DOI: 10.16338/j.issn.1009-1319.20190365

深度强化学习在作战任务规划中的应用

姚 桐　王 越　董 岩　戚 锦　耿修堂

摘　要　针对军事(战术级)作战任务规划面临的战场环境复杂、信息不完全、不确定性大、策略复杂度高等挑战，梳理了作战任务规划基本概念和流程框架，介绍了深度强化学习的基本原理和发展现状，分析了深度强化学习在作战任务规划中场景识别、目标检测、行为判断、威胁评估、路径规划、火力分配等方面的应用，为作战任务规划的智能化发展提供了新的研究思路。

关键词　深度学习　强化学习　任务规划　智能规划

引　言

随着作战任务的日趋复杂，作战方式正由传统的单一兵种作战向陆海空天电网一体化作战转变，而大数据、机器学习等人工智能技术的快速发展，又推动着未来战争形态向无人化、智能化方向快速发展。在这个过程中，对复杂作战任务进行规划，实现多兵种多武器平台之间的高效协同具有非常重要的意义。

本文从作战任务规划的概念和框架出发，简要介绍了深度强化学习的原理和发展，分析了深度强化学习在作战任务规划中的应用，为未来作战任务规划信息化、智能化发展提出了新的研究思路。

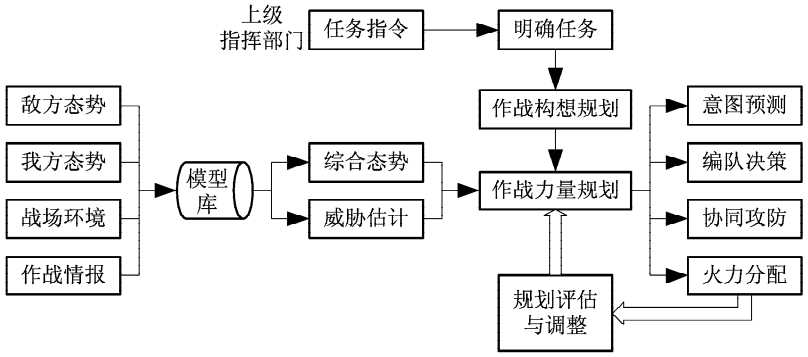

图1　作战任务规划框架结构图

1　作战任务规划框架

任务规划是指特种部队针对战场复杂环境，在完成环境感知和态势认知的基础上，对目标选择、路线规划、火力分配、作战方式等作战要素和作战活动进行统筹规划的战术过程。在外军中，任务规划定位在战术系统级、武器单元级，如战斧巡航导弹任务规划系统。目前，所有的任务规划系统基本都是针对战术飞行器的，直到2014年，美国陆军才着手开发陆军任务规划工具，用于装甲营、运输营等的行动规划。作战任务规划是作战指挥部队运用先进的现代作战信息技术，围绕作战目标和任务，全面分析敌我态势、作战资源和复杂战场环境，有效评估作战效能，系统分析和优化决策规划，合理实施作战行动的进程，结合作战仿真和兵棋推演消除可能出现的矛盾和冲突，提供合理的作战方案和可行的作战计划，选择最优作战方案，配合先进技术完成作战任务。

作战任务规划框架结构如图1所示。上级作战指挥部门根据下达的指令明确作战目标和任务，结合战场情报信息、基础技

基金项目：装备预研领域基金项目(61403120205)
本文2020-01-17收到，姚桐、王越分别系西北机电工程研究所助理工程师、高级工程师

术统计数据、目标规划模型，对敌我态势和行动以及战场外部环境的变化进行感知，估计敌方目标潜在威胁等级，利用任务规划模型对作战构想、各编队的主战、作战保障和后装保障等作战力量进行规划，完成目标检测识别、目标攻击意图预测、侦察编队决策、武器编队决策、协同攻防决策等，最后利用推演和评估分析技术对作战任务规划系统进行验证，根据结果对作战任务规划系统的功能进行调整和完善。

现阶段常用的作战任务规划系统多以人为核心，依靠指挥员的经验和知识对战场态势进行理解和判断，进而做出威胁评估、编队协同、火力分配等决策。但未来战争作战方式多域混合，参与作战的武器装备种类和数量众多，战场态势错综复杂且瞬息万变，战场情报和信息量迅速增多，这些因素决定了未来战场态势复杂多变且具有不确定性。传统的作战任务规划方式已越来越难以完全满足及时准确掌握战场态势、快速精准做出决策规划的需求，迫切需要智能化作战任务规划方式来提高其态势感知和自主决策的能力，发挥武器装备的最大作战效能。

2 深度强化学习概述

近年来，以深度强化学习为代表的人工智能技术的快速发展使得直接从战场原始数据中快速提取特征从而对战场态势进行描述、感知和进一步决策成了可能。

深度学习的本质是通过构造多层神经网络，提取输入数据的特征，挖掘和获取数据中包含的深层信息。自 2012 年 AlexNet 卷积神经网络在 ImageNet 大赛取得了重大突破以后，随着计算能力的提高和算法性能的发展，深度学习在计算机视觉领域得到了越来越广泛的应用，如人脸检测、图像分类、行为识别等，其典型算法有基于区域选择的 RC-NN 系列和端到端的 YOLO、SSD 系列。与传统的特征提取方法相比，深度学习避免了人工提取特征的主观性，通过深层次的神经网络提取多尺度特征并进行融合，提高了对目标的描述能力和感知能力。

强化学习是当前机器学习领域中的一个重要研究领域，它以试错机制与环境反馈进行交互，通过最大化累积奖赏的方式来选择最优策略。智能体在当前状态 $s(t)$ 下，根据策略 π 来选择动作 $\alpha(t)$，环境接收到该动作并转移至下一状态 $s(t+1)$，智能体接收环境反馈回来的奖赏 $r(t)$ 并根据策略选择下一步动作。强化学习不需要监督信号，可以在未知环境中进行探索和应用，其主要算法有蒙特卡洛强化学习、时间差分学习、策略梯度等。

早期的强化学习主要依赖人工提取特征，缺乏客观性和全面性，难以处理复杂状态下的问题。随着深度学习的发展，可以通过深度神经网络从原始数据中直接提取特征。深度学习提取特征和描述感知的能力较强，但推理决策相对较弱；强化学习的推理决策能力较强，但描述感知相对缺乏。因此，深度强化学习等将两者结合起来的算法便逐渐成为研究热点。深度强化学习是在训练中进行试错，奖励和惩罚作为反馈调整神经网络，得到更好的策略模型。

2015 年，DeepMind 团队提出了深度 Q 网络(DQN)，将深度卷积神经网络和 Q 学习结合到一起，只使用游戏的原始图像作为数据的输入，不依赖于人工提取特征，是一种端到端的学习方式，在 Atari 系列游戏上达到了人类的决策和控制效果。AlphaGo 就通过 DQN 在自我博弈中实现奖励积累的最大化，据此选择各个状态下的最好走法。这一算法更加符合现实世界中人类的决策思维，在智能机器人控制、无人驾驶、游戏通关、棋类对弈等多类决策和控制问题中得到广泛应用。2017 年，DeepMind 团队公布的 Alpha 系列最新研究成果AlphaZero，采取了一种自我博弈策略简化算法，与 AlphaGo 相比泛化能力更好，可使用完全相同的算法和超参数，不需要任何先验知识的情况下依靠自我博弈，只需几小时训练模型，就可以在围棋、国际象棋、日本将棋三种棋类对弈中战胜顶尖人工智能程序；2017 年，Ruslan 首次提出将记忆引入深度强化学习的思想，通过位置信息感知避免过多的记忆重写，使模型在不同环境下都有较高的效率和较好的效果。AlphaZero 的策略简化和 Ruslan 的记忆引入思想，都反映出深度强化学习的研究热点主要集中在模型性能效果和泛化能力

的提升上。

表 1 列举了典型的几种深度强化学习算法，并对其特点和性能进行了比较。

表 1　深度强化学习典型算法特点和性能比较

算法	特点	Atari 表现
DQN	采用记忆回放解决数据关联性问题，分离目标网络	100%(基准)
Dueling DQN	采用价值函数和优势函数网络结构	151.72%
A3C	采用多线程方法异步并发学习，避免经验回放相关性过强	163.07%
ACKTR	融入分布式 Kronecker 因子分解提升样本效率和可扩展性	353.87%
PPO	解决了 Policy Gradient 算法步长难以确定的问题，更易求解	46.26%

3　深度强化学习在作战任务规划中的应用

作战任务规划的 OODA 循环如图 2 所示，各作战单元实现战场态势信息的实时感知和共享，并在此基础上对战场整体态势进行理解，实现分布式的动态规划与协调。态势理解介于态势感知与规划调整之间，依据态势感知进行态势理解，再应用态势理解进行作战规划的调整。可以看出，态势感知和态势理解是作战任务规划的基础和前提。由上节对深度强化学习的简要介绍可知，在作战任务规划系统中，深度学习可用于战场态势感知及浅层知识推理，强化学习可用于战术决策及任务规划。

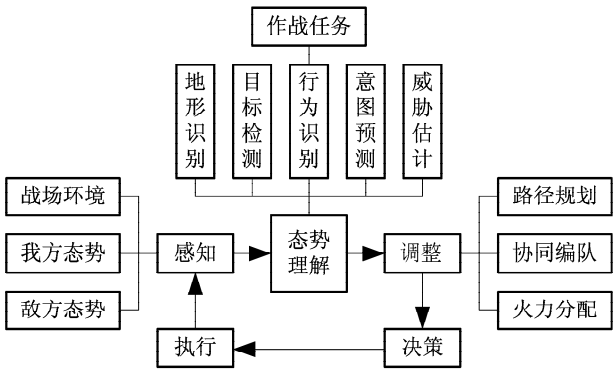

图 2　作战任务规划 OODA 循环

3.1　态势理解

态势理解也称作态势认知，是在多源信息融合的基础上对战场态势的感知、分析和预测，如目标检测、行动判断、威胁估计、意图预测等。通过获取海量战场信息和数据，实时感知并准确理解战场态势，挖掘复杂态势中的隐藏信息，是指挥员做出正确决策规划的基础。

近年来，快速发展的深度学习技术通过构建多层卷积神经网络，采用无监督逐层训练方法，提取目标的多尺度特征，避免了梯度发散和局部最优问题，在目标检测、行为识别、无人驾驶等诸多领域都得到了广泛应用。将深度学习应用在态势理解中，通过多层神经网络逐层进行训练，实现对战场态势的认知和理解。训练数据的来源包括历史数据、演习训练、靶场试验、战场仿真等。

战场态势理解框架流程如图 3 所示，外部情报、自身装备、协同平台等多源数据作为输入，通过预先训练好的多个深度模型，提取出态势元素的多尺度特征，完成战场环境识别(城市、山地、丛林等)、战场目标的类型识别(装甲、坦克、无人机等)和行为识别(静止、机动、速度等)，在此基础上分析敌方目标的战场兵力部署，分析和预测敌方下一步行动，综合得到对敌方目标攻击意图和威胁度的判断估计，形成完整的战场态势图。

3.1.1　场景识别

战场环境是战争的载体，影响着作战的全过程。对战场环境的准确分析，是态势感知的基础。当前对战场场景的识别和分析多以人为主，但在未来战争中，战场涉及海陆空多域，单纯依靠人的主观判断将难以满足实际作战需求，而深度学习等人工智能技术在场景识别领域已经取得了不错的研究成果。

2016 年，麻省理工大学计算机科学与人工智能实验室公开了场景数据集 Places，包含百万张不同场景图片，同时用 AlexNet、GoogleNet、VGGNet 等深度网络在该数据集上进行训练，得到的模型显著提高了场景识别的准确

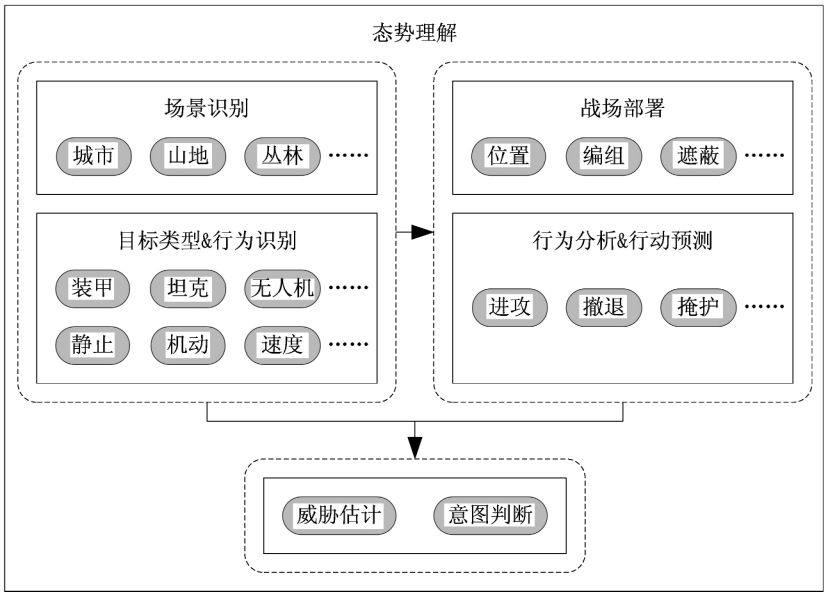

图 3　战场态势理解框架流程图

率，并公开了网络训练的模型参数，可将这些网络模型作为预训练模型在小样本战场场景数据集上进行迁移学习，找出输出结果可用的层，用这些层的输出作为输入对网络结构进行微调，训练一个参数量更小、规模也更小的网络模型，提高模型准确率和鲁棒性的同时缩短训练时长。

3.1.2　目标检测

在未来战争中，敌方目标跨越多域多兵种，目标的多样性、隐蔽性、伪装性以及战场环境的复杂性都大大提升了目标检测识别的难度，传统的人工提取特征设计分类器的方式难以满足作战中快速准确识别敌方目标的需求，而深度学习的快速发展以及大量公开图像数据集的建立，使得战场目标成为可能。

与传统目标检测方法相比，基于深度学习的目标检测方法显著提高了检测性能，是近年来计算机视觉领域的研究热点。基于区域选择的 R-CNN 系列与端到端的 YOLO 和 SSD 系列在检测精度和检测速度上均有了较大提高，部分模型已能够达到特定应用需求。

美国国防预先研究计划局(DARPA)开展的心灵之眼研究项目，采用深度学习等智能方法对视频和图像进行感知和推断；由 DARPA 和 AFRL 共同发起的 MSTAR SAR-ATR 系统，采用图像识别技术提高对 SAR 图像识别的准确率；2017 年，美国成立算法战跨职能小组(AWCFT)，通过数据标注、算法优化等技术手段，将深度学习等人工智能技术用在无人机上，以支持对 ISIS 的作战。

3.1.3　行为判断

在搜索并检测到目标后，对目标的行为做进一步判断，从而对目标的意图进行预测，判断出静止、攻击、撤退、侦察等，作为威胁评估和打击决策的输入依据，是态势理解中的重要一环。

行为判断是通过深层神经网络提取出不同行为的多尺度特征，经过分类回归将待检测的行为划分到行为集合中的某一类。特征提取、行为描述和回归分类是行为判断的 3 个环节。

Ji 等首先把卷积神经网络扩展到三维，提出三维卷积用于行为识别，通过三维卷积同时编码三维视频数据中的空间信息和时间信息；Tran 等对三维卷积做了进一步加深和改进，促进了其在行为识别领域的发展；Simonyan 等提出双流卷积神经网络方法，使用基于 RGB 图片的空间流卷积神经网络和基于光流图的时间神经网络分别提取静态特征和动态特征，最后将双流信息进行融合用于分类识别，大幅提高了行为识别的准确率。

3.1.4　威胁估计

威胁估计是根据当前战场态势对敌方军事力量的杀伤能力和对我方的威胁程度的一种综合评估。在战场态势感知的基础上，依据敌我双方的兵力部署、武器性能、敌方攻击意图和我方作战策略，以定量分析的方式对敌方威胁等级进行评估。

现代战场的快节奏要求指挥员必须具备更快的超前反应和作战能力，从战场上海量多源数据信息中实时分析和估计目标威胁等级，从而尽可能使指挥员针对敌方作战行动场景提前进行决策规划，而不是在敌方行动后再做出反应，这种超前行动方式更适

合复杂、多变、突发性强的未来作战场景，具有非常重要的意义。因此，及时准确地评估敌方目标威胁度，并根据我方的作战方案和武器系统的性能，提前进行科学火力分配和打击决策，是提高战场致胜能力的关键。

DARPA 开展的洞察项目，通过对多类传感器数据进行融合和推理，实现对潜在威胁的评估和预测。

3.2 策略规划

策略规划是作战任务规划的核心。随着战争向信息化、智能化转变，越来越需要智能决策和规划系统来辅助指挥员进行作战规划和指挥。智能决策规划消除了人的主观因素影响，具有自我学习、修正、推理和决策的能力，显著提高了作战任务规划的准确性和实时性，提升了情报分析、辅助决策和指挥控制能力。

DARPA 开展的深绿计划，采用计算机模拟仿真技术，演示不同作战方案可能产生的结果，通过预估敌方行动，缩短制定作战计划的时间，辅助指挥员快速做出正确的决策；DeepMind 建立的强化学习策略系统，不仅可以下围棋、玩游戏，还可以在多种战略任务的执行中达到与人类匹敌的效果。而 DeepMind 正在开发的战争策略游戏“星际争霸”，游戏过程更加接近于真实的作战场景。

3.2.1 路径规划

路径规划可分为全局路径规划和局部路径规划。全局路径规划的前提是区域地图信息已知，根据目的地的位置信息在地图上确定可行区域，选择最优路径；局部路径规划是在车辆传感器能够探测的范围内进行规划，根据探测到的外部环境信息、道路条件和意外事件等，快速准确地做出判断并规划出当前最优的行驶路线。

目前对于已知环境的全局路径规划已有很多相对成熟的算法，可实现无碰撞到达目标地点，但作战场景涉及城市、山地、丛林等各类复杂地形，难以提前预知精确的地图信息，如何根据传感器实时探测到的战场环境信息，进行局部路径规划和实时调整，实现战场上各武器指挥平台的快速调度和协同，是亟待解决的问题。

与其它路径规划算法相比，强化学习的一个重要优势在于强化学习不依赖环境建模，不需要对环境的先验知识，只需要给出奖励信号，智能体就可以采用试错的方式，通过与周围环境不断进行交互找出最优策略。强化学习方法将传感器所感知到的外界环境状态反馈映射到执行器动作，从而对外界环境变化快速响应，实现自主路径规划，具有实时、快速和鲁棒性强的优点。

Beom 等将模糊逻辑与强化学习结合，实现了地面移动机器人的导航；Deisenroth 等将高斯过程与基于模型的策略搜索强化学习结合，应用在了机器人的移动控制中；Xue 等建立两级分层控制框架，采用深度强化学习方法对两层控制策略进行训练，实现了不同地形变化条件下的自主导航。

3.2.2 火力分配

火力分配是在威胁估计和目标排序的基础上，我方武器平台对敌方目标的打击方式决策。因为作战场景中敌我双方武器装备的多样性、对抗性和不确定性等，传统的火力分配方法（矩阵对策法、优势函数法、优化指向向量法等）难以快速准确地完成最优火力分配。

在战场动态环境中，通过建立目标 Agent 模型，将环境事件和行为规则作为输入，不断根据环境变化修正对当前态势的认知，并通过获取外部环境对当前动作的回报奖励，利用强化学习方法不断更新策略知识库，根据毁伤效能最大化原则生成适应于当前战场环境的火力分配策略，并通过作战结果的反馈对火力分配策略进行修正，将生成的新的火力分配策略保存到策略库中。

结合了强化学习的火力分配方法能够感知自身所处的战场环境，并通过奖励反馈自适应于外部环境，从而构建更加准确合理的战场火力分配模型。

4 结束语

深度强化学习等新技术的兴起表明人工智能已经进入了快速发展阶段，也预示着未来智能化战争的到来。本文着眼于智能化战争中发展作战任务规划的迫切军事需求，结合对于深度学习和强化学习相关内容的阐述，分析了深度强化学习在作战任务规划中的应用，为未来作战任务规划信息化、智能化发展提出了新的研究思路。

参考文献

[1] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014

[2] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection. Proceedings of the IEEE Conference on Computer vision and Pattern Recognition, 2016

[3] Liu W, Anguelov D, Erhan D, et al. Ssd: single shot multibox detector. European Conference on Computer Vision Springer, Cham, 2016

[4] 唐振韬，邵坤，赵冬斌. 深度强化学习进展：从 AlphaGo 到 Alpha Zero. 控制理论与应用，2017，34(12)

[5] 刘全，翟建伟，章宗长. 深度强化学习综述. 计算机学报，2018，41(1)

[6] Hausknecht M, Stone P. Deep recurrent q-learning for partially observable mdps. AAAI Fall Symposium Series, 2015

[7] Silver D, Hubert T, Schrittwieser J, et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. arXiv preprint arXiv: 1712.01815, 2017

[8] Dai Z, Yang Z, Yang F, et al. Good semi-supervised learning that requires a bad gan. Advances in Neural Information Processing Systems, 2017

[9] Ji S, Xu W, Yang M, et al. 3D convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 35(1)

[10] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3D convolutional networks. Proceedings of the IEEE International Conference on Computer vision, 2015

[11] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. Advances in Neural Information Processing Systems. 2014

[12] Peng X B, Berseth G, Yin K K, et al. Deeploco: dynamic locomotion skills using hierarchical deep reinforcement learning. ACM Transactions on Graphics (ToG), 2017, 36(4)

(上接第 15 页)

[3] Kareem A, Kenneth P. Strategy in the age of artificial intelligence. Journal of Strategic Studies, http://www.tandfonline.com/doi/figure/10.1080/01402390.2015.1088838?scroll=top&needAccess=true, 2015-11-08

[4] Center for a New American Security. Why America needs a new way of war. https://www.CNAS.org, 2019-06-17

[5] Cheryl P. Work: human-machine teaming represents defense technology future. DoD News, https://www.defense.gov/News/Article/Article/628154/work-human-machine-teamingrepresents-defense-technology-future/, 2015-11-25

[6] 刁联旺. 美国 DARPA 有关“算法战”项目的发展分析与认识. 第一届空中交通管理系统技术学术年会论文集，2018

[7] MIT Committee to Evaluate the Innovation Deficit. The future postponed why declining investment in basic research threatens a US innovation deficit. https://dc.mit.edu, 2015-04-27

[8] US Air Force. 2019 the United States Air Force artificial intelligence annex to the department of defense artificial intelligence strategy. https://www.USAF.gov, 2019-09-29