计算机应用
Journal of Computer Applications
ISSN 1001-9081,CN 51-1307/TP

《计算机应用》网络首发论文

题目：基于奖励高速路网络的多智能体强化学习中的全局信用分配算法
作者：姚兴虎，谭晓阳
收稿日期：2020-05-31
网络首发日期：2020-10-09
引用格式：姚兴虎，谭晓阳．基于奖励高速路网络的多智能体强化学习中的全局信用分配算法[J/OL]．计算机应用.
https://kns.cnki.net/kcms/detail/51.1307.TP.20200930.1727.004.html

Journal of Computer Applications ISSN 1001-9081
计算机应用 CODEN JYIIDU http://www.joca.cn

文章编号:1001-9081(****)**-0000-00 doi:10.11772/j.issn.1001-9081.2020061009

基于奖励高速路网络的多智能体强化学习中的全局信用分配算法

姚兴虎[1,2,3]，谭晓阳[1,2,3*]

(1. 南京航空航天大学计算机科学与技术学院，南京 211106；

(2. 南京航空航天大学模式分析与机器智能工业和信息化部重点实验室，南京 211106；

(3. 南京航空航天大学软件新技术与产业化协同创新中心， 南京 211106)

(*通信作者电子邮箱 x.tan@nuaa.edu.cn)

摘 要: 针对多智能体系统中面临的联合动作空间随智能体数量的增加指数爆炸的问题，采用“中心训练-分散执行的架构”来解决联合动作空间的维数灾难及降低算法的优化代价。针对众多多智能体强化学习场景下，环境仅给出所有智能体的联合行为所对应的全局奖励这一问题，提出一种新的全局信用分配机制——奖励高速路网络。通过在原有算法的奖励分配机制上引入奖励高速路连接，将每个智能体的值函数与全局奖励直接建立联系，进而使得每个智能体在进行策略选择时能够综合全局的奖励信号与其自身实际分得的奖励值。 首先，在训练过程中，通过中心化的值函数结构对每个智能体进行协调；此外，这一中心化的结构同时也能起到全局奖励分配的作用；然后，在中心值函数结构中引入奖励高速路链接来辅助进行全局奖励分配，从而构建出奖励高速路网络；之后，在执行阶段，每个智能体的策略仅仅依赖于其自身的值函数。星际争霸微操作平台上的实验结果表明，本文所提出的奖励高速路网络能够取得很好的性能提升，相比当前最先进的反直觉的策略梯度算法(Coma)和单调 Q 值函数分解(Qmix)算法，在四个复杂的地图上的胜率能够提升超过 20%。更重要的是，在智能体数量较多且种类不同的 3s5z 和 3s6z 场景中，在所需样本数量为 Qmix，Coma 等算法 30%的情形下便能取得更好的结果。

关键词: 深度学习；深度强化学习；多智能体强化学习；多智能体系统；全局信用分配

中图分类号: TP181 **文献标志码**：A

Reward Highway Network Based Global Credit Assignment Algorithm in Multi-agent Reinforcement Learning

Xinghu Yao[1,2,3], Xiaoyang Tan[1,2,3*]

(1. *College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106;*

(2. *MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, Nanjing University of Aeronautics and Astronautics, Nanjing 211106;*

(3. *Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing University of Aeronautics and Astronautics，Nanjing 211106*)

Keywords: In order to solve the problem of the exponential explosion of joint action space with the increase of the number of agents in multi-agent systems, the "central training-decentralized execution" was adopted to solve the dimensional disaster of joint action space and reduce the optimization cost of the algorithm. A new global credit allocation mechanism, reward highway network, was proposed to solve the problem that the global reward is given only for the joint behavior of all agents in multi-agent reinforcement learning scenarios. By introducing the reward highway connection in the global credit assignment mechanism, the value function of each agent was directly connected with the global reward, so that each agent can use the global reward signal and its own credit when making decisions. Firstly, in the training process, each agent was coordinated through a centralized value function. In addition, this centralized structure can also play a role in global credit assignment; Then, the reward highway was introduced in the central value function structure to assist the global credit assignment, thus a new reward highway network was proposed. Then, in the execution phase, each agent's strategy depends only on its own value function; Experimental results on the StarCraft Multi-agent Challenge show that the reward highway network achieves a performance improvement of more than 20% over four complex maps compared to the advanced

收稿日期: 2020-05-31; 修回日期:2020-09-24; 录用日期: 2020-09-28。

基金项目：国家自然科学基金(61976115,61672280, 61732006)；装备预研基金(6140312020413)；南京航空航天大学人工智能+项目(56XZA18009)；全军共用信息系统装备预研(315025305)；南京航空航天大学研究生创新基金((Kfjj20191608)。

作者简介：姚兴虎(1996—)，男，山东济宁人，硕士研究生在读，CCF 会员，主要研究方向：深度强化学习、多智能体强化学习；谭晓阳(1971—)，男，江苏南京人，教授，博士，CCF 会员，主要研究方向：机器学习、强化学习。

Coma and Qmix algorithms. More importantly, in 3s5z and 3s6z scenarios with a large number and different types of agents, better results can be achieved with 30% of the required number of samples of algorithms such as Qmix, Coma, etc.

Keywords: deep learning; deep reinforcement learning; multi-agent reinforcement learning; multi-agent system; global credit assignment

0 引言

近年来，深度强化学习在游戏人工智能[1][2]，机器人自动控制[3]等领域取得了很大的进步。 然而，许多现实世界的真实场景需要多个智能体在同一个环境中与环境进行交互，这类问题场景可以建模为多智能体系统[4][5]。 常见的多智能体系统包括：多智能体协同规划[6]，信号灯的控制[7]以及多玩家电子游戏[8]等。 然而，多智能体系统的复杂性使得多智能体系统面临着诸多单智能体系统中没有的问题，这些问题使得简单地将单智能体强化学习算法移植到多智能体场景中不会取得令人满意的效果。具体来说，多智能体系统中面临的主要问题包括：每个智能体只能观测到环境的一部分所导致的对环境的部分可观测问题[9]；环境本身所具有的更强的非马尔可夫性[10]；多个智能体与环境进行不断的交互所导致的环境不稳定问题[11]；多个智能体的联合动作空间随着智能体数量的增加所导致的指数爆炸[12-15]；以及如何将环境反馈的针对环境中所有智能体联合动作的全局奖励分配给每个独立的智能体（称之为全局信用分配问题）[12-15]。 这些问题的存在不仅使得无法将所有的智能体建模为一个单智能体然后利用单智能体算法进行训练，而且也不适合将其他智能体看成环境的一部分从而为每个智能体单独进行建模。

近年来，由于概念上简单并且执行效率高，“中心训练-分散执行”的方式已经成为求解多智能体强化学习问题的一个标准范式[12-15]。 所谓“中心训练”，指的是在训练的过程中通过一个中心化的值函数来与环境直接进行交互；所谓“分散执行”，指的是每个智能体都有自己单独的值函数网络或者策略网络，因此在执行阶段每个智能体可以根据其自身的观测独立地执行动作。 在这一范式中，中心化的值函数直接接收环境给出的奖励信号，之后通过适当的全局信用分配机制。因此，中心化的值函数建立了每个智能体与环境进行交互的桥梁并在整个框架中处于核心地位。

如何设计中心化值函数与每个智能体的值函数之间的约束关系是设计整个信用分配机制的核心。一个合适的约束关系不仅能够有利于对全局信用进行一个良好的分配，还应使得整个算法复杂度不易过高.若采用简单的信用分配机制（比如“值分解网络(value decompose network, Vdn)[13]”中的加性方式），则会限制中心化值函数的表达能力并进一步影响到奖励分配过程；若设计复杂的奖励分配机制（比如“反直觉的多智能体策略梯度法(counterfactual multi-agent policy gradient, Coma[12])”和“Q 值变换网络(Qtran[15]”）则会增加优化求解的复杂度。

此外，Vdn[13]，Qmix[14]以及 Qtran[15]算法均假设全局最优的联合动作等价于每个智能体按照自己的值函数求得的局部最优动作的联合。 然而，复杂场景下的全局最优动作可能需要某些智能体做出一些牺牲其个人利益的行为。因而，基于这一假设的算法最终会收敛到问题的一个局部最优解。

针对多智能体强化学习问题中全局信用分配机制存在的上述问题，在“中心训练-分散执行”的框架下，本文提出了一种新的全局信用分配方法，称之为：奖励高速路网络（reward highway network, RHWNet）。RHWNet 将中心化值函数与每个智能体的值函数之间的耦合分为两部分：一方面通过混合网络来实现全局的奖励分配，这一方式能够对不同的智能体进行特异性的奖励分配；另一方面利用奖励高速路连接将全局奖励信号桥接到每个智能体值函数的训练过程中，从而实现全局信用的二次分配，这将使得单个智能体在最大化自身奖励值的同时兼顾其行为对全局奖励的影响。在算法复杂度方面，本文所提出的全局信用的二次分配过程几乎不需要额外的优化代价。在星际争霸微操作平台上的实验结果表明：本文所提出的方法在多个复杂的场景下能够获得很好的性能提升，并且具有更高的样本利用效率。。

本文的后续结构如下：首先本文在第 1 节介绍相关工作，然后在第 2 节介绍多智能体强化学习的一些背景知识，第 3 节提出奖励高速路网络这一全局信用分配结构。第 4 节给出所提算法的主要实验结果和消融研究，第 5 节对全文进行总结并给出一些未来的工作。

1 相关工作

近年来，随着深度强化学习方法的流行，多智能体强化学习算法的研究已从简单的环境过渡到复杂的场景。

由于概念上简单并且执行的高效，“中心训练-分散执行”的方式已经成为求解多智能体强化学习问题的一个标准范式。这类方法通常假设每个智能体的局部最优动作的拼接等价于联合的最优动作.其中代表性的方法有：Coma[12]，Vdn[13]，Qmix[14]和 Qtran[15]。 Coma 是一种在策略的“演员-评论家”算法，其通过一个精心设计的反直觉的基准来实现全局信用的分配，但是这一基准需要额外的计算代价。 Vdn，Qmix 和 Qtran 则是利用值函数迭代的方式，首先学习中心化的值函数，然后利用中心化值函数与非中心化值函数之间的约束关系完成全局信用的分配。值函数之间不同程度的约束关系使得 Vdn，Qmix 和 Qtran 三种方法的信用分配机制的复杂程度和优化求解难度有所不同。Smix(丨)[16]旨在学习一个更为灵活和更强泛化能力的中心化值函数结构，其未改变原有算法的奖励分配机制。

本文所提出的基于奖励高速路网络的信用分配机制同样属于“中心训练-分散执行”的框架，但是其重点在于如何在

不引入额外的信息以及不增加优化代价的前提下进行更为有效的信用分配。

此外，为智能体之间建立通信信道或者建立智能体之间的协调配合机制可以为单个智能体的决策提供更多的环境信息或者环境中其他智能体的信息。建立通信信道的方法主要包括[17][18]等；智能体之间的协调配合机制可以通过在智能体之间引入注意力机制[19][20]或者利用图神经网络[21][22]来实现。每个智能体利用更多的信息进行决策所产生的行为将会间接影响到整个系统的奖励分配。而本文所提出的方法在不考虑更多信息的条件下改善已有的信用分配机制。因此，这类方法与本文所提出的方法是互补的。

2 背景知识

在本节中，本文将介绍多智能体强化学习的相关背景知识。其中，2.1 节给出了多智能体强化学习的相关符号与问题建模；2.2 节介绍了本文所提算法的值函数的基本形式-深度循环 Q 函数网络(Deep Recurrent Q Network, DRQN)[23]；2.3 节介绍了三种流行的基于值函数迭代的多智能体强化学习算法，Vdn[13]，Qmix[14]和 Qtran[15]。

2.1 问题建立

本文考虑完全合作场景下的多智能体强化学习问题，其可被描述为非中心化部分可观测马尔可夫决策过程(Decentralized partial observable Markov Decision Process，Dec-POMDP)[24] 的一个变种。具体来说，本文可以用八元组 $G=(S,A,P,r,Z,O,N,\gamma)$ 来描述这一问题，其中 $s\in S$ 表示环境的真实状态，A 是每个智能体的所能采取的动作的集合。其中智能体的数目的总数是 N，γ 是奖励折扣因子。在每个时刻，每个智能体 $i\in\{1,2,\ldots,N\}$ 分别选取动作 $a^i\in A$ 从而拼成联合动作向量 $\boldsymbol{a}=\{a^1,a^2,\ldots,a^N\}\in A^N$。本文考虑一个部分可观测的场景，其中每个智能体 i 只能通过观测函数 $Z(s,i):S\times N\to O$ 得到部分信息 $o\in O$。每个智能体 i 历史的观测和动作序列为 $\tau^i\in T\equiv(O\times A)^*$。每个智能体将依据历史的观测和动作序列 τ 来进行决策。策略函数可分为静态策略函数和随机策略函数，其中随机策略函数可以定义为：$\pi[a|\tau]:T\times A\to[0,1]$。

在“中心训练-分散执行”的框架下，训练阶段利用环境的全局状态 s 和各个智能体的历史观测信息 $\boldsymbol{\tau}=\{\tau^1,\tau^2,\ldots,\tau^N\}$ 学习一个中心化的动作值函数 $Q([s,\boldsymbol{\tau}],a)$（简记为 $Q(\boldsymbol{\tau},a)$）。在执行阶段，每个智能体的策略函数 p^i 仅仅依赖于其自身的观测和动作历史序列 τ^i。所有智能体的共同目标是最大化所能从环境中得到的全局折扣奖励和：$\mathbf{E}_{a_0,\boldsymbol{\pi},s_0}\left[\sum_{t=0}^{\infty}\gamma^t r(s,\boldsymbol{a})\right]$。在下文中，为了简化记号，本文用黑体字符表示所有智能体的联合行为，并且在不引起歧义的情况下，省略每个智能体的序号 i。

2.2 深度循环 Q 网络

正在复杂的现实世界中的问题场景下，通常不能得到完整的状态信息并且观测的数据往往是具有噪声的，这种部分可观测的问题在多智能体场景下更为严重。此外，多智能体环境所天然具有的非马尔可夫性使得每个智能体需要考虑更多的历史信息来进行当前时刻的决策。文献[23]的结果表明，传统的深度 Q 网络在处理部分可观测的 MDP 问题中会出现性能下降而深度循环 Q 网络更为适合处理部分可观测以及非马尔可夫的环境。

深度循环 Q 网络通过引入 GRU[25]或者 LSTM[26]等循环神经网络结构来实现对历史信息的融合从而计算状态动作值。一方面，多智能体环境面临更严重的部分可观测性，采用这一循环神经网络结构能够更好地对历史信息进行融合，从而缓解对环境的部分可观测问题。另一方面，序列决策问题中当前的策略可能受到之前多步的状态和动作的影响，因此这一循环神经网络结构还能有助于处理序列决策问题场景下的非马尔可夫问题。与深度 Q 网络[1]相同，DRQN[23]也利用一个数据缓存区（replay buffer）来存储经验数据 $(\boldsymbol{\tau},\boldsymbol{a},r,\boldsymbol{\tau}',\gamma)$，其中 $\boldsymbol{\tau}'$ 是在联合的局部观测 $\boldsymbol{\tau}$ 下智能体采取联合动作 $\boldsymbol{a}$ 后获得全局奖励值 r 所得到的下一个联合观测值。DRQN 通过最小化如下的均方时间差分损失来进行学习：

$$L(\theta)=\sum_{i=1}^{b}\left[y_i^{DRQN}-Q(\boldsymbol{\tau},\boldsymbol{a};\theta)\right]^2, \tag{1}$$

其中 θ 是值函数网络的参数。$y_i^{DRQN}=r+\gamma\max_{a'}Q(\boldsymbol{\tau}',\boldsymbol{a}';\theta^-)$，$\theta^-$ 是目标网络(target network)的参数，其更新方式为每隔固定的迭代次数将主网络的参数 θ 直接复制。

2.3 Vdn, Qmix 和 Qtran

多智能体系统中联合的动作空间随着智能体数量的增加指数爆炸，因此直接优化联合的动作值函数代价巨大。为了降低算法的复杂度，众多算法假设智能体的联合最优动作等价于每个智能体依据其自身的值函数进行贪心地动作选择所得到的局部最优值的拼接，即：

$$\underset{\boldsymbol{a}}{argmax}\,Q_{tot}(\boldsymbol{\tau},\boldsymbol{a})=\begin{pmatrix}\underset{a^1}{argmax}\,Q^1(\tau^1,a^1)\\ \underset{a^2}{argmax}\,Q^2(\tau^2,a^2)\\ \vdots\\ \underset{a^N}{argmax}\,Q^N(\tau^N,a^N)\end{pmatrix}. \tag{2}$$

值函数分解网络（value decomposition network, Vdn）[13]限制中心化的值函数 $Q_{tot}(\boldsymbol{\tau},\mathbf{a})$ 为每个智能体的值函数的和，即：

$$Q_{tot}(\boldsymbol{\tau},\boldsymbol{a})=\sum_{i=1}^{N}Q^i(\tau^i,a^i;\theta^i). \tag{3}$$

Vdn 算法的损失函数和(1)相同，这一方法的优势在于其结构简单，但是这一简单的结构限制了中心化值函数的表达能力和全局信用分配的有效性。 Qmix[14]将这一线性分解拓展到了单调非线性分解。 具体来说，Qmix 假设中心化的值函数是每个智能体值函数的非负线性组合，即：

$$\frac{\partial Q_{tot}}{\partial Q^i}\geq 0,\ \forall i\in\{1,2,\cdots,N\}. \tag{4}$$

Qmix 算法通过建立每个智能体的值函数网络，一个混合网络和一系列的超网络来实现上述约束。 并且 Qmix 算法在超网络中输入全局的状态来辅助中心化值函数的训练。

Vdn 和 Qmix 算法的约束都是假设(2)的一个充分条件，Qtran[15]算法则进一步对约束进行松弛从而直接优化假设(2)的一个充要条件. 尽管 Qtran 工作在一个更大的假设空间，但是这一方法需要求解联合动作空间中的优化问题，这将带来庞大的计算代价。因此 Qtran 并不适用于复杂的多智能体场景。

3 本文方法

3.1 奖励高速路连接

残差网络[27]通过在深度神经网络中增加跳跃连接来缓解深度神经网络在信息传递的过程中所造成的信息丢失与损耗。高速路网络则是利用门控机制，将当前的信息选择性的进行传递。 本文利用残差学习的观点，将每个智能体应分到的奖励分为两部分：贪心奖励和合作奖励。 所谓贪心奖励是指按照假设(2)进行信用分配所分给每个智能体的奖励，记作 r_g，仅仅采用这种分配方式将使得每个智能体依据其自身的值函数进行贪心地策略选择；所谓合作奖励指的是每个智能体还应考虑的全局奖励部分，记作 r_c。 r_c 可通过对全局奖励 R 进行部分桥接得到，即 $r_c=\lambda\times R,\lambda\in[0,1]$。 本文称这种全局奖励直达的连接方式为奖励高速路连接。经过这两种形式的奖励分配后，训练过程中单个智能体 i 的实际收到的奖励信号为 $r^i=r_g^i+r_c^i$。 记环境所给的外部奖励为 R，则在一个有 N 个智能体的多智能体环境中，r_g,r_c 与 R 之间的关系为：

$$R=F(r_g^1,r_g^2,\cdots,r_g^N)+\sum_{i=1}^{N}r_g^i, \tag{5}$$

其中 F 为满足假设(2)所进行的全局信用分配函数，其可以是简单的所有贪心奖励 r_g 的和（对应于 Vdn），或者是所有 r_g 的非负组合（对应于 Qmix）。 上述二路的奖励分配方式及其与残差网络的结构对比可以用图 1 来描述。

从图 1 可以看出，残差连接[27]和奖励高速路连接均是在深度网络中添加一些跳过某些中间层的跳跃连接。这种跳跃连接的方式几乎不会带来额外的优化代价，但更多的信息将通过跳跃连接进行传递。两种结构不同之处在于：残差连接的信息流向是从前往后的，这样上一阶段的信息能够对后续阶段产生影响；而奖励分配的方式是从后往前的，这将使得两路奖励信号都被用来训练每个智能体的值函数网络，从而使得单独的智能体在考虑最优化其自身的利益的同时最大化全局奖励值。

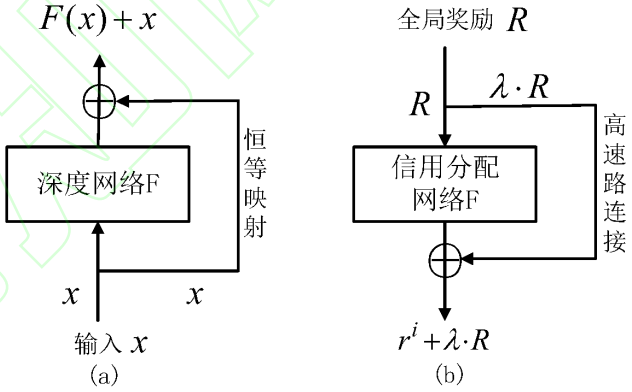

图1 残差连接和奖励高速路连接对比（(a)为残差块示意图，(b)为本文所提出的奖励高速路示意图）

Fig. 1 Comparison of residual network and reward highway network

3.2 本文所提算法

本文采用 Qmix 的网络结构作为本文算法的基本网络结构。 Qmix 采用混合网络和一系列的超网络来构造信用分配网络 F。每个超网络接受全局状态作为输入，输出的非负值作为混合网络的权重。本文称在这一信用分配网络 F 上加入奖励高速路连接所得到的网络为奖励高速路网络(Reward Highway Network, RHWNet)。RHWNet 的示意图如图 2 所示，与 Qmix 相同，在每个智能体单独的值函数网络中加入 GRU 来实现对历史信息的利用，并且所有智能体的值函数网络是参数共享的。通过图 2 可以看出，奖励高速路连接并不会引入额外的神经网络参数，因此 RHWNet 并没有额外的优化代价。

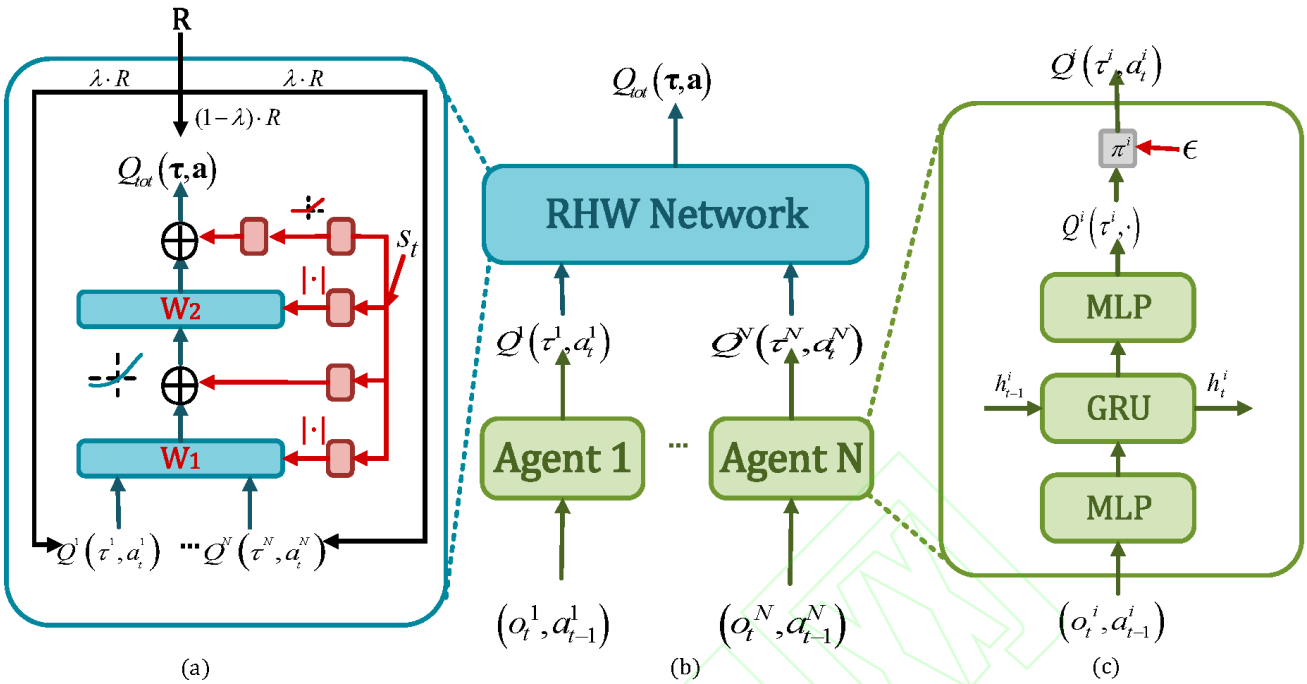

图2 本文所提算法的网络结构图（(a)为奖励高速路网络 RHWNet，(b)为算法的总体结构，(c)为单个智能体的值函数结构）

Fig. 2 The network structure of the proposed algorithm ((a): reward highway network, (b): overall structure of the proposed algorithm, (c): value function structure of a single agent)

在实现过程中，本文算法通过最小化如下的损失函数进行端到端的训练：

$$L(\theta)=\sum_{b=1}^{N_b}\left[(1-\omega)\left(y_b^{tot}-Q_{tot}^{p}(\boldsymbol{\tau},\boldsymbol{a};\theta,\psi)\right)^2+\omega\sum_{i=1}^{N}\left(y_b^{tot}-Q^i(\tau^i,a^i;\theta)\right)\right] \tag{6}$$

其中 N_b 为采样批量(batch)的大小，ω 为将全局奖励通过奖励高速路网络输送到每个智能体上的权重。θ 为所有智能体非中心化的值函数网络的参数。ψ 为奖励高速路网络的参数。$y^{tot}=r+\gamma\max_{\boldsymbol{a}'}\left(\boldsymbol{\tau}',\boldsymbol{a}',s;\theta^-,\psi^-\right)$，其中 γ 是奖励折扣因子，θ^-,ψ^- 是与标准的深度 Q 学习算法中相同的目标网络(target network)的参数。 所有的神经网络都是通过端到端的方式进行训练的。

4 实验与结果

在这一节，本文首先给出本文所提算法的实验环境和算法的实现细节，然后给出实验结果和消融分析。

4.1 实验环境

本文在星际争霸多智能体挑战(StarCraft Multi-Agent Challenge, SMAC)[28] 环境上对本文所提的 RHWNet 进行评估。SMAC 是基于星际争霸 II 游戏的一个实验环境，与完整的星际争霸 II 游戏相比，SMAC 侧重研究每个智能体的微操作。微操作指的是 SMAC 重点关注如何控制每个士兵去战胜敌方，而不考虑如何发展经济以及进行资源的调度等高层次的宏观操作。

表1 实验中所考虑的不同场景

Tab. 1 The scenarios considered in our experiments

场景	盟军	敌军	类型
2s_vs_1sc	2 Stalkers	1 Spine Crawler	asymmetric
3s5z	3 Stalkers & 5 Zealots	3 Stalkers & 5 Zealots	heterogeneous & symmetric
1c3s5z	1 Colossi, 3 3 Stalkers & 5 Zealots	1 Colossi, 3 1 Stalkers & 5 Zealots	heterogeneous & symmetric
3s6z	3 Stalkers & 6 Zealots	3 Stalkers & 6 Zealots	heterogeneous & symmetric

SMAC 提供了多种复杂的微操作场景来探究智能体之间的合作行为。在每个场景中，开始时刻两组敌对的士兵被分配到战场中的随机位置。战场中的每个士兵只能在其视野范围内搜集到关于战场环境的局部信息，这将带来严重的对环境的部分可观测性。环境仅根据智能体所采取的联合动作来给出一个全局的奖励信号。本文采用强化学习算法来控制战场中的一组士兵（同盟单元），来与内置的基于启发式规则的游戏 AI 控制的另一组士兵进行对抗。 在实验中，内置 AI 的难度被设置为“非常困难”来验证本文算法的有效性。

本文所提算法旨在优化合作场景下的全局奖励分配问题。 因此重点考虑非对称(asymmetic)场景（敌我双方士兵构成不同）以及非齐次且对称(heterogeneous and symmetic)场景

下（敌我双方士兵人员组成相同，但均由不同种类的士兵构成）的对抗。 表 1 列出了实验所考虑的四种实验场景。

4.2 实现细节

每个智能体的值函数网络由以下结构构成：首先从环境中得到的观测传入一层维度为 64 维的全连接层，经过 ReLU[29]激活函数后，输入到维度为 64 的 GRU 模块进行当前信息与历史信息的整合，GRU 模块的输出传入到一层维度为 64 的全连接层，之后再经过 ReLU 激活函数得到当前智能体的动作值向量 $Q^i(t^i,\cdot)$。然后根据这一动作值函数进行 ε-贪心的策略选择，随着训练的进行，ε 的取值从 1.0 线性衰减到 0.05。 为了降低网络的参数数量，所有智能体共享同一个动作值函数网络。

之后每个智能体的 Q 值传入奖励高速路网络，奖励高速路网络中的混合网络部分采用与 Qmix 算法相同的结构。 全局奖励值经过高速路传输的多少由公式(6)中的 λ 参数控制，在本文的所有实验场景中本文均设置 $\lambda=0.2$。

本文采用 RMSprop 方法来最小化损失函数(6)，其参数设置为：$lr=0.0005,\alpha=0.99$，奖励折扣因子 $\gamma=0.99$。每经过 200 局游戏对目标网络的参数进行一次更新。

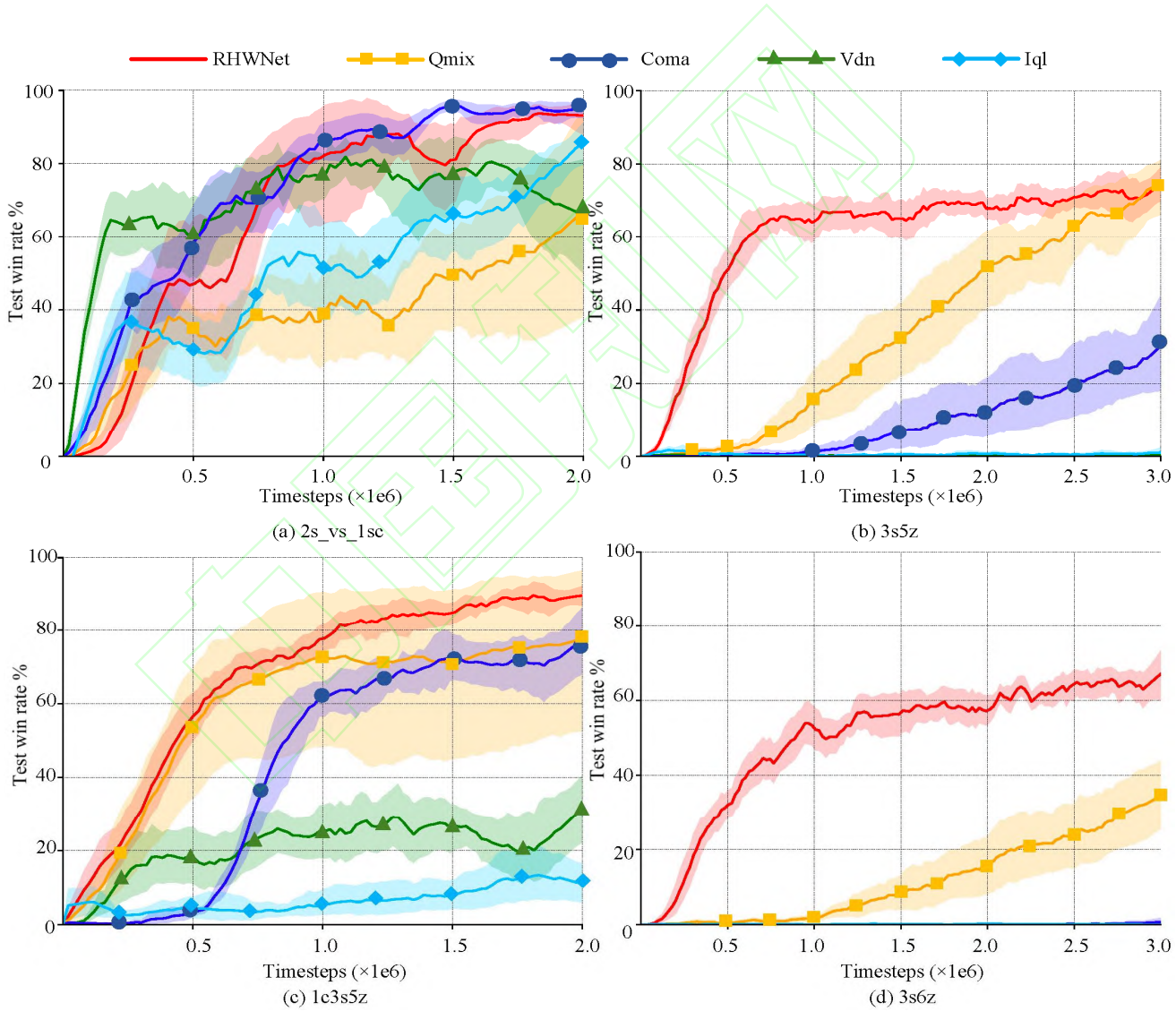

图3 本文所提算法与其他算法(Iql, Vdn, Qmix, Coma)，在四个场景下的测试胜率对比.

Fig. 3 Test win rates for our method and comparison methods (Iql, Vdn, Qmix,Coma).

4.3 主要实验结果

本文将所提算法与 SMAC 平台上最先进的算法 Coma 和 Qmix 进行对比，并与不进行全局信用分配的独立 Q 学习算法(independent Q learning, Iql)和只进行简单全局信用分配的 VDN 算法进行对比，主要实验结果如图 3 所示。本文将每个

算法在所有不同的场景中均独立训练十次，得到的线条和阴影部分分别表示平均测试胜率及对应胜率方差的95%的置信区间。阴影部分的面积大小可以作为衡量算法稳定性和鲁棒性的评价指标，阴影面积越小意味着算法的性能方差越小从而算法的稳定性和鲁棒性越好。子图(a)中的图例适用于所有子图。

可以看到，在所有的非齐次对称场景下(3s5z, 1c3s5z, 3s6z)，本文提出的算法能够取得最优的性能，并且在较为简单的场景(2s_vs_1sc)下也能获得接近最优的性能。此外RHWNet的性能提升不仅体现在最终的胜率上，还体现在学习的效率上。

具体来说，在智能体数量较少的 2s_vs_1sc 场景下，本文可以看出采用较为复杂奖励分配机制的 Qmix 性能要明显差于结构更简单的 Vdn 和 IQL 算法。这意味着 QMIX 这一较为复杂的全局信用分配机制在某些较为简单的场景下也有可能失效。而通过奖励高速路连接之后，RHWNet 算法在这一场景下的得到了很大的性能提升。

在 1c3s5z 场景下，每个团队中都有三种不同类型的智能体。 如图 3(c)所示，在这一场景下，Qmix 和 Vdn 算法性能都出现了较大的波动（对应的阴影部分面积增大）。 而 RHWNet 在取得性能提升的同时还具有更小的性能上的方差，这意味着 RHWNet 在复杂的问题场景下依然具有很好的鲁棒性。

在 3s5z 场景下，本文可以看到采用更为复杂奖励分配方式的 Qmix 算法性能要大大优于采用简单信用分配方式的 Vdn 算法以及不进行信用分配的 Iql 算法。尤其需要指出的是，Vdn 算法可看作 Qmix 算法的简化版本，这意味着在这一复杂的场景下，Qmix 所采用的更复杂的结构更有效。 然而这些基准算法都存在样本利用率低，学习速度慢的问题。而 RHWNet 则能大大加快算法的学习速度和样本利用的效率。同样的结果可以在更为复杂的 3s6z 场景下得到。在 3s6z 场景中，Coma，Vdn 和 Iql 的训练基本无效，Qmix 也不能得到令人满意的结果。而 RHWNet 在仅需要 Qmix 算法所需样本数量的三分之一的情形下，最终胜率能达到 Qmix 算法的 1.5 倍。这表明在 3s6z 这一智能体数量和种类较多的复杂场景下，已有算法的奖励分配机制不能有效地进行全局奖励分配，而奖励高速路连接为这种复杂场景引入了一个更好的奖励分配机制，从而取得了最终性能和样本效率的提升。

4.4 消融测试

在这一部分本文重点探究通过奖励高速路网络传递的全局奖励的比例对最终的实验性能所产生的影响。公式(6)中参数 λ 的作用其实起到了平衡原有的端到端的奖励分配方式和直接利用全局奖励的作用。当公式(6)中的 λ 取值较小时，每个智能体所获得的奖励信号更多地来源于直接的全局奖励；当 λ 取值较大时，每个智能体的奖励信号则更多地来源于混合网络的信用分配结果。

图 4 显示了在 3s5z 场景下，λ 的不同取值所获得的实验结果。其中实线和阴影表示独立进行 10 次实验的均值和 95% 的置信区间。从这一实验结果可以看出，当 λ =0.2,0.4,0.6 时，RHWNet 均能得到明显的性能提升。但是当 λ 的值进一步增大时，反而会出现性能下降。因此，通过信息高速路网络进行传输的全局奖励值的比例实际上起到了对原有信用分配机制与仅考虑全局奖励的平衡。实验结果表明，λ =0.2 是一个比较鲁棒的值。因此本文的所有实验场景都采用 λ =0.2 作为奖励高速路链接网络的权重。

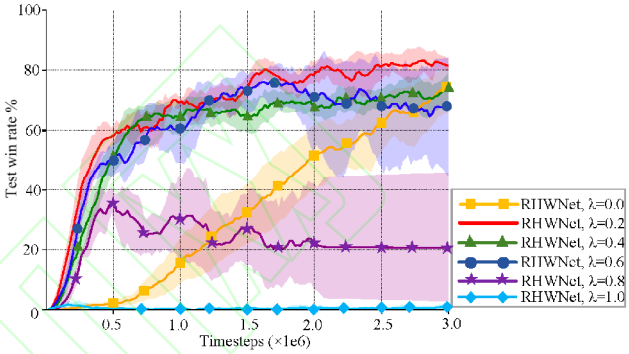

图 4 在 3s5z 场景中，所提出的算法对超参数 λ 的敏感性。

Fig. 4 Sensitivity of the proposed method to selected hyperparameter λ in 3s5z scenario.

5 总结和展望

在“中心训练-分散执行”的多智能体强化学习框架下，全局信用的分配可以通过对中心化值函数和非中心化值函数之间施加约束来实现。然而，不同的约束关系不仅决定了算法的复杂程度，还直接决定了奖励分配机制的有效性。 本文提出了一种基于奖励高速路网络的全局信用分配算法 RHWNet，通过在奖励分配机制上引入奖励高速路连接，能够达到：(1)每个智能体的决策行为能够考虑其自身所分得的局部奖励和整个团队的全局奖励；(2)奖励高速路连接结构简单，几乎不会引入额外的优化代价；(3)在多个复杂的场景下，RHWNet 相比原有的先进算法能够取得很好的性能提升。

本文的后续工作将会研究限制条件下的全局奖励分配问题（比如智能体之间存在资源竞争的关系），以及为智能体之间建立通信机制来进行协调配合。

参考文献

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

[2] 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. (LIU Q, ZHAI J W, ZHANG Z C, et al. A survey on deep reinforcement learning [J]. Chinese journal of computers, 2018, 41(1): 1–27.)

[3] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms [EB/OL]. [2020-09-03]. https://arxiv.org/pdf/1707.06347. pdf.

[4] 殷昌盛, 杨若鹏, 朱巍, 等. 多智能体分层强化学习综述[J]. 智能系统学报, 2020, 15(4): 1－10. (YIN CHANGSHENG, YANG RUOPENG, ZHU WEI, et al. A survey on multi-agent hierarchical reinforcement learning [J]. CAAI transactions on intelligent systems, 2020, 15(4): 1–10.)

[5] 孙长银, 穆朝絮. 多智能体深度强化学习的若干关键科学问题. 自动化学报, 2020, 46(7):1301-1312. (SUN CHANG-YIN, MU CHAO-XU.Important scientific problems of multi-agent deep reinforcement learning. Acta Automatica Sinica, 2020, 46(7): 1301−1312)

[6] 王冲, 景宁, 李军, 等. 一种基于多 Agent 强化学习的多星协同任务规划算法 [J]. 国防科技大学学报, 2011,33(1): 53–58. (WANG CHONG, JING NING, LI JUN, et al. An algorithm of cooperative multiple satellites mission planning based on multi-agent reinforcement learning [J]. Journal of national university of defense technology, 2011, 33(1): 53–58.)

[7] VAN DER POL E, OLIEHOEK F A. Coordinated deep reinforcement learners for traffic light control [C]// Proceedings of the 2016 Advances in Neural Information Processing Systems' Workshop on Learning, Inference and Control of Multi-Agent Systems. Cambridge: MIT Press, 2016: 1-8

[8] JADERBERG M, CZARNECKI W M, DUNNING I, et al. Human-level performance in 3D multiplayer games with population-based reinforcement learning [J]. Science, 2019, 364(6443): 859-865

[9] NAIR R, TAMBE M, YOKOO M, et al. Taming decentralized POMDPs: Towards efficient policy computation for multiagent settings [C]// Proceedings of the 2003 International Joint Conference on Artificial Intelligence. Piscataway: IEEE, 2003: 705-711

[10] LAURENT G J, MATIGNON L, FORT-PIAT L. The world of independent learners is not Markovian [J]. International Journal of Knowledge-based and Intelligent Engineering Systems, 2011, 15(1): 55-64

[11] LOWE R, WU Y I, TAMAR A, et al. Multi-agent actor-critic for mixed cooperative-competitive environments [C]// Proceedings of the 2017 Advances in neural information processing systems. Cambridge: MIT Press, 2017: 6379-6390

[12] FOERSTER J N, FARQUHAR G, AFOURAS T, et al. Counterfactual multi-agent policy gradients [C]// Proceedings of the thirty-second AAAI conference on artificial intelligence. Menlo Park, AAAI, 2018: 2974-2982

[13] SUNEHAG P, LEVER G, GRUSLYS A, et al. Value-decomposition networks for cooperative multi-agent learning based on team reward [C]// Proceedings of the 17th international conference on autonomous agents and multiagent systems. Cham: Springer, 2018: 2085-2087

[14] RASHID T, SAMVELYAN M, SCHROEDER C, et al. QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning [C]// Proceedings of the 2018 International Conference on Machine Learning. New York: JMLR. org, 2018: 4295-4304

[15] 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 2019 International Conference on Machine Learning. New York: JMLR. org, 2019: 5887-5896

[16] YAO X, WEN C, WANG Y, et al. SMIX (丨): Enhancing Centralized Value Functions for Cooperative Multi-Agent Reinforcement Learning [EB/OL]. [2020-09-03]. https://arxiv.org/pdf/1911.04094.pdf

[17] SUKHBAATAR S, FERGUS R. Learning multiagent communication with backpropagation [C]// Proceedings of the 2016 Advances in neural information processing systems. Cambridge: MIT Press, 2016: 2244-2252.

[18] FOERSTER J, ASSAEL I A, DE FREITAS N, et al. Learning to communicate with deep multi-agent reinforcement learning [C]// Proceedings of the 2016 Advances in neural information processing systems. Cambridge: MIT Press, 2016: 2137-2145

[19] IQBAL S, SHA F. Actor-Attention-Critic for Multi-Agent Reinforcement Learning [C]// Proceedings of the 2019 International Conference on Machine Learning. New York: JMLR. org, 2019: 2961-2970

[20] JIANG J, LU Z. Learning attentional communication for multi-agent cooperation [C]// Proceedings of the 2018 Advances in neural information processing systems. Cambridge: MIT Press, 2018: 7254-7264

[21] JIANG J, DUN C, HUANG T, et al. Graph Convolutional Reinforcement Learning [EB/OL]. [2020-09-03]. https://arxiv.org/pdf/1810.0920 2.pdf

[22] LIU Y, WANG W, HU Y, et al. Multi-Agent Game Abstraction via Graph Attention Neural Network [EB/OL]. [2020-09-03]. https://arxiv.org/pdf/1911.10715.pdf

[23] HAUSKNECHT M, STONE P. Deep recurrent q-learning for partially observable mdps [C]// Proceedings of the 2015 AAAI Fall Symposium Series. Menlo Park, AAAI, 2015: 29-37

[24] OLIEHOEK F A, AMATO C. A concise introduction to decentralized POMDPs [M]. Cham: Springer, 2016: 11-30

[25] CHO K, VAN MERRIËNBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [EB/OL]. [2020-09-03]. https://arxiv.org/pdf/1406.1078.pdf

[26] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural computation, 1997, 9(8): 1735-1780

[27] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C]. Proceedings of the 2016 IEEE conference on computer vision and pattern recognition. Piscataway: IEEE, 2016: 770-778

[28] SAMVELYAN M, RASHID T, SCHROEDER DE WITT C, et al. The starcraft multi-agent challenge [C]. Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems. Cham: Springer, 2019: 2186-2188.

[29] HAHNLOSER R H R, SARPESHKAR R, MAHOWALD M A, et al. Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit [J]. Nature, 2000, 405(6789): 947-951

This work is partially supported by, National science foundation of China (61976115,61672280, 61732006), Equipment pre-research fund (6140312020413), Artificial intelligence+ Project of Nanjing University of Aeronautics and Astronautics (56XZA18009)，Pre research on military shared information system equipment (315025305), Graduate Innovation Foundation of Nanjing University of Aeronautics and Astronautics (Kfjj20191608).

Yao Xinghu, born in 1996. Master candidate. His main research interests include deep reinforcement learning, multi-agent reinforcement learning, machine learning.

Tan Xiaoyang, born in 1971. PhD, professor, PhD supervisor. His main research interests include computer vision, deep reinforcement learning, machine learning.