文章编号：2095-2163(2022)02-0174-04 中图分类号：TP181 文献标志码：A

基于群落学习的空中博弈对抗模型

沈贤杰

（中国电子科技南湖研究院 JS 大脑实验室，杭州 314000）

摘 要：近年来，许多强化学习模型取得了令人满意的成绩。然而，其大多数还要求有较大量的对战训练数据，否则很容易产生模型冷启动、过拟合等一系列问题。针对这些问题，该文针对空战环境，提出了一种更为稳定有效的空战环境下行动策略设计。在融合自注意力机制的同时改进了群落学习（Population-based Learning，PBT）在现有强化学习模型训练中的应用。本文设计模型 PSA-Air（Population-based Self-attention Air Combat Model），在尚未结束的 2021 首届全国空中智能博弈对抗大赛中取得了优秀的成绩。经实验证明，本文算法设计在收敛速度以及最终性能上具有一定的优越性。

关键词：强化学习；自注意力机制；群落学习

Air combat model based on population-based learning

SHEN Xianjie

（JS Brain Lab，China Nanhu Academy of Electronics And Information Technology，Hangzhou 314000，China）

【Abstract】 In recent years，many reinforcement learning model has achieved promising performance. However，most of them still require a large number of training data. Otherwise，problems such as cold starting or over-fitting are easily to occur. To solve these problems，this paper proposed a more stable and effective action strategy designed for the air combat environment. Self-attention mechanism is combined in this model，and population-based learning are adaptively designed. The model proposed in this paper has achieved great performance in the unfinished 2021 First National Aerial Intelligence Game Contest by getting in the top 16 among more than a hundred teams. Besides，the experiments in this paper also demonstrate the superiority in convergence speed and final performance.

【Key words】 reinforcement learning；self-attention mechanism；population-based learning

0 引 言

博弈对抗算法在现实生活中应用场景非常广泛，例如棋类、商业投标、作战等。对于棋类等存在大量的融合人类专家先验知识的局内数据作为训练数据的场景，即使不采用强化学习，只采用监督学习即可获得接近甚至超越普通人类的表现[1-2]。然而，对于其它一些难以获得大量实际数据的场景，现有表现较好的解决方案，是使用结合先验知识的强化学习模型进行自博弈对抗。

PSA-Air 模型首先针对空中博弈对抗的场景，设计了多阶段的强化学习模型训练，并将原先朴素的行动策略方案改进为一种更稳定有效的基于相对位置的行动方案；针对不同阶段的模型训练设计不同的群落学习机制，来解决模型训练的冷启动、过拟合等问题。此外，该模型利用 Transformer[3] 中的自注意力机制，对多智能体环境状态进行编码，实验证明相比 LSTM[4] 具有更高的性能。

1 环境描述

本文算法模型解决的问题环境为态势完全透明的 5V5 空战问题。双方均由 1 架有人机与 4 架无人机组成，双方性能完全对等；可行动空域长宽均为 300 km，高度约为 10 km 的矩形。初始状态双方各从空域俯视图的正方形一对顶角，同一高度同时出发，每架飞机各携 2 枚导弹。在限制时间 20 min 内，若一方无人机被击落或者全部导弹已被发射则判负；若超过限制时间，当前剩于战力（飞机总架数、导弹剩余总数量）多者获胜；若剩余战力相等，则占据对战空域中间部分时间较长的一方获胜。对战过程中，内部机群之间的机载雷达可以互相提供制导功能。

2 先验知识

2.1 群落学习

群落学习技术[5] 最初由 DeepMind 提出，用于挑选神经网络最优超参数。具体地说，多个被随机赋予超参数的神经网络模型并行地训练。类似于遗

作者简介：沈贤杰（1997-），男，硕士，研发工程师，主要研究方向：强化学习博弈对抗、自然语言处理。

收稿日期：2021-10-20

传学习[6]，在每轮训练中获得较好表现的网络超参数组合，会被用于改进现有的超参数组合，表现较差的超参数组合则会被放弃。在 AlphaGo[7] 中同样存在类似的思想，在自博弈阶段会初始化一系列不同参数的对手用于对抗学习，来防止训练阶段的过拟合问题。

2.2　自注意力机制

Transformer 在自然语言处理等多个领域取得了非常优秀的成绩，其主要归功于其对于自注意力机制的应用。假设输入为 $[m,n]$ 维的矩阵，则需要 3 个均为 $[n,d]$ 维的矩阵 $\boldsymbol{K}$、$\boldsymbol{Q}$、$\boldsymbol{V}$，分别代表 ***Key***、***Query*** 与 ***Value***，将输入矩阵转换成 $[m,d]$ 维的矩阵。输入矩阵经过 $\boldsymbol{Q}$ 矩阵得到其 ***Query*** 矩阵，将该矩阵与经过 $\boldsymbol{K}$ 矩阵得到的 ***Key*** 矩阵进行内积，再与经过 $\boldsymbol{V}$ 矩阵转换后的 ***Value*** 矩阵相乘后，得到处理后的输入。该机制主要意义在于将输入视为 m 个 n 维的向量，使向量之间进行交互，挖掘输入之间的关系，凸显更重要的输入维度。具体公式如式(1) 所示，分母中的 d 用于防止矩阵内积结果过大。

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}})V \qquad (1)$$

3　算法模型描述

3.1　模型结构与输入输出

PSA-Air 模型主要由 Critic 和 Actor[8] 组成，并遵循 MADDPG[9] 集中式评价、分布式训练的原则，使用经验回放池以及目标网络机制[10]。其中，Critic 接收整体环境的输入、编码得到当前环境的嵌入向量，并输出衡量当前优劣情况的值；Actor 接收各个行动体的局部环境，并输出对应的行动策略。

每架飞机的状态由一个 10 维的向量表示，分别为 X、Y、Z 轴方向的方位，表示飞机飞行角度的航向角、俯仰角、横滚角，以及纵向加速度、切向加速度、导弹的剩余量等。以上参数均被标准化到 0~1 之间，使训练过程更为高效稳定。若该飞机已被击落则全设为 0。每一架飞机由一个独立的 Actor 控制，以往模型对于 Actor 会直接输出三维空间下各个角度的偏转角以及加速度，但采用这种方式，在训练期间会有较多的不稳定性，且由于在训练初期飞机容易飞离指定空域，所以训练效率较低。

为了解决上述问题，Actor 模型利用飞机间的相对位置进行导航指向，将 Actor 的最后一层设为一个维度为 10 的 Softmax 层，分别表示向各个飞机的相对方位的移动权重。设 Softmax 层输出为 $o_i(i=1,2,\ldots,10)$，第 i 架飞机在 X 方位上的坐标为 x_i，则其在该时刻的移动目标点为 $\sum_{j=1}^{10} o_j x_j$。

Actor 和 Critic 接收输入后，都会经过全连接层连接的若干个自注意力层，将不同行动体的状态向量进行交互，再经过带非线性激活函数的全连接层进行编码。同时该模型避免了使用 LSTM 对历史数据进行编码存储，主要考虑到 LSTM 的训练速度慢且对于强化学习模型训练难度较高[11]。为了利用历史数据，模型的输入会同时得到该时刻以及上一时刻的数据，虽然输入层维度会翻倍，但大大降低了对计算资源的需求以及训练难度。

3.2　对战阶段设定

对于一局对战，主要分为开局阶段以及开火阶段。开局阶段定义为：双方机群之间，两架飞机之间距离大于 3 倍最大开火范围时，认为处于开局阶段。该阶段的主要任务是机群内部组成一个良好的队形，使其能够在很大程度上保护有人机，并且利于攻击敌方。经过开局阶段后，会进入对战开火阶段。主要表现为无人机之间的短兵相接以及有人机的适当介入。在两个阶段分别会使用不同的策略网络。

3.3　奖励函数设置

PSA-Air 模型主要使用如下 3 种奖励函数。

第一种：使用最终对战结果的胜负奖励。胜平负分别对应+1、0、-1。然而一场对局往往需要经过上百次行动决策，仅有终局奖励太过稀疏。

第二种：对于当前战力的消耗进行评估。若某一时刻无人机被击落，则会给予负向奖励 0.5；若导弹发射，但并未击落目标，也会给予负向奖励 0.16；反之，对另一方则会进行正向奖励。

第三种：奖励用于指导保护己方有人机以及攻击敌方无人机。具体来说，对每一个时刻 t 都会记录一个环境值，其值为己方有人机距对方最近无人机距离与对方有人机距己方最近无人机距离的比值。若该比值较大，则说明己方有人机处在相对更安全的位置(只考虑仍然携带剩余导弹的无人机)，反之则说明己方有人机有被击落的风险。t 时刻的该奖励为 t 时刻的比值与 $t-1$ 时刻比值的变化值。

3.4　训练流程

PSA-Air 模型的训练主要分为预热阶段与自博弈训练阶段。预热阶段包括整个模型群落的预热训练，自博弈训练为群落内的不同智能体之间进行对抗训练。

3.4.1 预热训练

模型预热训练阶段对战的是基于规则的模型。规则模型在开局阶段会让有人机在原地打转一定时间,其余无人机往敌方有人机飞行,这样可让有人机处在相对安全的位置又不至于脱离集群太远。当敌方有人机进入攻击范围,则会使用贪心法让最近距离的不在攻击状态的飞机攻击。若被敌方飞机攻击,则有一定概率放弃攻击,自主进行绕圈飞行躲避攻击。同时在每一步的行为中增加一定的随机性以提升鲁棒性。

3.4.2 融合群落学习的自博弈训练

模型自博弈阶段,会使用群落学习的概念,随机初始化一组策略网络,用于和经过预热训练的模型进行自博弈训练,并同时训练两边对战的模型。在以往使用群落学习的强化学习模型中,每一轮的对手策略网络都会被随机选择,然而这样训练的效率较低,会浪费许多训练资源。PopAir 模型提出运用上限置信区间公式(Upper Confidence Bound, UCB)[12]对策略网络群落进行采样。UCB 公式常被用于蒙特卡洛搜索树中的节点采样,以提升搜索效率。具体如式(2)所示:

$$P = v_i + c\sqrt{\frac{2\ln(\sum_i T_i)}{T_i}} \tag{2}$$

式中,P 表示每轮被挑选的概率;v_i 为该网络的对战胜率;T_i 表示各网络的对战次数。若该策略网络的对战胜率较高或参与对战的次数较少,则被挑选的概率越大。

3.4.3 训练细节

模型在预热及自博弈训练阶段,都使用时间差分误差(TD error)版本的策略梯度下降法,TD error 的具体定义如式(3)所示:

$$\delta_\theta(s,a,s') = R(s,a,s') + \gamma v_\theta(s',a) - v_\theta(s',a') \tag{3}$$

其中,R 是立即回报;γ 是折扣系数;v_θ 是价值网络的输出。

策略网络 μ 和价值网络 v 更新如式(4)、式(5)所示。

$$\nabla_\theta J(\mu) = E_{s,a,s',a'\sim D}[\nabla_\theta \mu(s,a)\delta_\theta(s,a,s')] \tag{4}$$

$$L(v) = E_{s,a,r,s',a'\sim D}[(v(s,a) - r - \gamma v(s',a'))^2] \tag{5}$$

其中,D 代表经验存储池。

4 实验结果分析

4.1 实验环境

本文实验在 Ubuntu 20.04 系统上进行,模型由 Pytorch 实现,训练流程使用单张 Quadro P4000 显卡。

4.2 基于 UCB 的采样算法分析

实验比较了本文基于 UCB 公式采样训练出来的模型与使用平均采样概率种群学习训练出来的模型之间的优劣,训练时间统一控制为前者自博弈训练 3 000 轮次后。共进行了 3 次训练,每次训练完的两个模型之间进行 100 局对战,综合胜负情况展示见表 1。当使用 UCB 公式对对手智能体进行采样时,训练所得的模型明显有更高的胜率。由表 1 中数据分析可知,由于 UCB 公式在有限的时间内能够更好的平衡各个对手智能体的对战权重,若当前模型对战某个随机初始化训练的模型胜率较低时,该公式则会鼓励当前模型多与该模型进行对战,尽快弥补缺点,因此提升了训练的效率。

表 1 对战平均采样群落学习 100 局表现

Tab. 1 Results of 100 games against average sampling PBL

次数	胜	负	平
1	78	16	6
2	75	21	4
3	74	22	4
平均	75.7	19.7	4.7

4.3 自注意力层分析

尝试将 LSTM 以及自注意力层进行替换,来验证在该问题中是否自注意力比 LSTM 更快地收敛并得到更优的解。如图 1 所示,使用 LSTM 的变体模型在训练时 Actor loss 的波动情况更加明显、更不稳定,且收敛速度更慢,最终收敛的 loss 值略大于使用自注意力层的模型。

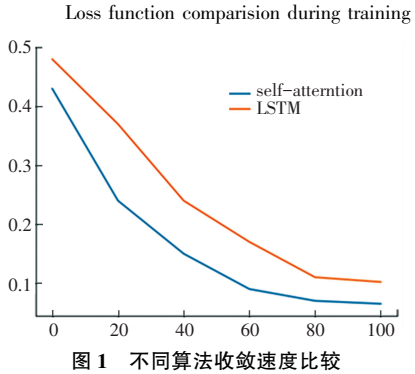

图 1 不同算法收敛速度比较

Fig. 1 Comparison of convergence rate with different algorithm

经分析得出,导致该问题的原因:一是 LSTM 结构的网络,在强化学习训练中训练难度较大,表现不稳定;二是在空战问题中,由于有较明显的开局、交火等不同阶段,行动策略有明显变化,而 LSTM 无法直接屏蔽上一阶段策略数据的影响,并且 LSTM 的

训练速度明显慢于使用近段数据进行自注意力机制交互对历史数据进行建模的方法。

5　结束语

本文针对空战博弈对抗问题提出了一种训练性能效率高，且性能优秀的强化学习模型 PSA－Air。该模型首先提出了一种基于智能体相对位置的行动方式，在处理环境输入时借鉴 Transformer 中的叠层自注意力机制，来进行各个智能体状态的交互解析。实验证明，PSA－Air 比直接使用 LSTM 进行解析有更快的收敛速度以及更好的表现。实验中，结合 UCB 公式的群落学习算法相比平均采样的变体更加适合于训练深度强化模型。

参考文献

[1] LEE G, LEE C, ROH B. Riverbed Modeler Reinforcement Learning M&S Framework Supported by Supervised Learning [C]//2021 International Conference on Information Networking (ICOIN). IEEE, 2021: 824-827.

[2] 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)：1620.

[3] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]//Advances in neural information processing systems. 2017: 5998-6008.

[4] ZHAO Z, CHEN W, WU X, et al. LSTM network: a deep learning approach for short－term traffic forecast [J]. IET Intelligent Transport Systems, 2017, 11(2): 68-75.

[5] PARKER-HOLDER J, PACCHIANO A, CHOROMANSKI K, et al. Effective Diversity in Population Based Reinforcement Learning [J]. arXiv preprint arXiv:2002.00632, 2020.

[6] 张韵，钟慧超，张春江，等. 基于机器学习的多策略并行遗传算法[J]. 计算机集成制造系统，2021.27(10)：2921-2928.

[7] SILVER D, HUANG A, MADDISON C J, et al. Mastering the game of Go with deep neural networks and tree search[J]. nature, 2016, 529(7587): 484-489.

[8] LOWE R, WU Y, TAMAR A, et al. Multi-agent actor-critic for mixed cooperative－competitive environments [J]. arXiv preprint arXiv:1706.02275, 2017.

[9] YUAN Y, LEI L, VU T X, et al. Actor－critic learning－based energy optimization for UAV access and backhaul networks [J]. EURASIP Journal on Wireless Communications and Networking, 2021, 2021(1): 1-27.

[10] LEE T Y, VAN BAAR J, WITTENBURG K, et al. Analysis of the contribution and temporal dependency of LSTM layers for reinforcement learning tasks[C]//CVPR Workshops. 2019: 99-102.

[11] XUE X, LI Z, ZHANG D, et al. A deep reinforcement learning method for mobile robot collision avoidance based on double dqn [C]//2019 IEEE 28th International Symposium on Industrial Electronics (ISIE). IEEE, 2019: 2131-2136.

[12] LI X, CAI Y, YU L, et al. A Modification of UCT Algorithm for WTN－EinStein würfelt nicht! Game [C]//2020 IEEE/CIC International Conference on Communications in China (ICCC). IEEE, 2020: 640-644.

(上接第 173 页)

客流智能预警方法，可快速确定车站发生突发大客流的瓶颈位置，有效指导地铁车站各瓶颈点突发大客流预案的制定及启动，对于保障城市轨道在大客流发生时的安全运营有重要的现实意义。

目前大多研究文件偏向于宏观预警，而本文提出的预警方案偏向微观预警，可以进一步研究建立宏观与微观相结合的更加便捷统一的预警指标体系。

参考文献

[1] 杨聚芬，王博，王奋，等. 地铁车站大客流预警系统分析[J]. 经贸实践，2018(14)：331.

[2] 赵保锋，邹晓磊，屈晓宜. 基于仿真的城市轨道交通站台客流滞留分级预警方法[J]. 城市轨道交通研究，2017，20(9)：107-110，115.

[3] 叶青，彭其渊. 考虑客流拥堵的城轨网络脆弱性评估[J]. 计算机应用研究，2016，33(10)：2923-2925，2945.

[4] 王雪梅，周立新，冯昊月. 城市轨道交通车站大客流预警及其疏解[J]. 城市建设理论研究(电子版)，2017(16)：208-209.

[5] 曹文超，干宏程. 基于 WiFi 数据的地铁车站客流预警模型[J]. 计算机工程与应用，2021，57(13)：233-238.

[6] 仇建华，尚凯，张亚岐，等. 基于相关向量机的城市轨道交通突发大客流预测[J]. 大连交通大学学报，2019，40(1)：13-17.

[7] 徐斌涛. 基于视频技术的轨道交通大客流检测方案[J]. 中国公共安全，2017(Z1)：71-73.

[8] 魏万旭，方勇，胡华，等. 基于视频数据挖掘的城市轨道交通车站行人交通行为特征提取系统研究[J]. 铁道运输与经济，2021，43(8)：119-125.

[9] 温念慈，倪少权，陈钉均，等. 城市轨道交通突发大客流协同应急决策研究[J]. 中国安全生产科学技术，2017，13(7)：48-54.

[10] ZHAO Dongdong, HU Xiaoyi. k－means clustering and kNN classification based on negative databases [J]. Applied Soft Computing Journal, 2021, 110.