(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 118587558 A
(43)申请公布日 2024.09.03

(21)申请号 202410693931.8

(22)申请日 2024.05.31

(71)申请人 北京航空航天大学
地址 100190 北京市海淀区学院路37号

(72)发明人 石荣晔 廖书昊 吴文峻

(74)专利代理机构 北京慕达星云知识产权代理事务所(特殊普通合伙) 11465
专利代理师 郭米琪

(51)Int.Cl.
G06V 10/82(2022.01)
G06N 3/092(2023.01)
G06N 7/01(2023.01)
G06N 3/08(2023.01)
G06N 3/063(2023.01)

权利要求书3页 说明书8页 附图3页

(54)发明名称

一种基于多智能体强化学习的振荡神经网络突触参数优化方法

(57)摘要

本发明公开了一种基于多智能体强化学习的振荡神经网络突触参数优化方法、设备及介质,方法包括:确定数字图像对应的存储模式,并作为学习任务;基于课程学习对学习任务进行拆分;基于拆分的数字图像识别任务集合,建立多智能体强化学习仿真环境;对多智能体强化学习进行训练,获得最优奖励的振荡神经网络突触权重;识别和存储所有预定的数字图像。在数字图像识别领域,这种方法使得单个网络能够识别更多的数字图像,并且增强了对复杂变化环境的动态适应性,显著提高了识别精度和处理效率。



CN 118587558 A

1.基于多智能体强化学习的振荡神经网络突触参数优化方法,其特征在于,包括:

确定数字图像对应的存储模式,并作为学习任务;

基于课程学习对学习任务进行拆分;

基于拆分的数字图像识别任务集合,建立多智能体强化学习仿真环境;

对多智能体强化学习进行训练,获得最优奖励的振荡神经网络突触权重;

识别和存储所有预定的数字图像。

2.根据权利要求1所述的基于多智能体强化学习的振荡神经网络突触参数优化方法,其特征在于,确定数字图像对应的存储模式,具体包括:将振荡神经网络需要存储的数字图像转换为二值模式矩阵,-1代表黑色,1代表白色,二值模式矩阵的集合作为最终学习任务。

3.根据权利要求1所述的基于多智能体强化学习的振荡神经网络突触参数优化方法,其特征在于,基于拆分的数字图像识别任务集合,建立多智能体强化学习仿真环境具体包括:

基于振荡神经网络定义状态、观测、动作、状态转移函数以及奖励,创建多智能体强化学习仿真环境,并将拆分的数字图像识别任务集合作为环境初始模式集合;

状态S_t在每个时间步t捕捉了ONN中所有神经元的数字图像中间相位信息;

每个智能体在时间步t的观测$o_{i,t}$提供了对ONN突触权重矩阵的部分观测;

每个智能体的动作$a_{i,t}$反映了其对第i个智能体相关联的突触权重的修改的局部调整;

第i个智能体从时间步长t到时间步长t+1的状态转移函数描述为:

$$S_{t+1} = S_t + \sum_{i=1}^{n} \left(a_{i,t} + a_{i,t}^{T}\right)$$

式中,S_t和S_{t+1}分别表示时间步长t和t+1的突触权重矩阵,$a_{i,t}$表示第i个智能体在时间步长t的行动,是一个n维矢量,$a_{i,t}$被添加到突触权重矩阵S_t的第i列,$a_{i,t}^{T}$被添加到突触权重矩阵S_t的第i行,对应于与第i个神经元连接的权重;

智能体i的奖励r_i计算公式为:

$r_i = r/n$

$$\varphi^{k^*} = \begin{bmatrix} \varphi_1^{k^*} \\ \varphi_2^{k^*} \\ \vdots \\ \varphi_n^{k^*} \end{bmatrix}, \delta\varphi^{k^*} = \varphi^{k^*} - \varphi_1^{k^*} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, k = 1,2,\ldots,P$$

$$r = P \times \mu + \sum_{k=1}^{P} - L_1\left(\delta\varphi^{k^*}, \delta\varphi^{k}\right)$$

$L_1(x,y) = |(x+2m\pi) - y|, x+2m\pi \in [-\pi,\pi]$

上式中,P指当前任务存储数字图像的数量,φ^{k^*}代表第k个输入模式的输出数字图像中间相位,$\varphi_i^{k^*}$指的是φ^{k^*}的第i个元素,$\delta\varphi^{k}$表示第k个输入的数字图像的相位偏差,m是使得

$x+2m\pi\in[-\pi,\pi]$的整数,x、y是两个不同的变量,n是智能体个数,μ表示用来区分不同任务的超参数。

4.根据权利要求3所述的基于多智能体强化学习的振荡神经网络突触参数优化方法,其特征在于,对多智能体强化学习进行训练具体包括:

初始化动作价值网络$Q_\phi^\pi(o,a)$,其中,$Q_\phi^\pi(o,a)$是智能体采用策略π,在观测o时,执行动作a所能获取的累计折扣奖励的期望,其中观测o是所有智能体观测o_i的集合,动作a是所有智能体动作a_i的集合;以及初始化策略网络;

将需要存储的数字图像添加随机噪声,智能体通过动作a修改突触权重矩阵并使用振荡神经网络振荡收敛得到奖励,产生样本(o,a,r,o'),其中,o表示所有智能体的观测,即$o=(o_1,o_2,\ldots,o_N)$,a表示所有智能体的动作$a=(a_1,a_2,\ldots,a_N)$,o'表示智能体i在执行动作a_i以后,所能到达的下一个观测o_i'的集合,i=1,2,3,...,N;r表示环境反馈的智能体奖励r_i的集合;

通过多智能体强化学习算法更新智能体策略$\pi^1,\pi_2,\ldots,\pi_N$,通过样本分别对策略网络和动作价值网络进行更新,直至模型收敛。

5.根据权利要求4所述的基于多智能体强化学习的振荡神经网络突触参数优化方法,其特征在于,策略网络的损失函数为:

$$\nabla_{\theta_i}J=\mathrm{E}_D\left[\nabla_{\theta_i}\pi_i(a_i|o_i)\nabla_{a_i}Q_\phi^\pi(o,a_1,\ldots,a_N)\big|_{a_i=\pi_i(o_i)}\right]$$

其中,$\nabla_{\theta_i}J$为策略网络的损失函数,E_D为经验缓存池,$\pi_i(a_i|o_i)$指智能体i从观测o_i选择动作a_i的概率,观测o_i是观测$o_{i,t}$在任一时刻的表示,动作a_i是动作$a_{i,t}$在任一时刻的表示,$Q_\phi^\pi(o,a)$是动作价值网络,用来估计在观测o下,采取动作a并且之后遵循策略π的情况下能获得的预期回报。

6.根据权利要求4所述的基于多智能体强化学习的振荡神经网络突触参数优化方法,其特征在于,对于动作价值网络$Q_\phi^\pi(o,a)$,计算部分可观测的马尔科夫决策过程算法的损失函数:

$$L_{MADDPG}(\phi)=\mathrm{E}_D\left[\left(Q_\phi^\pi(o,a_1,\ldots,a_N)-y\right)^2\right]$$

$$y=r_i+\gamma Q_\phi^{\pi'}(o',a_1',\ldots,a_N')\big|_{a_j'=\pi_j'(o_j)}$$

其中,$L_{MADDPG}(\varphi)$表示动作价值网络的损失,E_D为经验缓存池,y表示Q_ϕ^π的目标值,γ为折扣因子,r_i为智能体i的奖励,由ONN识别出的数字图像和输入的数字图像计算而出,体现了识别已存储数字图像的效果,$Q_\phi^{\pi'}(o',a_1',\ldots,a_N')$是对于下一个观测o'和动作序列a'的Q值的估计,$a'_j$是在下一步策略$\pi'_j$下得出的,动作序列a'是下一步的动作$a'_j$的集合,j=1,…,n。

7.一种计算机设备,其特征在于,包括:存储器和处理器,存储器中存储有可在处理器

上运行的计算机程序,处理器执行计算机程序时,实现权利要求1至6任一项方法的步骤。

8.一种计算机可读存储介质,其特征在于,存储介质上存储有计算机程序,该计算机程序被处理器执行时,实现权利要求1至6中任一项方法的步骤。

一种基于多智能体强化学习的振荡神经网络突触参数优化方法

技术领域

[0001] 本发明涉及神经形态计算技术领域,更具体的说是涉及基于多智能体强化学习的振荡神经网络数字图像识别方法。

背景技术

[0002] 近年来,数字图像识别技术得到了迅速发展,并在诸如自动驾驶、医疗影像分析、工业检测等多个领域得到了广泛应用。传统的图像识别方法多依赖于卷积神经网络(CNN)等深度学习模型,这些模型虽然性能卓越,但在计算资源和能耗方面存在较大需求,限制了其在资源受限环境中的应用。

[0003] 神经形态计算是受人脑结构和功能启发而发展的技术领域,旨在模拟人脑的神经网络来解决复杂的计算问题。在这一领域中,振荡神经网络(ONN)作为一种重要的实现方式,通过模拟人脑神经元的振荡同步行为,为高效的数据处理和模式识别提供了新的可能性。ONN利用相位同步来编码信息,有望在降低能耗和提高计算效率方面展现出巨大潜力,特别是在边缘计算领域。

[0004] ONN由于其低能耗和高效计算的特性,被认为是一种有潜力的替代方案。研究者们对ONN在数字图像识别中的应用表现出了浓厚兴趣,试图通过ONN的独特优势来提升图像识别的效率和准确性。例如,ONN已被应用于图像边缘检测、目标识别等任务中,并展示出了优异的性能。然而,尽管有诸多研究,ONN在实际应用中仍面临诸多挑战。传统上,突触权重的设计主要依赖于Hebbian学习规则,即如果一个神经元A经常足够接近地在时间上促使另一个神经元B激活,那么某些增长过程或代谢变化将发生在其中一个或两个神经元中,以致于A的效率作为B的一个激发因素将增加。这种学习机制在神经科学和神经网络理论中广泛采用,但在处理复杂的数字图像识别任务时存在局限性,如容易产生虚假模式且未能有效扩展存储容量或优化性能。

[0005] 因此,如何有效地优化振荡神经网络以应对数字图像识别任务中的复杂性,成为当前研究的重要方向之一。多智能体强化学习作为一种新的学习范式,通过多个智能体之间的协同和竞争,可以显著提升学习效率和模型的适应性,为解决传统方法的局限性提供了一种可行的途径。

发明内容

[0006] 有鉴于此,本发明提供了基于多智能体强化学习的振荡神经网络突触参数优化方法、设备及介质,在数字图像识别领域,这种方法使得单个网络能够识别更多的数字图像,并且增强了对复杂变化环境的动态适应性,显著提高了识别精度和处理效率。

[0007] 基于多智能体强化学习的振荡神经网络突触参数优化方法,包括:

[0008] 确定数字图像对应的存储模式,并作为学习任务;

[0009] 基于课程学习对学习任务进行拆分;

[0010] 基于拆分的数字图像识别任务集合,建立多智能体强化学习仿真环境;

[0011] 对多智能体强化学习进行训练,获得最优奖励的振荡神经网络突触权重;

[0012] 识别和存储所有预定的数字图像。

[0013] 优选的,确定数字图像对应的存储模式,具体包括:将振荡神经网络需要存储的数字图像转换为二值模式矩阵,-1代表黑色,1代表白色,二值模式矩阵的集合作为最终学习任务。

[0014] 优选的,基于拆分的数字图像识别任务集合,建立多智能体强化学习仿真环境具体包括:

[0015] 基于振荡神经网络定义状态、观测、动作、状态转移函数以及奖励,创建多智能体强化学习仿真环境,并将拆分的数字图像识别任务集合作为环境初始模式集合;

[0016] 状态S_t在每个时间步t捕捉了ONN中所有神经元的数字图像中间相位信息;

[0017] 每个智能体在时间步t的观测$o_{i,t}$提供了对ONN突触权重矩阵的部分观测;

[0018] 每个智能体的动作$a_{i,t}$反映了其对第i个智能体相关联的突触权重的修改的局部调整;

[0019] 第i个智能体从时间步长t到时间步长t+1的状态转移函数描述为:

[0020] $$S_{t+1} = S_t + \sum_{i=1}^{n} \left(a_{i,t} + a_{i,t}^{T}\right)$$

[0021] 式中,S_t和S_{t+1}分别表示时间步长t和t+1的突触权重矩阵,$a_{i,t}$表示第i个智能体在时间步长t的行动,是一个n维矢量,$a_{i,t}$被添加到突触权重矩阵S_t的第i列,$a_{i,t}^{T}$被添加到突触权重矩阵S_t的第i行,对应于与第i个神经元连接的权重;

[0022] 智能体i的奖励r_i计算公式为:

[0023] $r_i = r/n$

[0024] $$\varphi^{k^*} = \begin{bmatrix} \varphi_1^{k^*} \\ \varphi_2^{k^*} \\ \vdots \\ \varphi_n^{k^*} \end{bmatrix}, \delta\varphi^{k^*} = \varphi^{k^*} - \varphi_1^{k^*} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, k = 1,2,\ldots,P$$

$$r = P \times \mu + \sum_{k=1}^{P} - L_1\left(\delta\varphi^{k^*}, \delta\varphi^{k}\right)$$

[0025] $L_1(x,y) = |(x+2m\pi) - y|, x+2m\pi \in [-\pi,\pi]$

[0026] 上式中,P指当前任务存储数字图像的数量,φ^{k^*}代表第k个输入模式的输出数字图像中间相位,$\varphi_i^{k^*}$指的是φ^{k^*}的第i个元素,$\delta\varphi^{k}$表示第k个输入的数字图像的相位偏差,m是使得$x+2m\pi \in [-\pi,\pi]$的整数,x、y是两个不同的变量,n是智能体个数,μ表示用来区分不同任务的超参数。

[0027] 优选的,对多智能体强化学习进行训练具体包括:

[0028] 初始化动作价值网络$Q_{\phi}^{\pi}(o,a)$，其中，$Q_{\phi}^{\pi}(o,a)$是智能体采用策略π，在状态s时，执行动作a所能获取的累计折扣奖励的期望，其中观测o是所有智能体观测o_i的集合，动作a是所有智能体动作a_i的集合；以及初始化策略网络；

[0029] 将需要存储的数字图像添加随机噪声，智能体通过动作a修改突触权重矩阵并使用振荡神经网络振荡收敛得到奖励，产生样本(o,a,r,o')，其中，o表示所有智能体的观测，即$o=(o_1,o_2,\ldots,o_N)$，a表示所有智能体的动作$a=(a_1,a_2,\ldots,a_N)$，o'表示智能体i在执行动作a_i以后，所能到达的下一个观测o_i'的集合，i＝1,2,3,...,N；r表示环境反馈的智能体奖励r_i的集合；

[0030] 通过多智能体强化学习算法更新智能体策略$\pi_1,\pi_2,\ldots,\pi_N$，通过样本分别对策略网络和动作价值网络进行更新，直至模型收敛。

[0031] 优选的，策略网络的损失函数为：

[0032] $$\nabla_{\theta_i}J=\mathrm{E}_{D}\left[\nabla_{\theta_i}\pi_i(a_i|o_i)\nabla_{a_i}Q_{\phi}^{\pi}(o,a_1,\ldots,a_N)\big|_{a_i=\pi_i(o_i)}\right]$$

[0033] 其中，$\nabla_{\theta_i}J$为策略网络的损失函数，E_D为经验缓存池，$\pi_i(a_i|o_i)$指智能体i从观测o_i选择动作a_i的概率，观测o_i是观测$o_{i,t}$在任一时刻的表示，动作a_i是动作$a_{i,t}$在任一时刻的表示，$Q_{\phi}^{\pi}(o,a)$是动作价值网络，用来估计在观测o下，采取动作a并且之后遵循策略π的情况下能获得的预期回报。

[0034] 优选的，对于动作价值网络$Q_{\phi}^{\pi}(o,a)$，计算部分可观测的马尔科夫决策过程算法的损失函数：

[0035] $$L_{MADDPG}(\phi)=\mathrm{E}_{D}\left[\left(Q_{\phi}^{\pi}(o,a_1,\ldots,a_N)-y\right)^2\right]$$

$$y=r_i+\gamma Q_{\phi}^{\pi'}(o',a_1',\ldots,a_N')\big|_{a_j'=\pi_j'(o_j)}$$

[0036] 其中，$L_{MADDPG}(\varphi)$表示动作价值网络的损失，E_D为经验缓存池，y表示Q_{ϕ}^{π}的目标值，γ为折扣因子，r_i为智能体i的奖励，由ONN识别出的数字图像和输入的数字图像计算而出，体现了识别已存储数字图像的效果，$Q_{\phi}^{\pi'}(o',a_1',\ldots,a_N')$是对于下一个观测o'和动作序列a'的Q值的估计，$a_j$'是在下一步策略$\pi'_j$下得出的，动作序列a'是下一步的动作$a_j$'的集合，j＝1,…,n。

[0037] 一种计算机设备，包括：存储器和处理器，存储器中存储有可在处理器上运行的计算机程序，处理器执行计算机程序时，实现基于多智能体强化学习的振荡神经网络突触参数优化方法的步骤。

[0038] 一种计算机可读存储介质，存储介质上存储有计算机程序，该计算机程序被处理器执行时，实现基于多智能体强化学习的振荡神经网络突触参数优化方法的步骤。

[0039] 经由上述的技术方案可知，与现有技术相比，本发明公开提供了基于多智能体强化学习的振荡神经网络突触参数优化方法、设备及介质，解决了边缘设备单个网络难以存储足量的数字图像以及难以动态调整存储需求的不足。

附图说明

[0040] 为了更清楚地说明本发明实施例或现有技术中的技术方案,下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍,显而易见地,下面描述中的附图仅仅是本发明的实施例,对于本领域普通技术人员来讲,在不付出创造性劳动的前提下,还可以根据提供的附图获得其他的附图。

[0041] 图1为本发明提供的基于多智能体强化学习的振荡神经网络数字图像识别优化方法流程图。

[0042] 图2为本发明提供的强化学习训练流程图;

[0043] 图3为本发明提供的ONN结构图;

[0044] 图4为本发明提供的课程学习流程图;

[0045] 图5为本发明提供的状态矩阵示意图,紫色为观测矩阵,黄色为使用0遮盖用于区分智能体的部分。

具体实施方式

[0046] 下面将结合本发明实施例中的附图,对本发明实施例中的技术方案进行清楚、完整地描述,显然,所描述的实施例仅仅是本发明一部分实施例,而不是全部的实施例。基于本发明中的实施例,本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例,都属于本发明保护的范围。

[0047] 本实施例采用的软硬件平台包括具有高性能计算能力的服务器,搭载适用于机器学习的高级图形处理单元(GPU)。软件环境包括支持Python编程的开发环境,以及用于实现MARL的专业库,如PyTorch和MARLlib。仿真环境设置旨在模拟实际振荡神经网络ONN运行条件,包括各种复杂度的模式识别任务。

[0048] 本发明一种基于多智能体强化学习的振荡神经网络突触参数优化方法,如图1所示,具体步骤如下:

[0049] 步骤1:确定数字图像对应的存储模式,并作为学习任务

[0050] 将ONN需要存储的数字图像转换为二值模式矩阵,-1代表黑色,1代表白色。这些二值模式矩阵的集合作为最终学习任务D,代表了网络应能够识别或存储的信息。

[0051] 步骤2:使用课程学习对学习任务进行拆分

[0052] 将步骤1中确定的学习任务D使用课程学习进行任务拆分,从简单到复杂,增加网络在学习过程中的适应性和效率。如图4所示,将学习任务D分割成n个子集$D=\langle d_1,d_2,\cdots,d_n\rangle$,子任务集$d_i$的模式复杂度逐渐提高。初始化训练任务$T_1=d_1$,随机初始化初始策略$\pi_1$。第k次将任务子集$d_k$添加到训练任务中。然后,根据步骤4中的强化学习算法和策略π_k,在任务T_k上进行训练直到收敛,训练完成后,存储学习到的策略π_k并用于初始化策略π_{k+1}。课程学习的方法可以使得训练的奖励逐步提高。

[0053] 步骤3:基于拆分的图像识别任务集合,创建多智能体强化学习仿真环境创建多智能体强化学习仿真环境需要定义状态,观测,动作,状态转移函数以及奖励;并根据步骤2课程学习拆分的数字图像识别任务集合设定环境初始模式集合。

[0054] 为解决多个数字图像难以同时存储以及不同数字图像识别任务需求不一致的问题,将振荡神经网络ONN突触权重矩阵调整问题定义为一个部分可观察马尔可夫决策过程

POMDP。其中多个智能体协同工作,通过相互作用学习如何根据数字图像识别任务需求动态调整ONN的突触权重并存储更多数字图像。POMDP是标准MDP的扩展,它考虑了部分可观测性的特征。形式上,POMDP表示为N,S,O,A,P,R,γ,其中N表示有限的智能体(ONN神经元)集合,S代表状态,O代表观测,A表示行动,P是状态转移概率,R是奖励函数,γ∈0,1是折扣因子。

[0055] 这种仿真环境允许在不同任务需求的动态变化下,学习并优化如何存储和处理更多的数字图像,解决了传统方法中图像存储容量有限和处理能力不足的问题。在此基础上,图像处理任务中的结果得到了显著提升,能够更高效地识别和分类复杂的数字图像,提高了系统的整体性能和适应性。

[0056] ONN的结构如图3所示,w_{ij}表示突触权重矩阵第i行j列元素值,每个相位控制振荡器代表一个神经元。其将任务中选定输入的数字图像通过突触权重w_{ij}计算并经过相位控制振荡器振荡收敛,使其趋向存储数字图像的相位。突触权重决定了神经元间的相互作用强度。在ONN中,第i个神经元输出的总相位可以表示为$\theta_i(t) = \Omega t + \varphi_i$,其中$\Omega$是自然频率,$\varphi_i$是数字图像相位偏差。ONN第i个神经元的动态$\dot{\varphi}_i$可以通过以下方程描述,这里,$H(\varphi_j - \varphi_i)$是循环滤波器的平均值,n是突触权重矩阵维度:

[0057] $$\dot{\varphi}_i = \sum_{j=1}^{n} w_{ij} H(\varphi_j - \varphi_i)$$

[0058] 在ONN结构的基础上定义POMDP,POMDP的详细内容如下:

[0059] 状态(State)

[0060] 如图5所示,系统的状态S_t在每个时间步t捕捉了ONN中所有神经元的数字图像中间相位信息,这些信息直接受突触权重w_{ij}的影响。状态的定义依据ONN的动态性质,确保了在MARL框架内能够全面反映不同数字图像存储任务的差异。

[0061] 观测(Observation)

[0062] 如图5所示,每个智能体在时间步t的观测$o_{i,t}$提供了对ONN突触权重矩阵的部分观测,对于第i个智能体,突触权重矩阵S_t的第i行和第i列被遮盖。这种部分观测性质允许智能体基于局部信息做出决策,从而降低了问题的复杂性,并允许智能体专注于与其直接相关的突触权重的调整。

[0063] 动作(Action)

[0064] 每个智能体的动作$a_{i,t}$反映了其对第i个智能体相关联的突触权重的修改的局部调整。这些调整直接影响ONN中相应神经元的相位行为,进而影响整个网络的相位同步和数字图像存储能力。

[0065] 状态转移函数(State Transition Function)

[0066] 状态转移是由所有智能体的行动对ONN突触权重矩阵的累积效果决定的。第i个智能体从时间步长t到t+1的状态转移函数描述为:

[0067] $$S_{t+1} = S_t + \sum_{i=1}^{n} \left(a_{i,t} + a_{i,t}^{T}\right)$$

[0068] 这里,S_t和S_{t+1}分别表示时间步长t和t+1的突触权重矩阵。$a_{i,t}$表示第i个智能体在

时间步长t的行动,这是一个n维矢量,n等于智能体个数,智能体个数和突触权重矩阵维度一致,$a_{i,t}$被添加到突触权重矩阵S_t的第i列,$a_{i,t}^T$被添加到突触权重矩阵S_t的第i行。这个行动矢量被添加到权重矩阵S_t的第i行和列,对应于与第i个神经元连接的权重。

[0069] 奖励(Reward)

[0070] 奖励机制r_i旨在激励智能体找到最优的突触权重调整策略,这些策略将使数字图像存储任务达到最佳效果。奖励的计算基于数字图像存储效果的评估。

[0071] 每个独立智能体的奖励r_i被定义为系统的奖励的平均。因此,智能体i的奖励r_i计算如下,其中μ是用来区分不同任务的超参数,n是智能体个数:

[0072] $r_i = r/n$

[0073] $$\varphi^{k^*} = \begin{bmatrix} \varphi_1^{k^*} \\ \varphi_2^{k^*} \\ \vdots \\ \varphi_n^{k^*} \end{bmatrix}, \delta\varphi^{k^*} = \varphi^{k^*} - \varphi_1^{k^*} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, k = 1,2,\ldots,P$$

$$r = P \times \mu + \sum_{k=1}^{P} - L_1\left(\delta\varphi^{k^*}, \delta\varphi^{k}\right)$$

[0074] $L_1(x,y) = |(x+2m\pi) - y|, x+2m\pi \in [-\pi,\pi]$

[0075] 在上述公式中,P指当前任务存储数字图像的数量,φ^{k^*}代表第k个输入模式的输出数字图像中间相位,$\varphi_i^{k^*}$指的是φ^{k^*}的第i个元素,$\delta\varphi^k$表示第k个输入的数字图像的相位偏差,m是使得$x+2m\pi \in [-\pi,\pi]$的整数。

[0076] 步骤4:对多智能体强化学习进行训练,获得最优奖励的振荡神经网络突触权重

[0077] 如图2所示,在步骤3中创建的仿真环境中执行MARL算法,让多个智能体通过探索和利用策略来寻找突触权重,学习如何根据任务需求动态调整ONN的突触权重并存储更多数字图像。

[0078] 初始化多智能体的策略网络:初始化N个智能体策略$\pi_1,\pi_2,\cdots,\pi_N$,其中π_i是将观测o_i映射为动作a_i的函数,即$a_i = \pi_i(o_i)$。具体地,使用参数为θ_i的神经网络代表策略π_i。初始化动作价值网络$Q_\phi^\pi(o,a)$,其中$Q_\phi^\pi(o,a)$是多智能体采用策略π,在状态s时,执行动作a所能获取的累计折扣奖励的期望,使用参数为φ的神经网络代表动作价值网络$Q_\phi^\pi(o,a)$。

[0079] 使用智能体策略$\pi_1,\pi_2,\ldots,\pi_N$和环境交互。先将需要存储的数字图像添加随机噪声,智能体使用动作a修改突触权重矩阵并使用ONN对输入的数字图像进行振荡收敛并得到奖励,产生样本(o,a,r,o'),其中o表示所有智能体的观测,即$o = (o_1,o_2,\ldots,o_N)$,a表示所有智能体的动作$a = (a_1,a_2,\ldots,a_N)$,o'表示智能体i在执行动作a_i以后,所能到达的下一个观测o_i'的集合,i=1,2,3,...,N;r表示环境反馈的智能体奖励r_i的集合。此处的观测o为步骤3中所定义,表示实时变化的突触矩阵权重;动作a为步骤3中所定义,用于修改突触权重;奖励r为步骤3中所定义,由ONN识别出的数字图像和输入的数字图像计算而出,体现了识别已存储数字图像的效果,用于指导策略训练。

[0080] 使用多智能体强化学习算法更新智能体策略$\pi_1,\pi_2,...,\pi_N$。使用多智能体强化学习算法利用采集的样本计算损失分别对策略网络和动作价值网络进行更新。

[0081] 以典型的多智能体强化学习算法MADDPG为例,描述更新过程。

[0082] 策略网络的损失函数为:

[0083] $$\nabla_{\theta_i} J = \mathrm{E}_D\left[\nabla_{\theta_i}\pi_i\left(a_i|\ o_i\right)\nabla_{a_i} Q_\phi^\pi\left(o,a_1,\ldots,a_N\right)\Big|_{a_i=\pi_i(o_i)}\right]$$

[0084] 其中$\nabla_{\theta_i} J$为策略网络的损失函数,E_D为经验缓存池,$\pi_i(a_i|o_i)$为智能体i从观测o_i选择动作a_i的概率,$Q_\phi^\pi(o,a)$是动作价值网络,用来估计在观测o下,采取动作a并且之后遵循策略π的情况下能获得的预期回报。

[0085] 对于动作价值网络$Q_\phi^\pi(o,a)$,计算MADDPG算法的损失函数:

$$L_{MADDPG}(\phi) = \mathrm{E}_D\left[\left(Q_\phi^\pi\left(o,a_1,\ldots,a_N\right)-y\right)^2\right]$$

[0086]

$$y = r_i + \gamma Q_\phi^{\pi'}\left(o',a_1',\ldots,a_N'\right)\Big|_{a_j'=\pi_j'(o_j)}$$

[0087] 其中,$L_{MADDPG}(\varphi)$表示动作价值网络的损失,E_D为经验缓存池,y表示Q_ϕ^π的目标值,γ为折扣因子一般设置为0.99,r_i为智能体i的奖励,由ONN识别出的数字图像和输入的数字图像计算而出,体现了识别已存储数字图像的效果。

[0088] 重复多轮交互和更新直到收敛。

[0089] 在训练过程中持续评估数字图像存储效果,记录并保存产生最佳奖励结果的突触权重设置。

[0090] 步骤5:重复第2步直到存储所有需要的数字图像

[0091] 循环执行步骤2-4的过程,直至网络能够有效存储和识别所有预定的数字图像。

[0092] 基于本发明上述方法,当应用在工厂作业环境下,需对产品质量进行检测。图像输入为边缘端设备拍摄的产品外观图像,并执行上述步骤1-步骤5,并可通过ONN识别产品的轮廓,是否出现缺陷、裂纹、缺口或其他不符合标准的形状,从而确保产品质量。

[0093] 本实施例提供了一种计算机设备,包括:存储器和处理器,存储器中存储有可在处理器上运行的计算机程序,处理器执行计算机程序时,实现一种基于多智能体强化学习的振荡神经网络突触参数优化方法的步骤。

[0094] 本实施例提供了一种计算机可读存储介质,存储介质上存储有计算机程序,该计算机程序被处理器执行时,实现一种基于多智能体强化学习的振荡神经网络突触参数优化方法的步骤。

[0095] 本领域普通技术人员可以理解:实现上述方法实施例的全部或部分步骤可以通过程序指令相关的硬件来完成,前述的程序可以存储于计算机可读取存储介质中,该程序在执行时,执行包括上述方法实施例的步骤;而前述的存储介质包括:移动存储设备、只读存储器(Read-Only Memory,ROM)、随机存取存储器(Random Access Memory,RAM)、磁碟或者光盘等各种可以存储程序代码的介质。

[0096] 本说明书中各个实施例采用递进的方式描述,每个实施例重点说明的都是与其他实施例的不同之处,各个实施例之间相同相似部分互相参见即可。对于实施例公开的装置而言,由于其与实施例公开的方法相对应,所以描述的比较简单,相关之处参见方法部分说明即可。

[0097] 对所公开的实施例的上述说明,使本领域专业技术人员能够实现或使用本发明。对这些实施例的多种修改对本领域的专业技术人员来说将是显而易见的,本文中所定义的一般原理可以在不脱离本发明的精神或范围的情况下,在其它实施例中实现。因此,本发明将不会被限制于本文所示的这些实施例,而是要符合与本文所公开的原理和新颖特点相一致的最宽的范围。

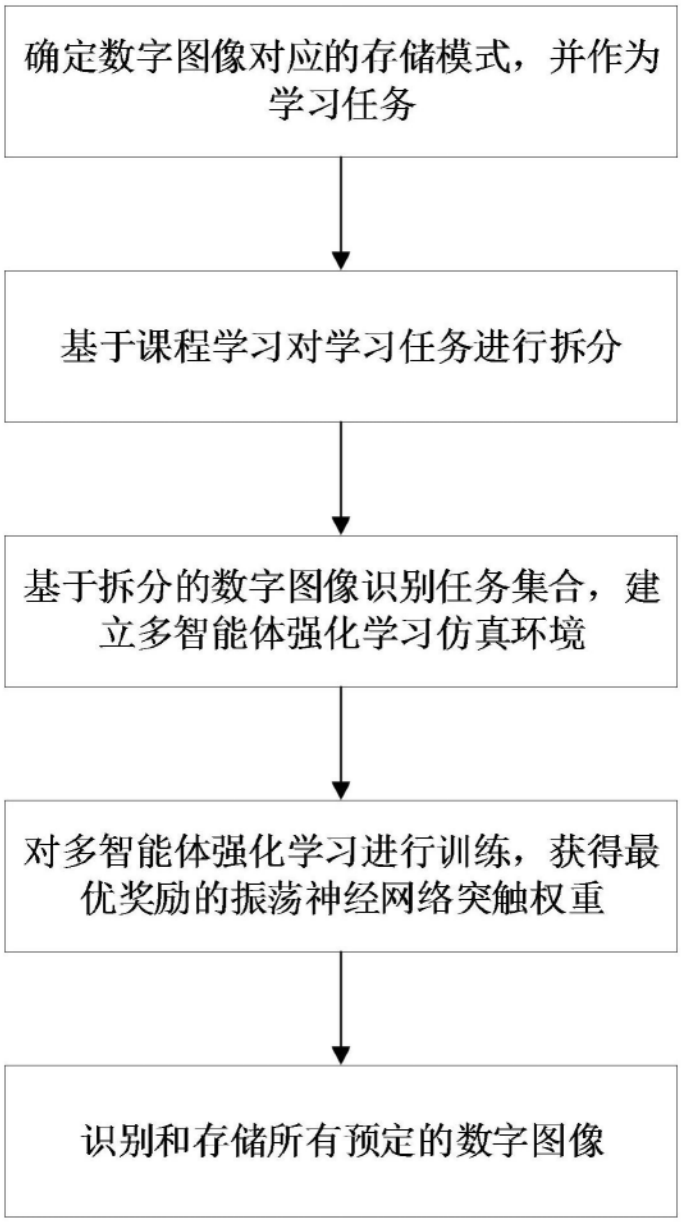

确定数字图像对应的存储模式，并作为学习任务
基于课程学习对学习任务进行拆分
基于拆分的数字图像识别任务集合，建立多智能体强化学习仿真环境
对多智能体强化学习进行训练，获得最优奖励的振荡神经网络突触权重
识别和存储所有预定的数字图像

图1

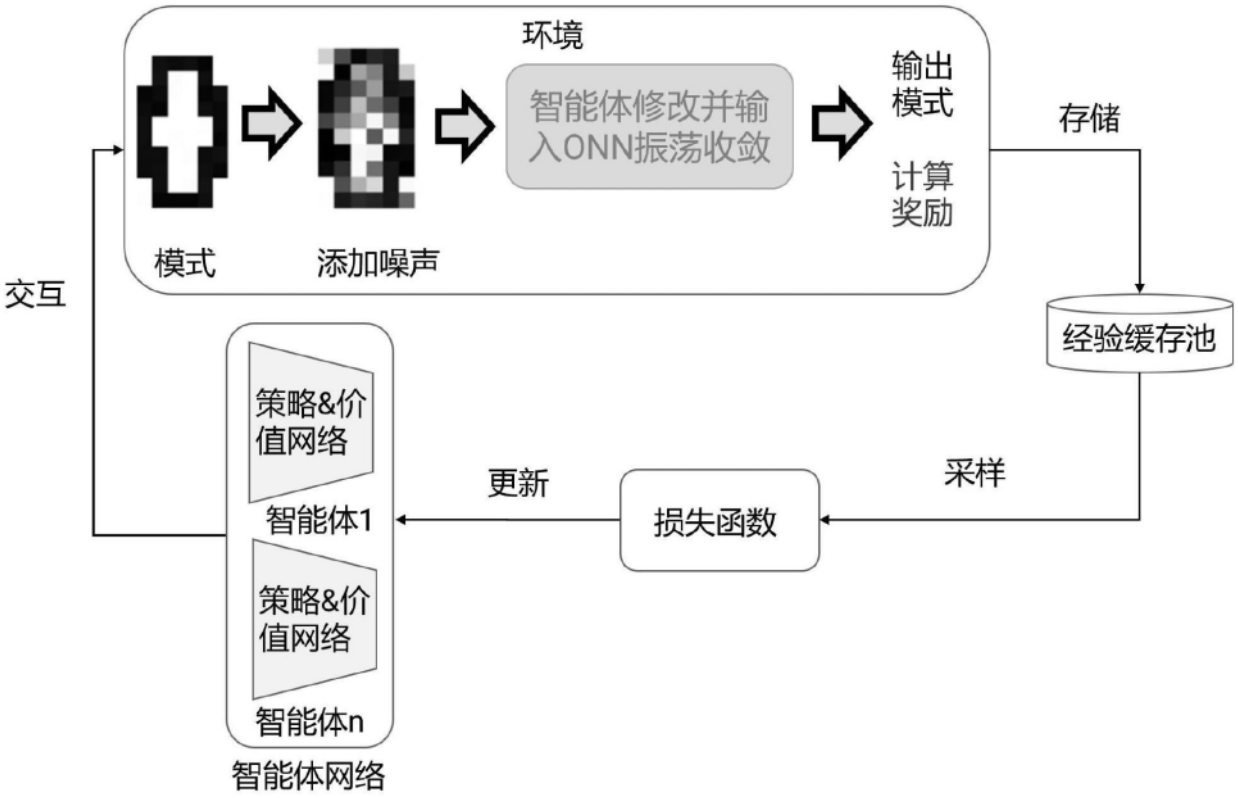

环境
智能体修改并输入ONN振荡收敛
输出模式
计算奖励
存储
模式
添加噪声
交互
经验缓存池
策略&价值网络
智能体1
策略&价值网络
智能体n
智能体网络
更新
损失函数
采样

图2

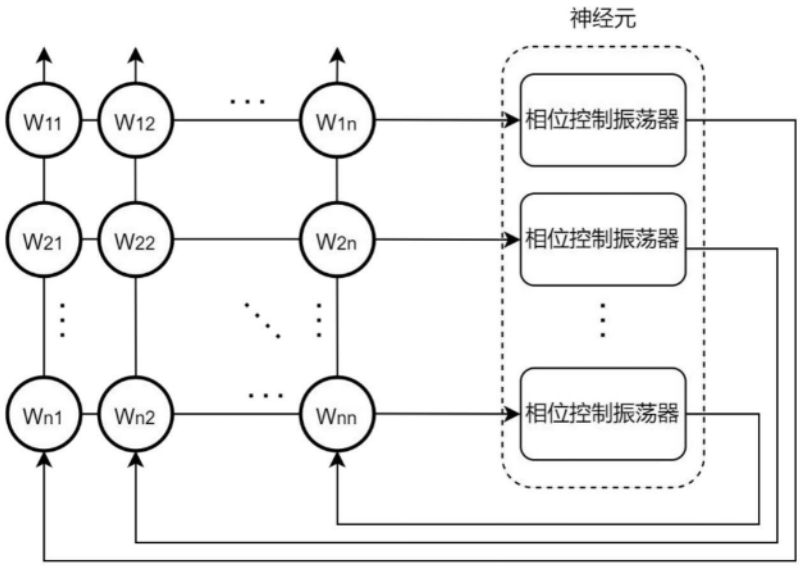

神经元
W11
W12
W1n
W21
W22
W2n
Wn1
Wn2
Wnn
相位控制振荡器
相位控制振荡器
相位控制振荡器

图3

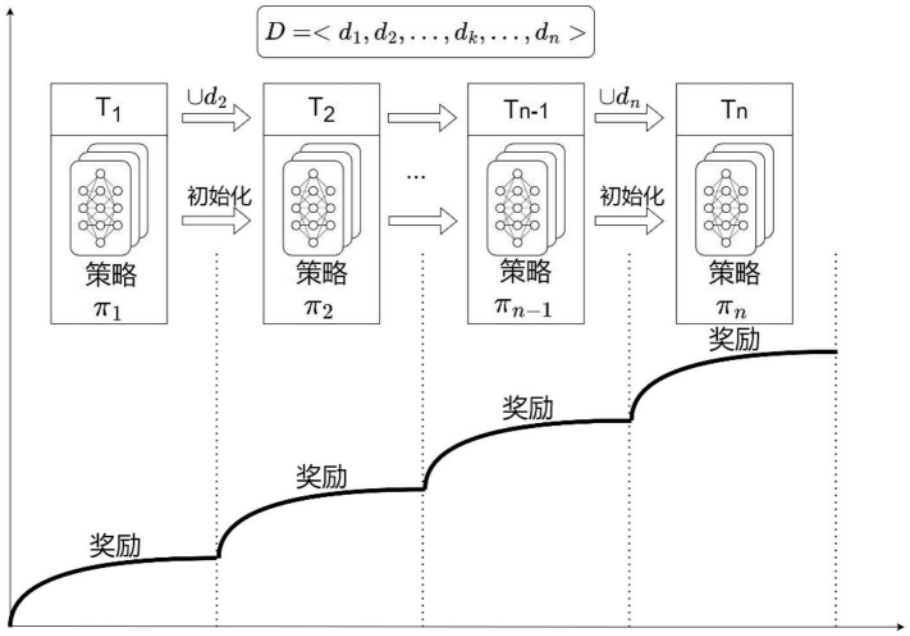

$D =< d_1, d_2, \ldots, d_k, \ldots, d_n >$
T1
$\cup d_2$
T2
Tn-1
$\cup d_n$
Tn
初始化
初始化
策略
π_1
策略
π_2
策略
π_{n-1}
策略
π_n
奖励
奖励
奖励
奖励

图4

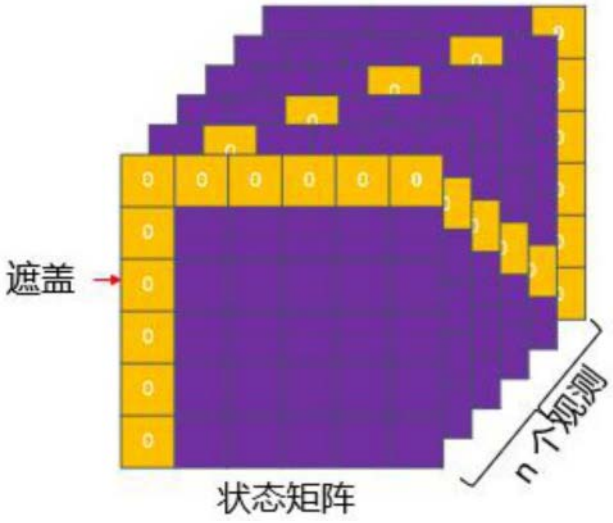

遮盖
n个观测
状态矩阵

图5