·智能博弈对抗专题·

doi：10.15908/j.cnki.cist.2021.02.003

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*

孙 彧[1,2] 李清伟[3] 徐志雄[4] 陈希亮[2]

（1 解放军31102部队 南京 210000）（2 陆军工程大学指挥控制工程学院 南京 210007）
（3 中国电子科技集团公司第二十八研究所 南京 210007）
（4 陆军边海防学院 西安 710100）

摘 要：鉴于多智能体深度强化学习在解决蜂群系统、能源分配和大型团队策略游戏等高维复杂动作空间以及多智能体决策问题中的良好表现，提出了一种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。在对多智能体深度强化学习基本概念和对空战策略生成的可行性分析的基础上，建立了基于多智能体马尔可夫决策过程空战配合策略的框架，从而生成最优对抗策略。实践表明，该模型可生成多种双机联合对抗策略，具有较高的研究价值和实际意义。

关键词：多智能体深度强化学习；多智能体马尔可夫决策；空战博弈对抗；战术决策

中图分类号：TP181 **文献标志码**：A **文章编号**：1674-909X(2021)02-0016-05

Game Confrontation Strategy Training Model for Air Combat Based on Multi-agent Deep Reinforcement Learning

SUN Yu[1,2] LI Qingwei[3] XU Zhixiong[4] CHEN Xiliang[2]

（1 Unit 31102 of PLA, Nanjing 210000, China）
（2 Command and Control Engineering College, Army Engineering University, Nanjing 210007, China）
（3 The 28th Research Institute of China Electronics Technology Group Corporation, Nanjing 210007, China）
（4 Army Academy of Border and Coastal Defence, Xi′an 710100, China）

Abstract: In view of the good performance of multi-agent deep reinforcement learning in solving the high-dimensional complex action space and multi-agent decision problems, such as the swarm systems, the energy distribution and the large team strategy games, a game confrontation strategy training model for air combat based on multi-agent deep reinforcement learning is proposed. Based on the analysis of the basic concept of multi-agent deep reinforcement learning and the feasibility of air combat strategy generation, the air combat coordination strategy framework of multi-agent Markov decision process is established to generate the optimal confrontation strategy. The practice shows that the model can generate a variety of dual-fighters joint confrontation strategies, thus has high research value and practical significance.

Key words: multi-agent deep reinforcement learning; multi-agent Markov decision; air combat game confrontation; tactical decision

* **基金项目**：国家自然科学基金（61806221）、国防科技创新特区163计划（1916311LZ00100301）、装备发展部“十三五”预研课题（31505550302）和国防科技重点实验室基金（6142101180304）资助项目。

收稿日期：2020-10-28

引用格式：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.

0　引　言

随着科学技术的不断发展，现代战争的节奏不断加快，战争复杂性不断上升，人脑决策已无法跟上战场态势快速更迭的步伐。空战领域作为未来战争的焦点，迫切需要快速决策、自动决策和自主决策。因此，构建智能化空战博弈对抗模型成为重点，主要体现在以下3个方面：1）训练空中编队指挥员或飞行员，丰富空战人员的战术经验；2）辅助地面领航员及空战编队作战人员进行引导或战术决策，减轻地面领航员及空战编队作战人员的压力；3）操控智能空战飞行员参与危险区域空战，降低飞行人员的风险性与空战成本。

近年来，以深度强化学习（DRL）[1]为基础的智能化求解方法在围棋[2]和雅达利游戏[3]等决策控制领域[4]取得了成功，并有望在指挥控制领域的智能化决策问题中取得突破。单智能体DRL算法应用于多智能体系统（MAS）[5]领域会产生环境非平稳性[6]、训练维度过大[7]及不完全信息博弈[8]等问题，导致学习效果欠佳。由于高精度和高仿真度的军事决策[9-10]往往采用多个相互配合的实体完成共同的作战任务，故与多智能体深度强化学习（MADRL）[11]方法较契合。因此，MADRL方法应用于小范围空战博弈对抗等以完成训练策略生成将成为未来人工智能作战模拟仿真的发展方向。

1　基本概念及可行性分析

多智能体强化学习（MARL）[12]最初是一类用于解决MAS中智能体协作完成配合任务的方法集，与传统强化学习（RL）[13]方法类似，MARL也是通过与环境不断交互及迭代来学习最优策略。MARL与环境的基本交互流程如图1所示。MARL通过控制作战实体与战场环境的不断交互来学习最优决策序列，并做出符合人类指挥员思维方式的作战策略规划。

传统MARL可在网格世界和九宫格等小型仿真系统中完成多个智能体间的简单合作任务，但不适用于高维复杂环境。此外，仿真试验中的单智能体DRL算法无法适用于多智能体环境，主要原因如下：1）多智能体环境中的状态和动作空间随智能体数量的增加呈现指数级增长趋势，从而导致学习过程难以收敛；2）智能体策略在训练过程中因互相影响会导致训练不平稳，严重制约了最优策略的生成；3）集中式学习器难以形成有效的回报分配方法，从而导致智能体陷入局部最优。

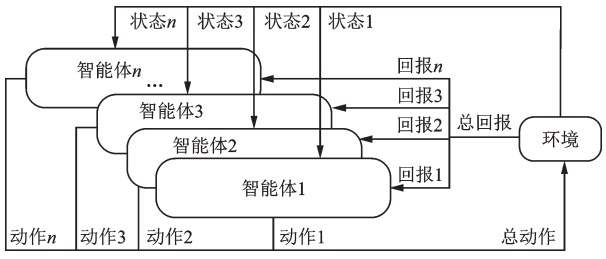


图1　MARL与环境的基本交互流程

为了解决上述问题，研究人员在传统DRL算法中融合了博弈论、分散式训练架构和分层训练等理念，从而强化了算法在多智能体环境中的适应性，同时也催生了MADRL。经过数年发展，MADRL诞生了众多算法、规则和框架，并在集群系统、资源分配、股票分析和路由协议等领域得到了广泛应用。例如，美国Open AI团队开发的多智能体深度确定性策略梯度（MADDPG）[14]算法通过设置中心化的评论家（critic）网络和分散化的演员（actor）网络，使环境中的每个智能体在感知全局状态情况下同时执行自我策略，既实现了智能体间的信息交互，又缓解了多智能体环境的非平稳性问题；反事实多智能体策略梯度（COMA）[15]算法在使用统一集中化训练网络的同时利用反事实基线（counterfactual baseline）衡量每个智能体对于环境的贡献值，并将其纳入回报分配系统，以解决多智能体环境中回报分配的问题。上述算法的使用显示了MADRL技术在不完全信息作战条件下的巨大潜力，也为解决智能决策问题提供了经验和参考。

2　空战博弈对抗策略训练模型

2.1　基于MAMDP的空战决策过程

采用MADRL算法解决空战博弈对抗策略生成问题，可认为是在连续动作状态空间上的及时决策过程，环境中的决策实体遵循多智能体马尔可夫博弈过程（MAMDP）[12]。MAMDP通常表示为以下五元组形式：

$$M=\left\langle \boldsymbol{S},\boldsymbol{A}_1,\cdots,\boldsymbol{A}_i,\cdots,\boldsymbol{A}_n,R_1,\cdots,R_i\cdots,R_n,f,\gamma\right\rangle \tag{1}$$

其中，$i=1,2,\cdots,n$，n为智能体数；$\boldsymbol{S}$为环境状态空间；$\boldsymbol{A}_i$为智能体i的动作空间；R_i为智能体i在执行动作$\boldsymbol{A}_i$后获得的瞬时回报；$f:\boldsymbol{S}\times\boldsymbol{A}\times\boldsymbol{S}\rightarrow[0,1]$表明智能体执行动作$\boldsymbol{A}$，从当前状态$\boldsymbol{s}$（$\boldsymbol{s}\in\boldsymbol{S}$）到下一个状态$\boldsymbol{s}'$转换函数的概率分布。

式（1）中各变量之间的关系可表示为：环境中的N个智能体在状态$\boldsymbol{S}$下做出联合动作$\boldsymbol{A}_1,\boldsymbol{A}_2,\cdots,\boldsymbol{A}_n$，

并从环境中获得瞬时回报$R_1, R_2, \cdots, R_n$，而智能体动作的选取则依据函数f的概率分布。MAMDP中，智能体的学习目标是找到最优联合策略，即最大化t到T的整体折扣回报值R，公式如下：

$$R=\sum_{1}^{n}\sum_{t}^{T}\gamma^{t} r_i^t \tag{2}$$

其中，γ为折扣系数，$\gamma\in[0,1]$。

2.2 空战场景建模与规则

2.2.1 空战场景建模

本文构建了2对2的视距内空战场景。该场景既可通过规则制定红蓝双方的作战行动序列，又可通过快速博弈对抗产生大量数据训练决策模型。对抗双方为红方和蓝方，其中，红方为由1架Su-30和1架J-10组成的双机编队（使用MADRL算法进行仿真）；蓝方为由1架F-16和1架F-18A组成的双机编队（使用固定战术规则）。双方战机的机载火控雷达工作模式包括扫描、跟踪和锁定3种状态；红方和蓝方战机分别搭载PL-15和AIM-120空空导弹。交战空域设置为矩形二维平面环境。双方战机随机设置出生点，并在指定对抗空域内自由空战。蓝方战机在空域内按规则巡逻机动，一旦感知红方战机威胁就伺机歼灭之，若目标歼灭或丢失后继续巡逻搜索；红方战机的任务是歼灭蓝方所有战机并取得预设空域制空权。

2.2.2 空战规则

每个时刻t的环境建模和交战规则如下：

1）设x_{A}^t和y_{A}^t分别为A战机在t时刻在二维平面交战空域的横纵坐标点；x_{B}^t和y_{B}^t分别为B战机在t时刻在二维平面交战空域的横纵坐标点。

2）设$D_{\mathrm{A,B}}^t$为战机A和B在t时刻的距离，本文使用二维平面欧式距离表示战机间距，公式如下；

$$D_{\mathrm{A,B}}^t=\sqrt{(x_{\mathrm{A}}^t-x_{\mathrm{B}}^t)^2+(y_{\mathrm{A}}^t-y_{\mathrm{B}}^t)^2} \tag{3}$$

3）设$\theta_{\mathrm{A,B}}^t$为t时刻A和B战机间的夹角，公式如下：

$$\theta_{\mathrm{A,B}}^t=\arctan\left|\frac{x_{\mathrm{A}}^t-x_{\mathrm{B}}^t}{y_{\mathrm{A}}^t-y_{\mathrm{B}}^t}\right| \tag{4}$$

4）设ϕ为A和B战机间在t时刻的最大可攻击方位角（恒定值）。

5）设δ_{A}、χ_{A}和λ_{A}分别为A战机机载火控雷达的扫描扇面张角、最大雷达探测距离和锁定时间（恒定值）。

6）设d_{A}、$\rho_{\mathrm{A}}^{\mathrm{s}}$和$\rho_{\mathrm{A}}^{\mathrm{h}}$分别为A战机空对空导弹的攻击距离、发射概率和命中概率（恒定值）。

当满足以下2个条件时，判定A战机摧毁B战机：1）A战机的机载火控雷达的扫描扇面持续覆盖B战机并达到锁定时间λ_{A}；2）A战机满足空对空导弹发射概率$\rho_{\mathrm{A}}^{\mathrm{s}}$，发射导弹满足命中概率$\rho_{\mathrm{A}}^{\mathrm{h}}$。

根据上述空战规则，本文给出了空战想定描述，如图2所示。

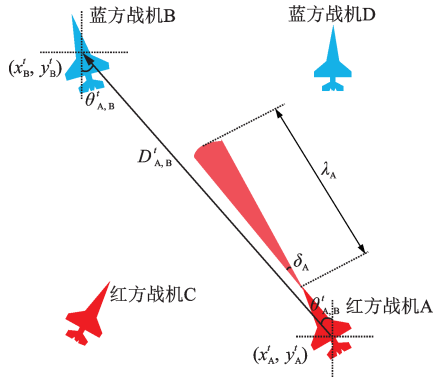


图2 空战想定描述

2.3 对抗策略训练流程

本文采用MADDPG算法对上述2对2空战想定进行建模。按照MADDPG算法的基本框架结构，环境中的每个智能体（红方战机）的训练网络分别由1个评论家网络和1个演员网络组成，其中演员网络根据自身观测值计算智能体动作，评论家网络用于评估演员网络获得动作的优劣。此外，算法还设置了经验回放缓存（ERB）[16]用于存储红方战机在环境中的训练经验，并以小样本抽样方式输入到每个智能体网络中以提升学习效率。MADDPG还在训练过程中引入了集中式训练分散式执行（CTDE）[14]训练框架，即每个智能体演员网络的输入只有自身观测值，在网络输入中的体现只有单个智能体（战机）的当前时刻状态$\boldsymbol{s}_t$，而评论家网络的输入则包括环境的全局信息，在网络输入中体现为环境中所有智能体的状态和动作信息，即$\boldsymbol{s}_t, \boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_N$，其中，$N$为一方战机数，本文想定中的$N=2$。评论家网络通过计算网络中的Q函数大小评价前者动作的优劣并指导其改进策略，本文仿真试验使用MADDPG算法中的梯度下降方法以求最优空战策略，其核心思想是通过迭代方式求得最高回报值对应的空战动作，该动作最终指导其智能体（战机）在不断迭代过程当中改进其空战策略。采用$\pi=(\pi(\theta_1), \pi(\theta_2), \cdots, \pi(\theta_N))$表明每个智能体的策略，其中，$\theta=(\theta_1, \theta_2, \cdots, \theta_N)$为网络参数。算法通过最小化损失函数求解每个智能体$i$的最优化策略，公式如下：

$$L_{\theta_i^Q}=\frac{1}{K}\sum_{t=1}^{K}(y_t-Q(\boldsymbol{s}_t, \boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_N, \theta_i^Q))^2 \tag{5}$$

其中，s_t为整个多智能体空战仿真环境在t时刻中的整体状态；$a_1, a_2, \cdots, a_N$为每个智能体（战机）在t时刻做出的动作。算法通过不断迭代方式最大化回报值，公式为：

$$y_t = r_t + Q_i'(s_{t+1}, a_1', a_2', \cdots, a_N', \theta_i^{Q'}) \tag{6}$$

其中，$\theta_i^{Q'}$为智能体i的目标评论家网络；$a_1', a_2', \cdots, a_N'$表明目标评论家网络的动作；r_t为t时刻的回报值。通过梯度下降方法更新演员网络参数公式如下：

$$\nabla_{\theta_i^\pi} L = \frac{1}{K}\sum_{t=1}^{K} \nabla_{\theta_i^\pi} \pi(o, \theta_i^\pi) \nabla_a Q(s, a_1, a_2, \cdots, a_N, \theta_i^Q) \tag{7}$$

其中，$\pi(o, \theta_i^\pi)$为演员网络；$Q(s, a_1, a_2, \cdots, a_N, \theta_i^Q)$为评论家网络；$\nabla_{\theta_i^\pi} L$表明损失函数朝着演员网络的方向下降。另外，MADDPG算法引入了目标网络架构以加速学习过程。这种双网络式算法可大幅度提升收敛效果，已经训练的稳定性和效率，目标网络架构更新方式如下：

$$\begin{cases} \theta_i^{Q'} = \tau\theta_i^Q + (1-\tau)\theta_i^{Q'} \\ \theta_i^{\pi'} = \tau\theta_i^\pi + (1-\tau)\theta_i^{\pi'} \end{cases} \tag{8}$$

其中，τ为控制更新频率的参数。

基于MADDPG算法的2对2对抗策略训练流程如3所示。

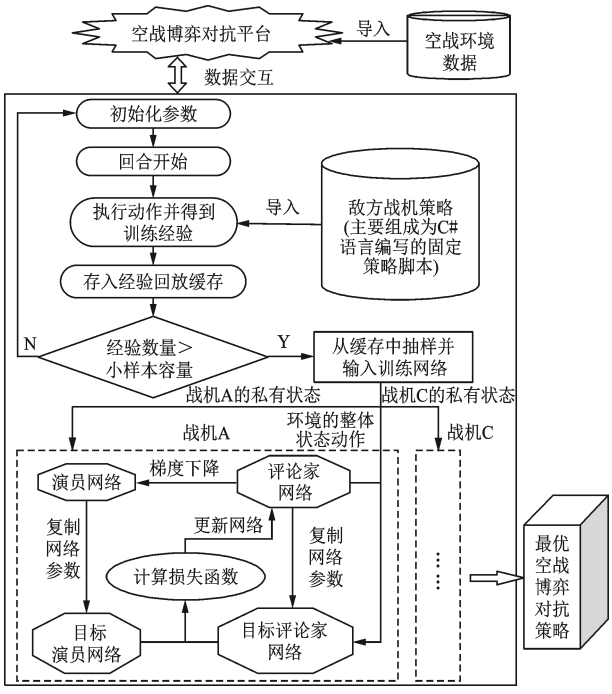
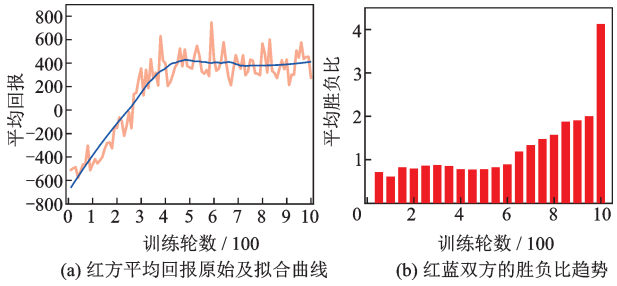


图3　基于MADDPG算法的2对2对抗策略训练流程

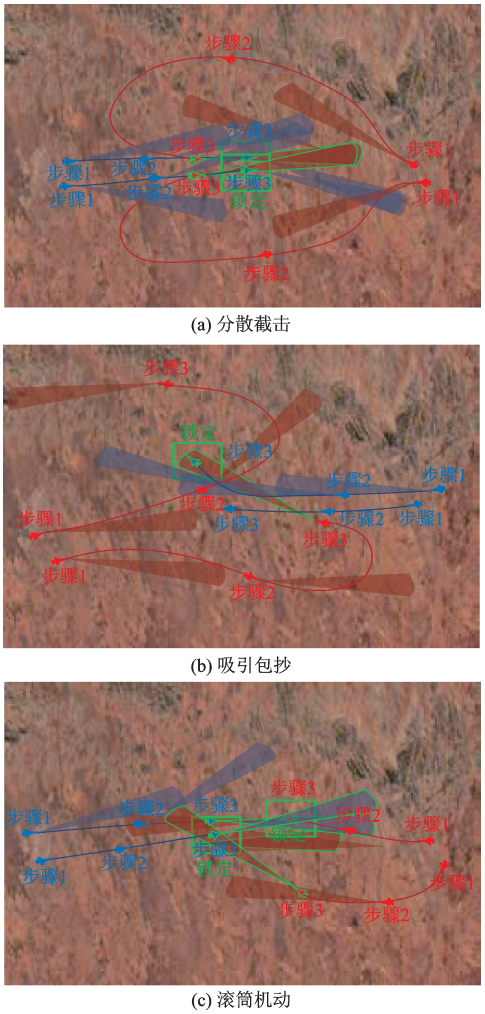
3　仿真试验与分析

本文以Ubuntu系统作为MADRL算法和空战平台的基本试验环境，Python3.5[17]作为开发语言，MADRL算法的运行机制基于TensorFlow1.7[18]。红蓝双方进行1 000轮对抗训练，其中，红方战机采用MADDPG算法，蓝方战机采用预编规则。1 000轮训练后红蓝双方重要参数对比如图4所示，可见，仿真训练在500轮后开始趋于收敛，在1 000轮后红方战机基本可以达到80%以上的胜率。通过对抗训练，红方战机学会3种空战配合歼击战术，其双机配合战术仿真轨迹如图5所示。



(a) 红方平均回报原始及拟合曲线　(b) 红蓝双方的胜负比趋势

图4　1 000轮训练后红蓝双方重要参数对比



(a) 分散截击

(b) 吸引包抄

(c) 滚筒机动

图5　3种战术的双机配合战术仿真轨迹

3种空域配合歼击战术如下：1）吸引包抄战术：指1架红方战机引诱1架蓝方战机离开编队，并与另一架红方战机围歼之；2）分散截击战术：指2架红方战机从1架蓝方战机的当面朝两侧散开，并从后方歼击之；3）滚筒机动战术：指红方战机躲避蓝方战机导弹后急速机动至其后方，并与同伴合力进行反击。

4 结束语

针对空战场景的博弈对抗团队策略生成问题，阐释了基于MAMDP的学习框架，提出了基于MADRL算法结合空战仿真场景的解决思路，并构建了一种基于MADDPG算法的2对2空战博弈对抗策略的训练模型。最后，对相关场景进行了模拟试验和分析。仿真试验表明，该训练模型可产生多种空战策略，具有较强的现实意义。后续将对如何实现仿真模型与人类飞行员实时交互和对抗进行研究。

参考文献(References)：

[1] 赵星宇，丁世飞．深度强化学习研究综述[J]．计算机科学，2018，45(7)：1-6.

[2] CHEN J X. The evolution of computing：AlphaGo[J]. Computing in Science & Engineering，2016，18(4)：4-7.

[3] CLARY K，TOSCH E，FOLEY J，et al. Let's play again：variability of deep reinforcement learning agents in Atari environments[EB/OL]. [2020-09-18]. https://people.cs.umass.edu/~kclary/NeurIPS_Critiquing_Trends_Workshop_2018_Variability_of_Deep_RL.pdf.

[4] DUAN Y，CHEN X，HOUTHOOFT R，et al. Benchmarking deep reinforcement learning for continuous control [EB/OL]. [2020-09-18]. https://arxiv. org/pdf/1604.06778.pdf.

[5] PIPATTANASOMPORN M，FEROZE H，RAHMAN S. Multi-agent systems in a distributed smart grid：design and implementation [C]//Proceedings of 2009 IEEE/PES Power Systems Conference & Exposition. New York：IEEE，2009：1-8.

[6] SON K，KIM D，KANG W J，et al. Qtran：learning to factorize with transformation for cooperative multi-agent reinforcement learning[C]//Proceedings of the 36th International Conference on Machine Learning. Long Beach：PMLR，2019：5887-5896.

[7] CHU T S，WANG J，CODECÀ L，et al. Multi-agent deep reinforcement learning for large-scale traffic signal control[J]. IEEE Transactions on Intelligent Transportation Systems，2019，21(3)：1086-1095.

[8] CELLI A，CICCONE M，BONGO R，et al. Coordination in adversarial sequential team games via multi-agent deep reinforcement learning [EB/OL]. [2020-09-18]. https://arxiv.org/pdf/1912.07712v1.

[9] 姚传明，王庆元，杨叶林．多平台协同作战任务系统建模[J]．指挥信息系统与技术，2017，8(3)：43-48.

[10] 宋艳波，何加浪，孙钧正，等．联合作战决心方案评估[J]．指挥信息系统与技术，2016，7(4)：49-54.

[11] HERNANDEZ-LEAL P，KARTAL B，TAYLOR M E，et al. A survey and critique of multi-agent deep reinforcement learning[J]. Autonomous Agents and Multi-Agent Systems，2019(33)：750-797.

[12] BUŞONIU L，BABUŠKA R，DE SCHUTTER B. Multi-agent reinforcement learning：an overview BT-innovations in multi-agent systems and applications[J]. Innovations in Multi-Agent Systems and Applications，2010(310)：183-221.

[13] 马骋乾，谢伟，孙伟杰．强化学习研究综述[J]．指挥控制与仿真，2018，40(6)：68-72.

[14] LOWE R，WU Y I，TAMAR A，et al. Multi-agent actor-critic for mixed cooperative-competitive environments[EB/OL]. [2020-09-18]. https://papers.nips.cc/paper/2017/file/68a9750337a418a86fe06c1991a1d64c-Paper.pdf.

[15] FOERSTER J N，FARQUHAR G，AFOURAS T，et al. Counterfactual multi-agent policy gradients [EB/OL]. [2020-09-18]. https://export.arxiv.org/pdf/1705.08926.

[16] TAMPUU A，MATIISEN T，KODELJA D，et al. Multi-agent cooperation and competition with deep reinforcement learning [EB/OL]. [2020-09-18]. https://pubmed.ncbi.nlm.nih.gov/28380078/.

[17] PEDREGOSA F，VAROQUAUX G，GRAMFORT A，et al. Scikit-learn：machine learning in Python[J]. The Journal of Machine Learning Research，2011(12)：2825-2830.

[18] ABADI M，BARHAM P，CHEN J，et al. Tensorflow：a system for large-scale machine learning[C]//Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation. Berkeley：USENIX Association，2016：265-283.

作者简介：

孙 彧，男(1993—)，助理工程师，研究方向为多智能体深度强化学习。

李清伟，男(1987—)，高级工程师，研究方向为指挥控制工程与智能空战博弈。

徐志雄，男(1994—)，工程师，研究方向为元深度强化学习。

陈希亮，男(1985—)，副教授，研究方向为指挥信息系统工程与深度强化学习。

(本文编辑：马 岚)