DOI: 10.11992/tis.201809033
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20190416.1007.002.html

深度强化学习中状态注意力机制的研究

申翔翔[1], 侯新文[2], 尹传环[1]
（1. 北京交通大学 交通数据分析与挖掘北京市重点实验室, 北京 100044; 2. 中国科学院自动化研究所 智能系统与工程研究中心, 北京 110016）

摘　要: 虽然在深度学习与强化学习结合后, 人工智能在棋类游戏和视频游戏等领域取得了超越人类水平的重大成就, 但是实时策略性游戏星际争霸由于其巨大的状态空间和动作空间, 对于人工智能研究者来说是一个巨大的挑战平台, 针对 Deepmind 在星际争霸 II 迷你游戏中利用经典的深度强化学习算法 A3C 训练出来的基线智能体的水平和普通业余玩家的水平相比还存在较大的差距的问题。通过采用更简化的网络结构以及把注意力机制与强化学习中的奖励结合起来的方法, 提出基于状态注意力的 A3C 算法, 所训练出来的智能体在个别星际迷你游戏中利用更少的特征图层取得的成绩最高, 高于 Deepmind 的基线智能体 71 分。

关键词: 深度学习; 强化学习; 注意力机制; A3C 算法; 星际争霸 II 迷你游戏; 智能体; 微型操作
中图分类号: TP183　　**文献标志码**: A　　**文章编号**: 1673−4785(2020)02−0317−06

中文引用格式: 申翔翔, 侯新文, 尹传环. 深度强化学习中状态注意力机制的研究 [J]. 智能系统学报, 2020, 15(2): 317–322.
英文引用格式: SHEN Xiangxiang, HOU Xinwen, YIN Chuanhuan. State attention in deep reinforcement learning[J]. CAAI transactions on intelligent systems, 2020, 15(2): 317–322.

State attention in deep reinforcement learning

SHEN Xiangxiang[1], HOU Xinwen[2], YIN Chuanhuan[1]
(1. Beijing Key Laboratory of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing 100044, China; 2. Center for Research on Intelligent System and Engineering, Institute of Automation, Chinese Academy of Sciences, Beijing 110016, China)

Abstract: Through artificial intelligence, significant achievements beyond the human level have been made in the field of board games and video games since the emergence of deep reinforcement learning. However, the real-time strategic game StarCraft is a huge challenging platform for artificial intelligence researchers due to its huge state space and action space. Considering that the level of baseline agents trained by DeepMind using classical deep reinforcement learning algorithm A3C in StarCraft II mini-game is still far from that of ordinary amateur players, by adopting a more simplified network structure and combining the attention mechanism with rewards in reinforcement learning, an A3C algorithm based on state attention is proposed to solve this problem. The trained agent achieves the highest score, which is 71 points higher than Deepmind's baseline agent in individual interplanetary mini games with fewer feature layers.
Keywords: deep learning; reinforcement learning; attention mechanism; A3C; StarCraft II mini-games; agent; micromanagement

近年来, 由于硬件的发展, 计算资源的增加, 深度学习在人工智能领域崛起。利用深度学习可以从高维原始数据中提取高层特征, 研究者们不再受手工选取特征的影响, 进而使得图像检测、语音识别和自然语言处理等领域的研究水平达到新高度[1]。

在强化学习和深度学习结合起来之后也获得了质的飞跃, 促进了游戏、机器人、金融管理、健康医疗和智慧交通等领域的发展。引人注意的是深度强化学习深度 Q 网络在 Atari 游戏上的应用取得了重大的突破, 达到人类水平[2]。深度强化

收稿日期: 2018−09−17.　**网络出版日期**: 2019−04−17.
基金项目: 中央高校基本科研业务费专项资金项目 (2018JBZ006); 国家自然科学基金项目 (61105056).
通信作者: 尹传环. E-mail: chyin@bjtu.edu.cn.

学习在人工智能领域的作用日益显著,“围棋专家”AlphaGo 到 AlphaZero 的水平也远远超过甚至碾压人类水平[3]。AlphaGo 和 AlphaZero 的主要贡献者 David Silver 在他的教程中明确指出人工智能就是深度学习与强化学习,并决定将其带领的 Deepmind 团队的重点研究转移到难度更大的实时策略性游戏星际争霸上,因此掀起了人工智能领域的另一波研究热潮。

星际争霸是一个微观操作和宏观计划相结合的战争对抗性实时策略游戏,游戏玩家在大面积而且部分信息可见的环境中必须学会控制大量的游戏单元以发展经济,建造建筑,建设军队,从而能够为获取战争的胜利打下坚实的基础。星际争霸状态空间和动作空间是十分巨大的,截至目前,学者们直接在整个游戏上研究是十分困难的,现在的研究主要集中在一些经典场景的微型操作中,期望成为研究整个游戏人工智能的基石。Deepmind 团队与游戏星际争霸的拥有者暴雪公司合作将星际Ⅱ发展成为研究人工智能的环境,并在文献 [4] 中详细介绍了星际争霸Ⅱ的学习环境 SC2LE(StarCraft Ⅱ Learning Environment) 并且还针对星际争霸Ⅱ中的迷你游戏运用经典的深度强化学习算法 A3C(Asynchronous Advantage Actor-Critic) 训练出一些基线智能体。事实上,在 Deepmind 团队决定研究星际争霸之前,其他研究者们在星际争霸上的研究工作就进行了很多年[5],只不过绝大部分的研究工作主要集中在星际争霸Ⅰ而不是星际争霸Ⅱ上,基于不同的机器学习算法在不同的方面进行研究并取得了一些成果,例如,对星际争霸中的战争结果进行估计[6-7],在宏观上进行管理[8],对智能体行为的可解释性进行研究[9-10],以及将强化学习应用到微操场景中去[11-13]。

注意力机制是自深度学习发展之后广泛应用在自然语言处理、图像检测、语音识别等领域的核心技术。神经网络中的注意机制[14] 是基于人类的视觉注意机制提出的,虽然存在不同的模型,但它们都基本上归结为能够以高分辨率聚焦在图像的某个区域,同时以低分辨率感知周围的图像区域,然后不断调整关注点。近年来注意力机制也开始被应用于循环神经网络 (recurrent neural networks)[15-16],主要涉及自然语言处理和图像检测等领域,主要思想是解码器在每一时间步中都能够关注到源输入序列的不同位置,重点是注意力模型可以关注到目前已经学习到的内容以及学习下一步应该主要关注的内容。本文创新灵感的来源主要是:在强化学习决策序列过程中,智能体需要关注到输入状态序列中有价值的状态。以星际争霸 II 中的 3 个经典迷你游戏作为测试平台,它们分别是战胜跳虫和毒爆虫 (DefeatZerglingsAndBanelings),奔向烽火处 (MoveToBeacon) 以及收集矿物碎片 (CollectMineralShards),在这些小游戏中,智能体只包含 4 种动作行为,即上下左右,这样就可以大大缩小动作空间。利用星际争霸 II 学习环境中提供的接口,可以获取很多状态特征,比如非空间特征和空间特征图层,进而来训练智能体,但是经过分析这 3 个迷你游戏发现智能体获取如此多的空间特征是没有必要的,因此我们挑选了部分特征,去掉冗余特征,以加快智能体的学习速度。

根据以上所谈论到的问题以及注意力机制的优势,本文的主要贡献包括到以下两个方面:

1) 采用的网络结构比 Deepmind 提供的基线智能体的网络结构更加简洁。

2) 将强化学习中的奖励与注意力机制结合起来,每一个时间步,智能体更加关注有价值的游戏状态。

通过以上两个方面相结合,不仅加速了智能体在星际争霸 II 中的学习速度,也使得智能体学习到更优的策略,取得更好的成绩。

1　强化学习

在本节中,首先回顾一下经典的强化学习场景以及算法 A3C[17]。

经典的强化学习场景中的一些基本概念描述如下:在某一时刻 t,智能体根据当前环境的状态 s_t 以及策略 π 发送动作信号 a_t 与环境交互,并且根据环境返回的状态信息 s_{t+1} 与奖励 $r_t = r(s_t, a_t)$ 信息不断更新自己的策略 π,获取的累计收益表示为

$$R_t = \sum_{k=0}^{\infty} \gamma^k r(s_{t+k}, a_{t+k})$$

其中 $\gamma \in (0,1)$ 表示折扣因子。智能体的目标可以表示为

$$\max_{\pi} E_{s\sim P_0}[R_t | s_t = s]$$

P_0 为状态 s 的先验分布,动作值函数表示为

$$Q^{\pi}(s,a) = E_{\pi}[R_t | s_t = s, a_t = a]$$

表示在状态 s 下根据策略 π 选择动作 a 的期望累计回报。同样的,状态值函数表示为

$$V^{\pi}(s) = E_{\pi}[R_t | s_t = s]$$

表示在策略 π 下状态 $s_t = s$ 的期望累计回报。

A3C 算法是将策略函数和价值函数相结合的强化学习方法,对目标函数式 (1):

$$J_{\theta}(s) = E_{\pi_{\theta}}[R_t | s_t = s] \tag{1}$$

运用梯度上升的方法以不断更新现有的策略 π_θ 的参数 θ, 期望获得使目标收益能够达到最大的最优策略, 则关于策略参数 θ 的梯度公式为

$$\nabla_\theta J_\theta(s)=E_{\pi_\theta}[\nabla_\theta \ln\pi_\theta(a_t|s_t)R_t|s_t=s]$$

但是只采用这种方式存在高方差的缺点, 因此 Williams 等[18] 提出了改进版本:

$$\nabla_\theta J_\theta(s)=E_{\pi_\theta}[\nabla_\theta \ln\pi_\theta(a_t|s_t)[R_t-b(s_t)]|s_t=s]$$

通常情况下,$Q^\pi(s_t,a_t)$ 与 $V^\pi(s_t)$ 来代替 R_t 和 $b(s_t)$, 所以梯度也可以表示为

$$\nabla_\theta J_\theta(s)=E_{\pi_\theta}\left[\ \nabla_\theta \ln\pi_\theta(a_t|s_t)[Q^\pi(s_t,a_t)-V^\pi(s_t)]|s_t=s\ \right]$$

其中 $A^\pi(s_t,a_t)=Q^\pi(s_t,a_t)-V^\pi(s_t)$ 表示优势函数。

基于价值函数的方法则主要是与环境进行交互后, 对动作值函数和价值函数进行估计, 然后获取较优的策略或者是促进策略优化, 在 A3C 算法中主要采用后者, 一般只对价值函数进行估计, 通常最小化此损失函数:

$$[G^\pi(s_t)-V^\pi_{\theta'}(s_t)]^2 \tag{2}$$

其中,

$$G^\pi(s_t)=E_{\pi_\theta}\left[\sum_{k=0}^{n}\gamma^k r(s_{t+k},a_{t+k})+\gamma^{n+1}V^\pi_{\theta'}(s_{t+n+1})\right]$$

为了提升策略的探索度, 通常在 A3C 算法中加入熵正则化项 H, 则 A3C 算法最大化目标函数为

$$J_\theta(s_t)=E_{\pi_\theta}[A^\pi(s_t,a_t)]+\delta H[\pi_\theta(s_t)] \tag{3}$$

其中 δ 为超参数, 其梯度公式为

$$\nabla_\theta J_\theta(s_t)=E_{\pi_\theta}[\nabla_\theta \ln\pi_\theta(a_t|s_t)[Q^\pi(s_t,a_t)-V^\pi(s_t)]+\delta\nabla_\theta H[\pi_\theta(s_t)]]$$

A3C 算法同时为了提高学习的稳定性并且加快学习速度, 利用异步的方法, 将多个智能体在不同的线程中运行, 共同更新一个策略网络。

2　基于状态注意力的 A3C 算法

在深度强化学习中, 是否采用游戏环境默认的奖励是一个值得探讨的问题, 即便是一些经典算法被提出以后, 在实际的源码实现中也对环境原始的奖励进行了缩放[19]。因此, 本文认为, 原始环境中定义的奖励只是起到了一定的基础作用, 并未真正体现出各个游戏状态的相对重要性, 为了让智能体学会关注更有价值的游戏状态, 引入了权重网络 w_ϑ, 为每个时刻下的奖励赋予不同的权值, 此时累计回报便表示为

$$R_t=\sum_{k=0}^{\infty}\gamma^k w_\vartheta(s_{t+k})r(s_{t+k},a_{t+k})$$

当引入权重网络 w_ϑ 后, $Q^\pi(s_t,a_t)$ 与 $V^\pi(s_t)$ 仍然满足贝尔曼方程式 (4) 和式 (5):

$$Q^\pi(s_t,a_t)=w_\vartheta(s_t)r(s_t,a_t)+\gamma Q^\pi(s_{t+1},a_{t+1}) \tag{4}$$

$$V^\pi(s_t)=E_{\pi_\theta}[w_\vartheta(s_t)r(s_t)+\gamma V^\pi(s_{t+1})] \tag{5}$$

由此可以看出, 此算法和 A3C 算法是很相似的。所以基于注意力机制的 A3C 算法最大化的目标函数分别为式 (3)、式 (6), 最小化目标函数为式 (2):

$$J'_{w_\vartheta}(s_t)=-\left[G^\pi_{w_\vartheta}(s_t)-V^\pi(s_t)\right]^2 \tag{6}$$

其中,

$$G^\pi_{w_\vartheta}(s_t)=E_{\pi_\theta}\left[\sum_{k=0}^{n}\gamma^k w_\vartheta(s_{t+k})r(s_{t+k},a_{t+k})+\gamma^{n+1}V^\pi(s_{t+n+1})\right]$$

其式 (2) 则主要体现在不断调整价值网络的参数, 使价值网络更靠近于真实的价值网络, 式 (5) 则主要体现在通过不断调整权重网络 w_ϑ 缩短真实的价值网络与训练过程中价值网络的差距, 进一步可以体现出通过对奖励的不同缩放来使训练过程中价值网络的更贴近真实的价值网络, 式 (5) 关于注意力权重网络 w_ϑ 的参数 ϑ 的梯度可以表示为

$$\nabla_\vartheta J'_{w_\vartheta}(s_t)=-2\left[G^\pi_{w_\vartheta}(s_t)-V^\pi(s_t)\right]\nabla_{w_\vartheta}G^\pi_{w_\vartheta}(s_t)\nabla_\vartheta w_\vartheta(s_t)$$

3　实验验证

在本节中,将本文提出的基于注意力机制的 A3C 算法在实时策略性游戏星际争霸 II 中的迷你游戏上进行实验验证网络结构与算法的有效性, 有关于战胜跳虫和毒爆虫、奔向烽火处和收集矿物碎片这 3 个小游戏的具体描述如下:

战胜跳虫和毒爆虫: 最初状态下, 在地图的两侧分别有 9 个陆战队员和 10 个虫子 (6 个跳虫和 4 个毒爆虫), 当任何一个跳虫和毒爆虫被陆战队员消灭, 智能体都会获得奖励, 当所有的跳虫和毒爆虫被消灭, 又会恢复到刚开始的 10 个, 此时也会额外增加 4 个满血状态的陆战队员, 其他陆战队员的血量还是保持原来的样子。与此同时虫子和陆战队员的位置会被重置到地图的两侧。

奔向烽火处: 地图上有一个烽火标记和一个陆战队员, 当陆战队员到达烽火标记的位置智能体就会获得奖励, 同时, 烽火的位置会重新设置。

收集矿物碎片: 地图上有两个陆战队员和 20 个分散在屏幕各处的矿物碎片, 当任何一个陆战队员移动到矿物碎片处智能体都会获得奖励, 当然最优的策略应该是两个陆战队员独立行动, 分开收集矿物, 当所有的矿物被收集完之后, 地图会继续随机生成 20 个矿物碎片。

更多关于星际争霸 II 迷你游戏的细节, 请参考文献 [20]。

3.1　网络结构

本文的学习环境与测试环境是基于 Deepmind 和暴雪合作的 SC2LE, 网络结构与传统的网络结构非常相似。如图 1 所示, 我们利用很简单的三层卷积神经网络和一层全连接网络, 将 SC2LE 中提供的部分屏幕特征图层 (单元类型、已选择、生命值) 输入到网络里, 3 个卷积层的过滤器的个数分别是 32、64、64, 大小分别是 8、4、3, 步长分别是 4、2、1, 每一层有 RELU 激活函数, 在全连接层中有 512 个隐层单元和 RELU 激活函数, 网络有 3 个输出, 分别输出策略、价值和基于注意力机制的 A3C 算法中的注意力权重, 我们使用 RMSProp 优化器, 每次网络输入量的大小为 32 批。实验具体硬件环境的条件是拥有 8 GB 显存的 GPU、16GB 内存以及 8 核 4.2 GHz 的 CPU。

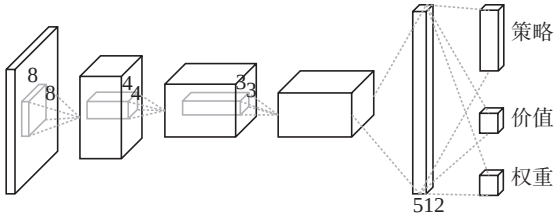

图 1　网络结构
Fig. 1　Network architecture

3.2　实验结果验证

从本节开始对游戏场景的介绍中可以知道, 智能体玩好游戏的第一步应该关注的是如何选择陆战队员。在游戏战胜跳虫和毒爆虫中, 有 10 个陆战队员, 首先选择哪一个陆战队员对他发出命令, 是否应该完全区别对待这 10 个陆战队员, 是一个值的考虑的问题, 本文认为随机选择陆战队员是一个不错的决策, 随机的选择意味着不再区分陆战队员, 所有的陆战队员将采取同一个策略, 增加了策略的鲁棒性。比如, 在图 2 中, 随机选择 2 个相同状态的陆战队员交换他们的位置之后的状态其实和交换之前的状态是完全一样的, 所以, 随机选择的策略意味着缩小了状态空间, 从实验过程中可以进一步发现, 随机的选择有利于陆战队员分散开来, 这种行为也有利于陆战队员击败虫子。从上面两种情况中可以看出, 在游戏战胜跳虫和毒爆虫中随机选择是一个不错的策略, 事实上, 虽然在游戏收集矿物碎片中也每次随机选择一个陆战队员执行命令, 但是这并不是一个明智的决策。比如, 在某一段时间内, 很有可能会出现一个陆战队员在忙碌地收集矿物碎片, 而另一个陆战队员却一直在等待的情况。

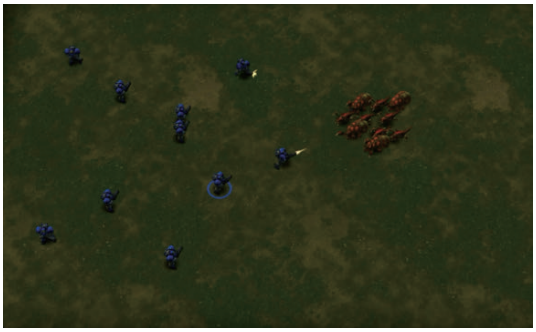

图 2　战胜跳虫和毒爆虫游戏界面截图
Fig. 2　The screenshot of DefeatZerglingsAndBanelings

为了保证游戏智能体与人类的成绩相比较时操作速度是相当的, 即对于人类是一场公平的竞争。Deepmind 在整个游戏实验中每 8 帧执行一个动作, 而在战胜跳虫和毒爆虫整个游戏中, 因为每次随机选择一个陆战队员, 然后对陆战队员发送命令。然而从 SC2LE 提供的程序接口 pysc2 中发现, 如果每次都要随机选取一个陆战队员, 那么每次选择陆战队员之前, 就必须选择全部陆战队员然后从这些陆战队员中选择其中的一个。因此每次这种操作就会让陆战队员速度变慢, 所以在战胜跳虫和毒爆虫实验中选择每 4 帧执行一个动作, 这种方式是很合理的, 毕竟人类并不会在选择某一个陆战队员之前选择全部的陆战队员。

实验结果如表 1 所示, 基于注意力机制的 A3C 算法的性能表现不错, 与目前 Deepmind 提供的基准智能体 ATARI 网络的分数相比较在战胜跳虫和毒爆虫的迷你游戏中得分显著提高。

表 1　人类和智能体获取的平均分数表
Table 1　Averaged results for human baselines and agents

人与智能体	战胜跳虫和毒爆虫	奔向烽火处	收集矿物碎片
DEEPMIND业余选手	729	26	133
星际争霸职业玩家	727	28	177
DEEPMIND随机策略	23	1	17
DEEPMIND的ATARI网络	81	**25**	96
随机策略	37	1.5	16
基于注意力机制的A3C算法	**152**	22	**97**

从表 1 可以看出, 随机策略在战胜跳虫和毒爆虫、收集矿物碎片迷你游戏中的平均成绩要比 Deepmind 随机策略的平均成绩要高一些, 由此可见, 虽然本文的网络结构比 Deepmind 的 Atari 网络结构简单一些, 但是对于星际中的这 3 个游戏场景来说, 简单的网络结构更适合。在游戏奔向

烽火处中基于注意力机制的 A3C 算法的成绩并没有比 Deepmind 的 Atari 网络的成绩高, 经过分析原因后发现, 由于基于注意力机制的 A3C 算法的智能体的可选择方向只包含上下左右, 所以陆战队员不能直线到达目标位置, 但是陆战队员所走的路线就是在规定方向的基础上的最短路径, 而通过 Deepmind 发表的视频中可以发现, 在这个小游戏上, 游戏智能体直接定位到目标位置, 陆战队员可以沿直线走过去, 在这个小游戏上也许是一个好的办法, 但是如果游戏中添加了障碍物, 也许这就不是一个好的方法了。虽然在战胜跳虫和毒爆虫的游戏分数上基于注意力机制的 A3C 算法取得了较大的提高, 但是与人类水平相比还存在较大的差距, 这也意味着还存在较大的空间值得我们研究与探索。

4 结束语

本文认为不同的游戏状态或者游戏帧有不同的重要性, 智能体理应关注更有价值的状态, 因此本文提出了基于注意力机制的 A3C 算法, 由此将注意力机制和强化学习中的奖励结合起来, 得到了一定的进步, 但是智能体比起人类水平还是存在较大差距, 深度强化学习的应用, 虽然在很多游戏上取得了成功, 但是在实时策略游戏上还面临很大的挑战。在战胜跳虫和毒爆虫迷你游戏中, 本文做法也存在不足之处: 1)人类不会采用随机选择陆战队员这样的策略, 比如, 大部分玩家会优先选择让受伤的陆战队员后退然后远距离攻击敌人, 而不是站在那里被敌人杀死。2)系统预先给定好的奖励是否是有利于深度强化学习算法进行学习的最优奖励, 这是不确定的, 应该采用一定的策略来优化这个默认的奖励。以上两点也是我们未来工作考虑的两个方面。

参考文献:

[1] LI Yuxi. Deep reinforcement learning: an overview [EB/OL]. [2018-01-17]https://arxiv.org/abs/1701.07274.

[2] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529–533.

[3] 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.

[4] VINYALS O, EWALDS T, BARTUNOV S, et al. StarCraft II: a new challenge for reinforcement learning [EB/OL]. [2018-01-17]https://arXiv: 1708.04782, 2017.

[5] ONTANON S, SYNNAEVE G, URIARTE A, et al. A survey of real-time strategy game AI research and competition in StarCraft[J]. IEEE transactions on computational intelligence and AI in games, 2013, 5(4): 293–311.

[6] SYNNAEVE G, BESSIERE P. A dataset for StarCraft AI & an example of armies clustering[C]//Artificial Intelligence in Adversarial Real-Time Games. Palo Alto, USA, 2012: 25–30.

[7] SYNNAEVE G, BESSIÈRE P. A Bayesian model for opening prediction in RTS games with application to StarCraft[C]//Proceedings of 2011 IEEE Conference on Computational Intelligence and Games. Seoul, South Korea, 2011: 281–288.

[8] JUSTESEN N, RISI S. Learning macromanagement in starcraft from replays using deep learning[C]//Proceedings of 2017 IEEE Conference on Computational Intelligence and Games. New York, USA, 2017: 162–169.

[9] DODGE J, PENNEY S, HILDERBRAND C, et al. How the experts do it: assessing and explaining agent behaviors in real-time strategy games[C]//Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. Montreal QC, Canada, 2018.

[10] PENNEY S, DODGE J, HILDERBRAND C, et al. Toward foraging for understanding of starcraft agents: an empirical study[C]//Proceedings of the 23rd International Conference on Intelligent User Interfaces. Tokyo, Japan, 2018: 225–237.

[11] PENG Peng, WEN Ying, YANG Yaodong, et al. Multiagent bidirectionally-coordinated nets for learning to play starcraft combat games[EB/OL]. [2018-01-17]https://arXiv: 1703.10069, 2017.

[12] SHAO Kun, ZHU Yuanheng, ZHAO Dongbin, et al. StarCraft micromanagement with reinforcement learning and curriculum transfer learning[J]. IEEE transactions on emerging topics in computational intelligence, 2019, 3(1): 73–84.

[13] WENDER S, WATSON I. Applying reinforcement learning to small scale combat in the real-time strategy game StarCraft: Broodwar[C]//Proceedings of 2012 IEEE Conference on Computational Intelligence and Games. Granada, Spain, 2012: 402–408.

[14] DENIL M, BAZZANI L, LAROCHELLE H, et al. Learning where to attend with deep architectures for image tracking[J]. Neural computation, 2012, 24(8): 2151–2184.

[15] BAHDANAU D, CHO K, BENGIO Y, et al. Neural machine translation by jointly learning to align and

translate[C]//Proceedings of International Conference on Learning Representations. 2015.

[16] MNIH V, HEESS N, GRAVES A, et al. Recurrent models of visual attention[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada, 2014: 2204–2212.

[17] MNIH V, BADIA A P, MIRZA M, et al. Asynchronous methods for deep reinforcement learning[C]//Proceedings of the 33rd International Conference on Machine Learning. New York USA, 2016: 1928-1937.

[18] WILLIAMS R J. Simple statistical gradient-following algorithms for connectionist reinforcement learning[J]. Machine learning, 1992, 8(3/4): 229–256.

[19] ILYAS A, ENGSTROM L, SANTURKAR S, et al. Are deep policy gradient algorithms truly policy gradient algorithms? [EB/OL]. [2018-01-17]https://arXiv: 1811.02553, 2018.

[20] DeepMind. DeepMind mini games[EB/OL]. (2017-08-10)[2018-09-10]. https://github.com/deepmind/pysc2/blob/master/docs/mini_games.md.

作者简介：

申翔翔，硕士研究生，主要研究方向为深度强化学习。

侯新文，项目研究员，主要研究方向为人脸检测和识别、机器学习、强化学习和博弈对抗。发表学术论文 40 余篇，Google Scholar 1 000 多次。

尹传环，副教授，主要研究方向为网络安全 (入侵检测)、数据挖掘、机器学习。

第四届亚洲人工智能技术大会
The 4th Asian Conference on Artificial Intelligence Technology

由中国人工智能学会、重庆市大数据应用发展管理局、重庆理工大学、重庆市巴南区人民政府联合主办，重庆理工大学期刊社、重庆市巴南区科学技术局、重庆市巴南区大数据应用发展管理局、重庆两江人工智能学院联合承办，重庆市人工智能学会协办，重庆市科学技术协会指导的“第四届亚洲人工智能技术大会（ACAIT 2020）”将作为 2020 年中国智能产业博览会期间的唯一国际学术会议在重庆召开。

征稿范围（但不局限于）：

1）人工智能理论基础；2）人工智能应用；3）模式识别；4）机器感知与虚拟现实；5）自然语言处理和机器翻译；6）图像和语音处理；7）计算机视觉；8）神经网络与计算智能；9）知识科学与知识工程；10）生物信息学与人工生命；11）机器学习；12）深度学习及其应用；13）数据挖掘；14）面向大数据的人工智能技术；15）智能控制与智能管理；16）粗糙集与软计算；17）智能搜索；18）智能推理；19）智能规划；20）智能信息处理；21）智能制造；22）智能机器人；23）物联网；24）工业互联网；25）智能通信与网络；26）人机交互/普适计算；27）智慧能源；28）自动程序设计。

联系方式：

联系人：贺柳、徐佳忆

电话：023-62561406

邮箱：cqznjs@126.com；xb@cqut.edu.cn