

海空集群对抗深度强化学习算法研究平台设计

刘宝宏

(湖南苍树航天科技有限公司, 湖南 长沙 410073)

✉ liubh_vip@163.com



摘要:针对当前海空集群对抗智能指挥控制决策研究缺乏有效的算法研究平台的问题,设计了深度强化学习算法研究平台。首先分析了应用需求,其次对平台组成、硬件架构进行了设计描述,最后针对想定设计、算法训练和算法评价进行了子流程设计。平台设计采用仿真内核和显示模块分离、并行与分布式运行的模式,支持深度强化学习算法的快速训练和决策控制效果的验证评价,避免直接进行海空装备实物实验耗费大量时间和面临未知风险,有效提高了算法研发效率。

关键词:集群对抗;深度强化学习;平台设计;并行与分布式运行

中图分类号:TP183 **文献标志码:**A

Design of Deep Reinforcement Learning Algorithm Research Platform for Sea and Air Swarm Confrontation

LIU Baohong

(Hunan Cangshu Aerospace Technology Co., Ltd., Changsha 410073, China)

✉ liubh_vip@163.com

Abstract: This paper proposes to design a deep reinforcement learning algorithm research platform to address the lack of effective algorithm research platforms for intelligent command and control decision-making in sea and air swarm confrontation. Firstly, the application requirements are analyzed and then the platform composition and hardware architecture are introduced. Finally, a sub-process is designed for scenario design, algorithm training, and algorithm evaluation. The platform design adopts a mode of separate simulation kernel and display module, and parallel and distributed operation, supporting rapid training of deep reinforcement learning algorithms and the verified evaluation of decision control effects. It avoids unknown risks and a large amount of time consumed by direct physical experiments of sea and air equipment, and it can effectively improve the efficiency of algorithm research and development.

Key words: swarm confrontation; deep reinforcement learning; platform design; parallel and distributed execution

0 引言 (Introduction)

当前,以深度强化学习为代表的人工智能方法在星际争霸^[1]、王者荣耀^[2]等即时策略游戏项目上的应用取得重大进展和突破。深度强化学习算法在复杂军事博弈对抗领域的应用正在加紧推进,相关科研项目 and 比赛亟需深度强化学习算法研究平台的支撑^[3-5]。对此,研究人员进行了初步探索和尝试,吴

昭欣等^[6]设计了基于深度强化学习技术的智能仿真平台,但其并不涉及对抗过程;卢锐轩等^[7]设计了基于人工智能技术的智能自博弈平台,支持智能体训练和自博弈训练,并进行了一对一空战仿真试验。海空集群对抗如果直接使用海空实际装备进行演习实验,不仅耗费巨大、效率低,而且风险极高^[8]。而如果采用深度强化学习算法解决对抗中的指挥决策问题,先利用仿真系统对算法进行训练和评价研究,再将算法迁移至实际装

备,这是一条可行路径。要想这样,一方面需要研究平台生成大量的样本数据用于算法训练,另一方面,需要研究平台合理评估算法性能,指导算法的演化改进。

根据海空集群对抗仿真实验的特点和研究人员的需要,本文设计了一种深度强化学习算法研究平台(以下简称平台)。在软件层面,平台支持设计作战想定对海空装备、战场环境及交战过程进行仿真;支持与深度强化学习算法进行交互,提供算法所需的训练数据并执行算法的输出命令;支持简单方便地进行算法修改或替换,平台运行效率高,支持分布式并行运行;在硬件层面,支持对海空集群的计算节点规模进行简单、方便的水平扩展。

1 平台架构设计(Platform architecture design)

1.1 应用需求分析

针对海空集群对抗问题具有的跨域跨平台、计算量大、快速推演等特点,平台主要需要实现以下功能。

1.1.1 海空集群对抗仿真

在海空装备实体模型和行为模型的基础上,能够针对作战场景想定,采用计算机仿真的方式对海空集群对抗过程进行仿真推演,并且深度强化学习算法能够与仿真进行交互,包括仿真的控制、仿真数据的获取及海空装备仿真实体的控制决策等。

1.1.2 深度强化学习算法设计

算法设计主要包括两方面的功能:一是提供典型的深度强化学习算法供用户直接使用,如深度 Q 网络(Deep Q-Network, DQN)^[9]、深度确定性梯度(Deep Deterministic Policy Gradient, DDPG)、近端策略优化(Proximal Policy Optimization, PPO)算法等,用户只需针对特定想定设计算法的输入、输出和奖励函数等,就可以直接进行训练和评价;二是提供算法自定义扩展开发的功能,用户能够自行设计开发相关新算法对智能指挥决策方法进行探索和研究。

平台需要支持算法训练和算法评价两个过程。算法训练要求平台能够对算法的相关参数进行设置,以及对算法过程和训练结果进行记录和展示;算法评价要求平台能够对算法训练结束后形成的模型进行有效性评估。

1.1.3 并行与分布式运行支撑

算法训练学习需要大量的样本数据,通过在仿真推演中不断试错实现学习、训练和提升,因此要求仿真能够在并行与分布式运行支撑下快速且高效地生成大量的数据样本,提高算法训练的效率。

1.2 平台软件架构

平台由海空集群对抗仿真系统(以下简称仿真系统)和深度强化学习系统组成,平台模块组成图如图 1 所示。仿真系统采用定步长的时间推进方式,在每个步长都可以通过仿真系统的外部访问接口控制仿真系统运行、获取战场态势和执行任务命令等。深度强化学习系统通过网络通信调用仿真系统的外部访问接口,实现对仿真系统的控制和信息获取。

仿真系统由想定模块、海空环境模型模块、海空装备模型

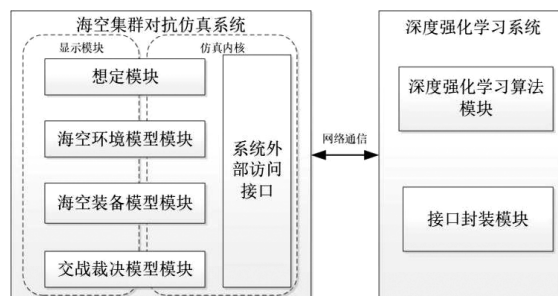


图 1 平台模块组成图

Fig. 1 Diagram of platform module composition

模块、交战裁决模型模块和系统外部访问接口等组成。

想定模块主要包括想定基本信息设置、兵力部署、条令规则设置、作战任务规划、想定打开与保存等功能,用于定义和设置作战对抗问题初始状态,如战场区域、推演方、作战时间、作战兵力、作战目标、作战行动等。

海空环境模型模块用于生成海空集群交战的战场环境信息,包括海域、岛礁和空域等三维模型,海空集群在此环境模型中进行作战活动。

海空装备模型模块由海上舰艇模型和空中战机模型组成,采用参数化建模框架,将模型向下一级分解,分为机动模型子模块、侦察模型子模块、火力模型子模块和任务处理子模块。机动模型子模块是指平台的机动能力,主要包括平均速度、最大速度、航程、爬升率等信息。侦察模型子模块是指平台具有的侦察能力,主要包括雷达、红外和可见光等侦察设备的侦察范围,以及对各类目标的发现概率等;火力模型子模块是指平台上所搭载的火力单元,主要包括各类炮、导弹等。各个型号的海空装备的参数化建模框架相同,只是具有不同的参数值。

交战裁决模型模块对海空装备的交火行为进行裁决,给出海空装备的受损信息,更新其状态。武器目标火力毁伤裁决过程如下所示。

- (1)加载相关信息,主要加载进攻武器战技指标参数、目标装甲防护能力、进攻武器到目标的距离、战场环境等信息。
- (2)将所需的参数传入击中概率计算规则进行计算,返回击中概率。
- (3)采用随机数生成器生成 0~1 的随机数。
- (4)比较随机数与击中概率,如果随机数小于等于击中概率,则判定为击中,否则为未击中,结束流程。
- (5)传入毁伤计算规则所需的参数进行毁伤计算,例如对舰船目标毁伤的规则如下:舰船被击中 1 发反舰导弹则判为失去动力无法机动,被击中 2 发反舰导弹则判为失去防空能力,被击中 3 发反舰导弹则判为击沉等。
- (6)输出毁伤结果,结束流程。

系统外部访问接口包括系统控制接口、态势获取接口和控制命令接口等。系统控制接口用于深度强化学习系统控制仿真系统的启动、停止、加载想定等;态势获取接口用于深度强化学习系统获取当前仿真系统内各个系统和实体的状态,如仿真的时间、双方海空装备的状态、环境信息等;控制命令接口用于

响应深度强化学习系统调用的任务指令,如机动、开火、侦察装备的开机和关机等控制命令。

仿真系统在具体实现时划分为仿真内核与显示模块,仿真内核关注仿真的高效推演计算,不包括界面显示,其推演过程展现由显示模块完成。仿真内核与显示模块的分离,使得算法在训练时只需要使用仿真内核而无需使用显示模块,这样就可以避免耗费计算渲染资源,加快仿真推演进程;在对算法模型进行评价分析时,同时运用仿真内核和显示模块详细展示海空集群对抗仿真全程,便于用户直观地理解战斗过程。

深度强化学习系统包括深度强化学习算法模块和接口封装模块。深度强化学习算法模块采用深度神经网络进行建模,用于控制仿真系统的运行,读取战场态势信息,包括战场环境信息、敌方兵力部署和状态信息、我方兵力部署和状态信息等,采用分布式强化学习算法进行训练,输出海空集群的联合动作,通过接口封装模块将动作转换为单个平台的控制命令,调用仿真系统的外部访问接口传输给仿真系统进行处理和响应。

1.3 平台硬件架构

平台采用分布式的网络架构,深度强化学习系统部署在一台服务器上,仿真系统部署在多台计算节点上,服务器和计算节点间通过网络连接,网络通信采用 gRPC 等协议,平台硬件架构图如图 2 所示。每台计算节点上可以运行多个仿真系统实例,实现分布式并行和加速,从而使训练样本的生成速度和算法的学习效率大大提高,具有良好的扩展性。服务器和计算节点可以根据需要配置图形处理单元 GPU,实现仿真和算法的高效运行和训练。

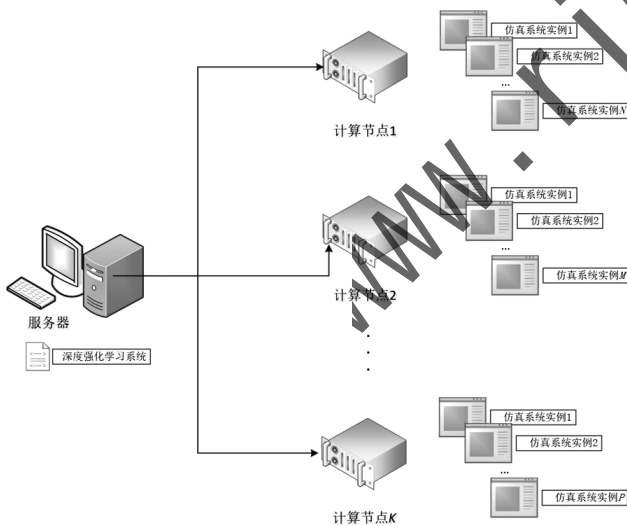


图 2 平台硬件架构图

Fig. 2 Platform hardware architecture

2 平台使用流程(Platform usage process)

平台使用流程依次分为想定设计、算法训练和算法评价三个子流程。

2.1 想定设计子流程

想定设计子流程(图 3)实现红蓝对抗场景的设计,主要包括以下步骤。

(1)设置推演方和时间:一般为对抗关系的红蓝双方,想定时间为对抗开始和结束时间。

(2)部署兵力:将推演方的海空装备作为兵力在战场环境模型中进行部署,包括海空装备的类型、挂载、数量、位置等信息。

(3)设置条令规则:对不同场景下的武器控制规则进行限制,如开火规则分为自由开火(向所有未识别为友方的单元开火)和谨慎开火(仅对识别为敌对方的单元开火),机动规则(设置海空装备受到攻击时,是否忽略计划航线)。

(4)规划作战任务:一般情况下,推演双方中一方的兵力由深度强化学习算法进行控制,另一方的兵力由预先制定的作战任务规划或者其他的算法进行控制。作战任务规划设计海空兵力的作战任务,如巡逻任务、打击任务、拦截任务等,并且规定在不同的条件下实施相应的作战任务。

(5)保存想定:保存设计的想定,用于算法训练和评价时进行加载。

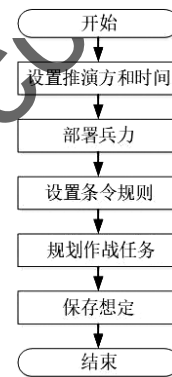


图 3 想定设计子流程

Fig. 3 Scenario design sub-process

2.2 算法训练子流程

算法训练子流程(图 4)实现算法训练生成模型,主要包括以下步骤。

(1)调用仿真系统的外部访问接口,开启 C 个仿真系统实例,这些实例可以运行于一台或多台计算的节点上,能够通过计算节点 IP 地址来指定。

(2)启动仿真系统实例,加载想定,初始化海空装备状态。

(3)初始化深度神经网络参数、学习率等超参数。

(4)在每个时间步,强化学习算法模块通过调用仿真系统的外部访问接口,用于获取当前仿真信息和收集训练样本。

(5)判断训练样本数量,当训练样本超过设定的阈值时,开始神经网络的训练,训练完成后保存网络模型,采用训练的神经网络为不同的仿真系统实例生成作战行动。当训练样本数不够时,使用初始化的神经网络为不同的仿真系统实例生成作战行动。

(6)采用接口封装模块对作战行动生成作战命令,调用仿真系统的外部访问接口传入命令。

(7)仿真系统内部执行命令、更新状态,如果对战完成则重新加载想定进行下一局的仿真推演。

(8)如果算法训练完成,则结束训练,获得保存的网络模型,供下一步的评价使用。

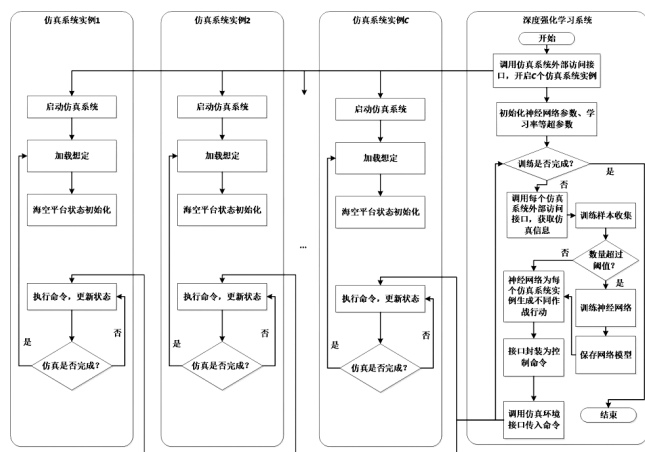


图4 深度强化学习算法训练子流程

Fig. 4 Sub-process of deep reinforcement learning algorithm training

2.3 算法评价子流程

算法评价子流程(图5)实现对算法的评价,主要包括以下步骤。

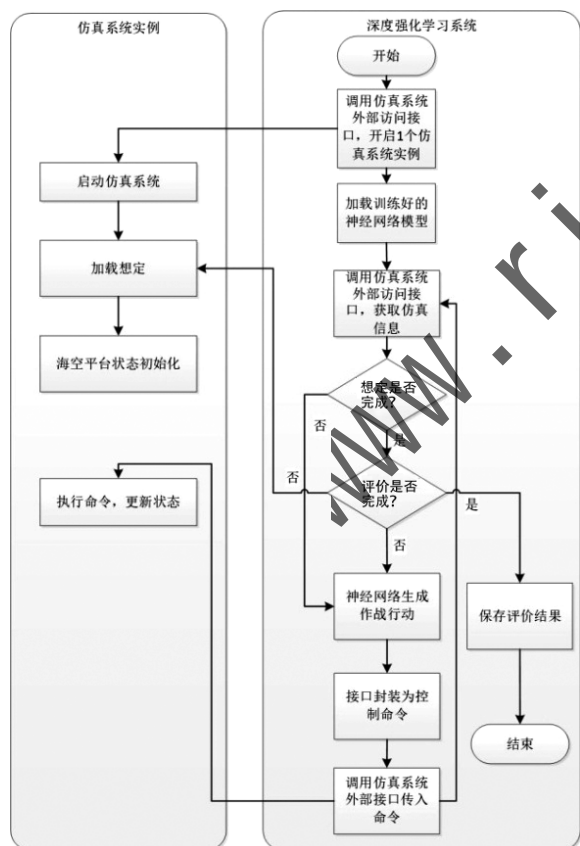


图5 深度强化学习算法评价子流程

Fig. 5 Sub-process of deep reinforcement learning algorithm evaluation

(1)首先调用仿真系统外部访问接口,开启一个仿真系统实例,然后启动仿真系统,加载想定,此时如果要展示战斗过

程,则可以接入显示模块。

(2)加载训练好的深度网络模型。

(3)深度强化学习算法调用仿真系统的外部访问接口,获取当前的仿真信息。

(4)判断想定是否完成,如果完成则判断评价是否完成,否则转到“步骤(6)”。

(5)判断评价是否完成,如果完成则保存评价结果,结束流程,否则加载想定,转到“步骤(6)”。

(6)神经网络模型生成作战行动。

(7)将作战行动封装为控制命令。

(8)调用仿真系统外部接口传入控制命令。

(9)仿真系统执行命令,更新状态。

3 结论(Conclusion)

研究海空集群对抗仿真决策控制的深度强化学习算法时,需要对所研究算法产生的大量样本进行高效训练并对其结果验证评价,针对这一问题,算法研究平台通过软件和硬件架构设计,仿真内核、显示模块分离设计及并行与分布式运行设计,可以快速对所研究的强化学习算法进行训练,还能对其决策控制效果进行验证评价。该算法研究平台避免了直接进行海空装备实物实验需要耗费大量时间和可能产生未知风险的问题,满足了算法研究人员对仿真平台的需求,能够提高海空集群对抗仿真过程中深度强化学习算法的研发效率。

参考文献(References)

- [1] VINYALS O, BABUSCHKIN I, CZARNECKI W M, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning[J]. Nature, 2019, 575: 350-354.
- [2] YE D H, LIU Z, SUN M F, et al. Mastering complex control in MOBA games with deep reinforcement learning[J]. Proceedings of the AAAI conference on artificial intelligence, 2020, 34(4): 6672-6679.
- [3] 马悦, 吴琳, 郭圣明. 智能博弈技术军事应用展望[J]. 指挥与控制学报, 2023, 9(2): 135-145.
- [4] 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.
- [5] 周来, 靳晓伟, 郑益凯. 基于深度强化学习的作战辅助决策研究[J]. 空天防御, 2018, 1(1): 31-35.
- [6] 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.
- [7] 卢锐轩, 孙莹, 杨奇, 等. 基于人工智能技术的智能自博弈平台研究[J]. 战术导弹技术, 2019(2): 47-52, 98.
- [8] 刘全, 翟建伟, 章宗长, 等. 深度强化学习综述[J]. 计算机学报, 2018, 41(1): 1-27.
- [9] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518: 529-533.

作者简介:

刘宝宏(1974-),男,博士,副教授。研究领域:人工智能与仿真,军用仿真。