System Simulation, 2005(6): 1469-1472, 1496.

[3] Payne D. Commond Decision Model Technologey Assess-ment: ADAY34926, 19-5[R]. [S.l.]: [s.n.]: 16-27.

[4] 李斌, 刘苏洋, 李春洪, 等. 探索性仿真实验仿真想定空间筛选[J]. 火力与指挥控制, 2013, 38(5): 152-156.
Li Bin, Liu Suyang, Li Chunhong, et al. Screening the Simulation Scenario Space in Exploratory Simulation Experiment [J]. Fire Control & Command Control, 2013, 38(5): 152-156.

[5] 喻飞飞, 赵志敏, 包俊. 探索性仿真分析框架下的实验点设计方法[J]. 指挥控制与仿真, 2014, 36(2): 80-84.
Yu Feifei, Zhao Zhimin, Bao Jun. Experiment Points Design Methods Under the Framework of Exploratory Simulation Analysis[J]. Command Control & Simulation, 2014, 36(2): 80-84.

[6] 姚桐, 王越, 董岩, 等. 深度强化学习在作战任务规划中的应用[J]. 飞航导弹, 2020(4): 16-21.

[7] 吴昭欣, 李辉, 王壮, 等. 基于深度强化学习的智能仿真平台设计[J]. 战术导弹技术, 2020(4): 193-200.
Wu Zhaoxin, Li Hui, Wang Zhuang, et al. The Design of Intelligence Simulation Platform Based on DRL[J]. Tactical Missile Technology, 2020(4): 193-200.

[8] 于博文, 吕明, 张捷. 基于分层强化学习的联合作战仿真作战决策算法[J]. 火力与指挥控制, 2021, 46(10): 140-146.
Yu Bowen, Lü Ming, Zhang Jie. Joint Operation Simulation Decision-making Algorithm Based on Hierarchical Reinforcement Learning[J]. Fire Control & Command Control, 2021, 46(10): 140-146.

[9] 石鼎, 燕雪峰, 宫丽娜, 等. 强化学习驱动的海战场多智能体协同作战仿真算法[J]. 系统仿真学报, 2023, 35(4): 786-796.
Shi Ding, Yan Xuefeng, Gong Lina, et al. Multi-agent Cooperative Combat Simulation in Naval Battlefield with Reinforcement Learning[J]. Journal of System Simulation, 2023, 35(4): 786-796.

[10] Lillicrap T P, Hunt J J, Pritzel A, et al. Continuous Control with Deep Reinforcement Learning[J]. (2019-07-05) [2022-06-26]. https://arxiv.org/abs/1509.02971.

[11] Schulman J, Wolski F, Dhariwal P, et al. Proximal Policy Optimization Algorithms[EB/OL]. (2017-08-28) [2022-06-26]. https://arxiv.org/abs/1707.06347.

[12] Mnih V, Adrià Puigdomènech Badia, Mirza M, et al. Asynchronous Methods for Deep Reinforcement Learning[C]//Proceedings of the 33rd International Conference on International Conference on Machine Learning. Cambridge: JMLR, 2016: 1928-1937.

[13] 曹雷. 基于深度强化学习的智能博弈对抗关键技术[J]. 指挥信息系统与技术, 2019, 10(5): 1-7.
Cao Lei. Key Technologies of Intelligent Game Confrontation Based on Deep Reinforcement Learning [J]. Command Information System and Technology, 2019, 10(5): 1-7.

[14] 孙长银, 穆朝絜. 多智能体深度强化学习的若干关键科学问题[J]. 自动化学报, 2020, 46(7): 1301-1312.
Sun Changyin, Mu Chaoxu. Important Scientific Problems of Multi-agent Deep Reinforcement Learning [J]. Acta Automatica Sinica, 2020, 46(7): 1301-1312.

[15] 孙彧, 李清伟, 徐志雄, 等. 基于多智能体深度强化学习的空战博弈对抗策略训练模型[J]. 指挥信息系统与技术, 2021, 12(2): 16-20.
Sun Yu, Li Qingwei, Xu Zhixiong, et al. Game Confrontation Strategy Training Model for Air Combat Based on Multi-agent Deep Reinforcement Learning[J]. Command Information System and Technology, 2021, 12 (2): 16-20.