(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2020 年 2 月 13 日 (13.02.2020)

WIPO | PCT

(10) 国际公布号
WO 2020/029095 A1

(51) 国际专利分类号:
G06N 3/08 (2006.01)

(21) 国际申请号: PCT/CN2018/099256

(22) 国际申请日: 2018 年 8 月 7 日 (07.08.2018)

(25) 申请语言: 中文

(26) 公布语言: 中文

(71) 申请人: 中国科学院深圳先进技术研究院 (**SHENZHEN INSTITUTES OF ADVANCED TECHNOLOGY CHINESE ACADEMY OF SCIENCES**) [CN/CN]; 中国广东省深圳市南山区深圳大学城学苑大道 1068 号, Guangdong 518055 (CN)。

(72) 发明人: 王峥 (**WANG, Zheng**); 中国广东省深圳市南山区深圳大学城学苑大道1068号, Guangdong 518055 (CN)。 梁明兰 (**LIANG, Minglan**); 中国广东省深圳市南山区深圳大学城学苑大道1068号, Guangdong 518055 (CN)。

(74) 代理人: 深圳智趣知识产权代理事务所 (普通合伙) (**SHENZHEN IPLUS INTELLECTUAL PROPERTY AGENCY (GENERAL PARTNERSHIP)**); 中国广东省深圳市福田区梅林街道梅都社区梅林路48号理想公馆2427, Guangdong 518000 (CN)。

(81) 指定国(除另有指明，要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国(除另有指明，要求每一种可提供的地区保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM,

(54) **Title:** REINFORCEMENT LEARNING NETWORK TRAINING METHOD, APPARATUS AND DEVICE, AND STORAGE MEDIUM

(54) 发明名称: 强化学习网络的训练方法、装置、训练设备及存储介质

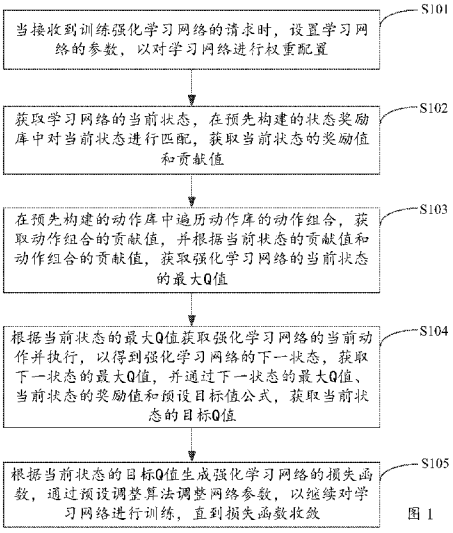

图 1

S101 Upon receipt of a request for training of a reinforcement learning network, set network parameters of the learning network, so as to perform weight configuration on the learning network;

S102 Acquire the current state of the learning network, perform, in a pre-formed state reward library, matching on the current state, and acquire the reward value and the contribution value of the current state;

S103 Traverse action combinations in a pre-formed action library, acquire contribution values of the action combinations, and acquire the maximum Q value of the current state of the reinforcement learning network according to the contribution value of the current state and the contribution values of the action combinations;

S104 Acquire the current action of the reinforcement learning network according to the maximum Q value of the current state and execute same, so as to obtain the next state of the reinforcement learning network, acquire the maximum Q of the next state, and acquire a target Q value of the current state by means of the maximum Q value of the next state, the reward value of the current state, and a preset target value formula;

S105 Generate a loss function of the reinforcement learning network according to the target Q value of the current state, and adjust the network parameters by means of a preset adjustment algorithm, so as to continue to train the learning network until the loss function converges.

(57) **Abstract:** The present invention is applicable to the technical field of machine learning, and provides a reinforcement learning network training method, apparatus and device, and a storage medium. Said method comprises: upon receipt of a request for training of a reinforcement learning network, setting network parameters of the reinforcement learning network, so as to perform weight configuration; acquiring the current state of the reinforcement learning network, and the reward value and the contribution value of the current state; acquiring the maximum Q value of the action combination in the current state by traversing action combinations in an action library; acquiring the current action according to the maximum Q value of the current state and executing same, and acquiring a target Q value of the current state by obtaining the maximum Q value of a next state; and generating a loss function of the reinforcement learning network, and adjusting the network parameters by means of a preset adjustment algorithm, so as to continue to train the reinforcement learning network until the loss function converges. The present invention reduces the calculation amount for reinforcement learning network training, thereby increasing the training speed of a reinforcement learning network and improving the training efficiency.

WO 2020/029095 A1

[见续页]

AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

(57) 摘要: 本发明适用机器学习领域，提供了一种强化学习网络的训练方法、装置、训练设备及存储介质，该方法包括：当接收到训练强化学习网络的请求时，设置强化学习网络的网络参数，以进行权重配置，获取强化学习网络的当前状态，以及当前状态的奖励值和贡献值，通过遍历动作库的动作组合，获取当前状态下的动作组合的最大Q值，根据当前状态的最大Q值获取当前动作并执行，通过得到下一状态的最大Q值，获取当前状态的目标Q值，生成强化学习网络的损失函数，通过预设调整算法调整网络参数，以继续对强化学习网络进行训练，直到损失函数收敛，从而降低了训练强化学习网络的计算量，进而加快了强化学习网络的训练速度、提高了训练效率。

强化学习网络的训练方法、装置、训练设备及存储介质

技术领域

本发明属于机器学习领域，尤其涉及一种强化学习网络的训练方法、装置、训练设备及存储介质。

背景技术

强化学习(reinforcement learning)，又称再励学习、评价学习，是一种重要的机器学习方法，是智能体（Agent）从环境到行为映射的学习，以使奖励信号(强化信号)函数值最大，强化学习不同于连接主义学习中的监督学习，主要表现在教师信号上，强化学习中由环境提供的强化信号是对产生动作的好坏作一种评价(通常为标量信号)，而不是告诉强化学习系统 RLS(reinforcement learning system)如何去产生正确的动作。由于外部环境提供的信息很少，RLS 必须靠自身的经历进行学习。通过这种方式，RLS 在行动-评价的环境中获得知识，改进行动方案以适应环境在智能控制机器人及分析预测等领域有许多应用。

近年来，强化学习广泛应用于机器人控制领域、计算机视觉领域、自然语言处理、博弈论领域、自动驾驶。训练强化学习网络过程通常在 CPU 与 GPU 设备上实现，其计算量相当大，在实际应用过程中，存在着占用资源多、运算速度慢、效率低等问题，并且因内存访问带宽的限制导致计算能力无法进一步提升。

发明内容

本发明的目的在于提供一种强化学习网络的训练方法、装置、训练设备以及存储介质，旨在解决由于现有技术无法提供一种有效的强化学习网络的训练

方法，导致训练计算量大、效率低的问题。

一方面，本发明提供了一种强化学习网络的训练方法，所述方法包括下述步骤：

当接收到训练强化学习网络的请求时，设置所述强化学习网络的网络参数，以对所述强化学习网络进行权重配置；

获取所述强化学习网络的当前状态，在预先构建的状态奖励库中对所述当前状态进行匹配，获取所述当前状态的奖励值和贡献值；

在预先构建的动作库中遍历所述动作库的动作组合，获取所述动作组合的贡献值，并根据所述当前状态的贡献值和所述动作组合的贡献值，获取所述强化学习网络的当前状态的最大 Q 值；

根据所述当前状态的最大 Q 值获取所述强化学习网络的当前动作并执行，以使所述强化学习网络进入下一状态，获取所述下一状态的最大 Q 值，并通过所述下一状态的最大 Q 值、所述当前状态的奖励值和预设目标值公式，获取所述当前状态的目标 Q 值；

根据所述当前状态的目标 Q 值生成所述强化学习网络的损失函数，通过预设调整算法调整所述强化学习网络的网络参数，以继续对所述强化学习网络进行训练，直到所述损失函数收敛。

另一方面，本发明提供了一种强化学习网络的训练装置，所述装置包括：

参数设置单元，用于当接收到训练强化学习网络的请求时，设置所述强化学习网络的网络参数，以对所述强化学习网络进行权重配置；

匹配获取单元，用于获取所述强化学习网络的当前状态，在预先构建的状态奖励库中对所述当前状态进行匹配，获取所述当前状态的奖励值和贡献值；

遍历获取单元，用于在预先构建的动作库中遍历所述动作库的动作组合，获取所述动作组合的贡献值，并根据所述当前状态的贡献值和所述动作组合的贡献值，获取所述强化学习网络的当前状态的最大 Q 值；

执行获取单元，用于根据所述当前状态的最大 Q 值获取所述强化学习网络

的当前动作并执行，以使所述强化学习网络进入下一状态，获取所述下一状态的最大 Q 值，并通过所述下一状态的最大 Q 值、所述当前状态的奖励值和预设目标值公式，获取所述当前状态的目标 Q 值；以及

生成调整单元，用于根据所述强化学习网络的目标 Q 值生成所述强化学习网络的损失函数，通过预设调整算法调整所述强化学习网络的网络参数，以继续对所述强化学习网络进行训练，直到所述损失函数收敛。

另一方面，本发明还提供了一种强化学习网络训练设备，包括存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机程序，所述处理器执行所述计算机程序时实现如上述强化学习网络的训练方法的步骤。

另一方面，本发明还提供了一种计算机可读存储介质，所述计算机可读存储介质存储有计算机程序，所述计算机程序被处理器执行时实现如上述强化学习网络的训练方法的步骤。

本发明当接收到训练强化学习网络的请求时，设置强化学习网络的网络参数，以进行权重配置，获取强化学习网络的当前状态，以及当前状态的奖励值和贡献值，通过遍历动作库的动作组合，获取当前状态下的动作组合的最大 Q 值，根据当前状态的最大 Q 值获取当前动作并执行，通过得到下一状态的最大 Q 值，获取当前状态的目标 Q 值，生成强化学习网络的损失函数，通过预设调整算法调整网络参数，以继续对强化学习网络进行训练，直到损失函数收敛，从而降低了训练强化学习网络的计算量，进而加快了强化学习网络的训练速度、提高了训练效率。

附图说明

图 1 是本发明实施例一提供的强化学习网络的训练方法的实现流程图；

图 2 是本发明实施例一提供的状态奖励库的优选存储结构示意图；

图 3 是本发明实施例一提供的动作库的优选存储结构示意图；

图 4 是本发明实施例二提供的强化学习网络的训练装置的结构示意图；

图 5 是本发明实施例三提供的强化学习网络的训练装置的结构示意图；

图 6 是本发明实施例四提供的一种强化学习网络训练设备的结构示意图；以及

图 7 是本发明实施例四提供的一种强化学习网络训练设备的优选结构示意图。

具体实施方式

为了使本发明的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本发明进行进一步详细说明。应当理解，此处所描述的具体实施例仅仅用以解释本发明，并不用于限定本发明。

以下结合具体实施例对本发明的具体实现进行详细描述：

实施例一：

图 1 示出了本发明实施例一提供的强化学习网络的训练方法的实现流程，为了便于说明，仅示出了与本发明实施例相关的部分，详述如下：

在步骤 S101 中，当接收到训练强化学习网络的请求时，设置强化学习网络的网络参数，以对强化学习网络进行权重配置。

本发明实施例适用于强化学习网络训练设备，例如，MATLAB（Matrix Laboratory，矩阵实验室）等训练设备。在本发明实施例中，当接收到训练强化学习网络的请求时，设置学习网络的网络参数，以对学习网络进行权重配置，具体地，先写入网络参数，当进行网络运算时，根据写入的网络参数启动强化学习网络相应神经元的计算模式，通过这种方式来配置每层网络的每个神经元的参数，从而实现数据并行处理，进而提高了数据处理效率。

在步骤 S102 中，获取强化学习网络的当前状态，在预先构建的状态奖励库中对当前状态进行匹配，获取当前状态的奖励值和贡献值。

在本发明实施例中，状态奖励库为预先构建的存储了状态节点和对应奖励值的集合，在接收到训练请求之后，获取强化学习网络的当前状态，并提取当

前状态的特征数据，通过该当前状态的特征数据计算得到当前状态的贡献值，然后，在状态奖励库中对当前状态进行匹配，得到当前状态的奖励值。

作为示例地，如图 2 所示，图中示出了状态奖励库的优选存储结构，状态奖励库分为 n 个奖励组，分别对应 n 个特殊状态的奖励值，数据的开头存储了奖励值组数 n，数据库的结尾存储了一般状态的奖励值，即第（n+1）个奖励值，每一个奖励组都包括不同的状态节点，即不同状态值，不同的状态节点对应着不同范围的状态值。

优选地，在预先构建的状态奖励库中对当前状态进行匹配时，将当前状态与状态奖励库中的预设数量个奖励组对应的所有状态节点进行匹配，当当前状态位于预设数量个奖励组中预设状态节点中时，将预设状态奖励组的奖励值设置为当前状态的奖励值，否则将当前状态的奖励值设置为预设一般状态奖励值，从而快速获取当前状态的即时奖励。具体地，由于当前状态只能位于一个状态节点中，或者，当前状态位于所有状态节点外，因此，在匹配状态节点时，可采用逐一匹配状态节点的方法进行匹配，当当前状态位于预设状态节点中时，停止匹配其他的状态节点，并将预设状态节点对应的奖励值设置为当前状态的奖励值，当逐一匹配所有状态节点后，都没有匹配成功，则将一般状态奖励值设置为当前状态的奖励值。

在步骤 S103 中，在预先构建的动作库中遍历动作库的动作组合，获取动作组合的贡献值，并根据当前状态的贡献值和动作组合的贡献值，获取强化学习网络的当前状态的最大 Q 值。

在本发明实施例中，动作库为预先构建的存储了学习网络可输出的所有动作的集合，Q 值为强化学习网络中状态映射到动作值的表征，遍历动作库的所有动作组合，获取每个动作组合（实时动作）的贡献值，在遍历动作库的动作组合时，每得到一个动作组合，将通过当前状态的贡献值和动作组合的贡献值计算每一个动作组合的Q值，从而可获得强化学习网络的当前状态的最大Q值。

作为示例地，如图 3 所示，图中示出了动作库的优选存储结构，动作库分

为动作内存模块和实时动作内存模块，动作内存模块用于存储所有动作的信息，具体有动作维数 n、每个动作维数的步长值、最大值和起始值，实时动作内存模块用于存储即将输出的动作信息，具体为 n 维动作中每个动作的动作值，作为示例地，在自动驾驶的强化学习网络中，动作有左转（第一维）、右转（第二维）、刹车（第三维）等，对应的动作值为（1,a）、（2,b）、（3,c）其中，1、2、3 分别代表动作的维度（例如，第一维、第二维和第三维），a、b、c 分别为第一、二、三维动作对应的度量值。

优选地，在预先构建的动作库中遍历动作库的动作组合时，将动作库中预设动作列表上的预设数量维动作的起始值，依次设置为动作库中预设实时动作表上的预设数量个实时动作值，获取预设动作列表上的预设第一维动作的步长值，并将预设第一维动作的步长值逐次累加到预设第一维动作对应的实时动作值，当对应的实时动作值逐次累加到预设第一维动作对应的范围之外时，获取预设动作列表上的预设第二维动作的步长值，并将预设第二维动作的步长值逐次累加到预设第二维动作对应的实时动作值，从而快速、准确地计算出每个实时动作对该学习网络的贡献值。其中，预设第一维动作和预设第二维动作都为预设数量维动作中的一维动作。

在步骤 S104 中，根据当前状态的最大 Q 值获取强化学习网络的当前动作并执行，以得到强化学习网络的下一状态，获取下一状态的最大 Q 值，并通过下一状态的最大 Q 值、当前状态的奖励值和预设目标值公式，获取当前状态的目标 Q 值。

在本发明实施例中，当前动作为当前状态时，强化学习网络需要执行的动作，预设目标值公式具体为 $Target\_Q(s,a;\theta)=r(s)+\gamma\max Q(s',a';\theta)$，其中，$Target\_Q(s,a;\theta)$ 为当前状态的目标 Q 值，s 为当前状态，a 为当前动作，$r(s)$ 为当前状态的奖励值，γ 为折扣因子，θ 为网络参数，$\max Q(s',a';\theta)$ 为下一状态的最大 Q 值。具体地，按照贪婪策略，根据当前状态的最大 Q 值获取强化学习网络的当前动作并执行，进入下一状态，此时，重复步骤 S102 和步骤 S103 的方

法，得到下一状态的最大 Q 值，再通过预设目标值公式得到当前状态的目标 Q 值。

优选地，在获取当前状态的目标 Q 值之后，将当前状态、当前动作、当前状态的奖励值和下一状态作为训练样本进行存储，从而加快了后续的收敛过程。

优选地，强化学习网络训练设备包含 2 个处理器，其中一个芯片为 AI 芯片，该 AI 芯片的架构介于 ASIC（Application Specific Integrated Circuit，专用集成电路）和 FPGA（Field－Programmable Gate Array，现场可编程逻辑门阵列）之间，用于处理强化学习网络训练过程中根据当前状态决策、响应当前动作的部分过程，从而通过提高内存的访问带宽提高强化学习网络的训练速度。

在步骤 S105 中，根据当前状态的目标 Q 值生成强化学习网络的损失函数，通过预设调整算法调整网络参数，以继续对学习网络进行训练，直到损失函数收敛。

在本发明实施例中，得到当前状态的目标 Q 值后，生成强化学习网络的损失函数，具体的，该损失函数为 $L(\theta)=E\left[(Target\_Q(s,a;\theta)-Q(s,a;\theta))^2\right]$，其中，$Target\_Q(s,a;\theta)$ 为当前状态的目标 Q 值，E 为均方差，$Q(s,a;\theta)$ 为实时 Q 值，s 为当前状态，a 为当前动作，θ 为网络参数，然后通过预设调整算法对神经网络参数进行调整，以继续对学习网络进行训练，直到损失函数收敛，从而最终完成强化学习网络的训练。具体地，预设调整算法为 SGD（stochastic gradient descent，随机梯度下降）算法。

在本发明实施例中，当接收到训练强化学习网络的请求时，设置强化学习网络的网络参数，以进行权重配置，获取强化学习网络的当前状态，以及当前状态的奖励值和贡献值，通过遍历动作库的动作组合，获取当前状态下的动作组合的最大 Q 值，根据当前状态的最大 Q 值获取当前动作并执行，通过得到下一状态的最大 Q 值，获取当前状态的目标 Q 值，生成强化学习网络的损失函数，通过预设调整算法调整网络参数，以继续对强化学习网络进行训练，直到损失函数收敛，从而降低了训练强化学习网络的计算量，进而加快了强化学习网络

的训练速度、提高了训练效率。

实施例二：

图 4 示出了本发明实施例二提供的强化学习网络的训练装置的结构，为了便于说明，仅示出了与本发明实施例相关的部分，其中包括：

参数设置单元 41，用于当接收到训练强化学习网络的请求时，设置强化学习网络的网络参数，以对强化学习网络进行权重配置；

匹配获取单元 42，用于获取强化学习网络的当前状态，在预先构建的状态奖励库中对当前状态进行匹配，获取当前状态的奖励值和贡献值；

遍历获取单元 43，用于在预先构建的动作库中遍历动作库的动作组合，获取动作组合的贡献值，并根据当前状态的贡献值和动作组合的贡献值，获取强化学习网络的当前状态的最大 Q 值；

执行获取单元 44，用于根据当前状态的最大 Q 值获取强化学习网络的当前动作并执行，以得到强化学习网络的下一状态，获取下一状态的最大 Q 值，并通过下一状态的最大 Q 值、当前状态的奖励值和预设目标值公式，获取当前状态的目标 Q 值；以及

生成调整单元 45，用于根据当前状态的目标 Q 值生成强化学习网络的损失函数，通过预设调整算法调整网络参数，以继续对学习网络进行训练，直到损失函数收敛。

在本发明实施例中，当接收到训练强化学习网络的请求时，设置强化学习网络的网络参数，以进行权重配置，获取强化学习网络的当前状态，以及当前状态的奖励值和贡献值，通过遍历动作库的动作组合，获取当前状态下的动作组合的最大 Q 值，根据当前状态的最大 Q 值获取当前动作并执行，通过得到下一状态的最大 Q 值，获取当前状态的目标 Q 值，生成强化学习网络的损失函数，通过预设调整算法调整网络参数，以继续对强化学习网络进行训练，直到损失函数收敛，从而降低了训练强化学习网络的计算量，进而加快了强化学习网络的训练速度、提高了训练效率。

在本发明实施例中，强化学习网络的训练装置的各单元可由相应的硬件或软件单元实现，各单元可以为独立的软、硬件单元，也可以集成为一个软、硬件单元，在此不用以限制本发明。各单元的具体实施方式可参考实施例一的描述，在此不再赘述。

实施例三：

图 5 示出了本发明实施例三提供的强化学习网络的训练装置的结构，为了便于说明，仅示出了与本发明实施例相关的部分，其中包括：

参数设置单元 51，用于当接收到训练强化学习网络的请求时，设置强化学习网络的网络参数，以对强化学习网络进行权重配置；

匹配获取单元 52，用于获取强化学习网络的当前状态，在预先构建的状态奖励库中对当前状态进行匹配，获取当前状态的奖励值和贡献值；

遍历获取单元 53，用于在预先构建的动作库中遍历动作库的动作组合，获取动作组合的贡献值，并根据当前状态的贡献值和动作组合的贡献值，获取强化学习网络的当前状态的最大 Q 值；

执行获取单元 54，用于根据当前状态的最大 Q 值获取强化学习网络的当前动作并执行，以得到强化学习网络的下一状态，获取下一状态的最大 Q 值，并通过下一状态的最大 Q 值、当前状态的奖励值和预设目标值公式，获取当前状态的目标 Q 值；

经验存储单元 55，用于将当前状态、当前动作、当前状态的奖励值和下一状态作为训练样本进行存储；以及

生成调整单元 56，用于根据当前状态的目标 Q 值生成强化学习网络的损失函数，通过预设调整算法调整网络参数，以继续对学习网络进行训练，直到损失函数收敛。

其中，匹配获取单元 52 包括：

匹配子单元 521，用于将当前状态与状态奖励库中的预设数量个奖励组对应的所有状态节点进行匹配；以及

状态值设置单元 522，用于当当前状态位于预设数量个奖励组中预设状态节点中时，将预设状态奖励组的奖励值设置为当前状态的奖励值，否则将当前状态的奖励值设置为预设一般状态奖励值。

遍历获取单元 53 包括：

起始值设置单元 531，用于将动作库中预设动作列表上的预设数量维动作的起始值，依次设置为动作库中预设实时动作表上的预设数量个实时动作值；

第一累加单元 532，用于获取预设动作列表上的预设第一维动作的步长值，并将预设第一维动作的步长值逐次累加到预设第一维动作对应的实时动作值；以及

第二累加单元 533，用于当对应的实时动作值逐次累加到预设第一维动作对应的范围之外时，获取预设动作列表上的预设第二维动作的步长值，并将预设第二维动作的步长值逐次累加到预设第二维动作对应的实时动作值。

在本发明实施例中，当接收到训练强化学习网络的请求时，设置强化学习网络的网络参数，以进行权重配置，获取强化学习网络的当前状态，以及当前状态的奖励值和贡献值，通过遍历动作库的动作组合，获取当前状态下的动作组合的最大 Q 值，根据当前状态的最大 Q 值获取当前动作并执行，通过得到下一状态的最大 Q 值，获取当前状态的目标 Q 值，生成强化学习网络的损失函数，通过预设调整算法调整网络参数，以继续对强化学习网络进行训练，直到损失函数收敛，从而降低了训练强化学习网络的计算量，进而加快了强化学习网络的训练速度、提高了训练效率。

在本发明实施例中，强化学习网络的训练装置的各单元可由相应的硬件或软件单元实现，各单元可以为独立的软、硬件单元，也可以集成为一个软、硬件单元，在此不用以限制本发明。各单元的具体实施方式可参考实施例一的描述，在此不再赘述。

实施例四：

图 6 示出了本发明实施例四提供的强化学习网络训练设备的结构，为了便

于说明，仅示出了与本发明实施例相关的部分，其中包括：

本发明实施例的强化学习网络训练设备 6 包括处理器 61、存储器 62 以及存储在存储器 62 中并可在处理器 61 上运行的计算机程序 63。该处理器 51 执行计算机程序 63 时实现上述强化学习网络的训练方法实施例中的步骤，例如图 1 所示的步骤 S101 至 S105。或者，处理器 61 执行计算机程序 63 时实现上述各个强化学习网络的训练装置实施例中各单元的功能，例如图 4 所示单元 41 至 45 以及图 5 所示单元 51 至 56 的功能。

如图 7 所示，强化学习网络训练设备的优选结构示意图。优选地，强化学习网络训练设备 7 包括第一处理器 711、第二处理器 712,、第一存储器 721、第二存储器 722、以及存储在存储器第一存储器 721 和第二存储器 722 中的计算机程序 73，计算机计算程序 73 可在第一处理器 711 和第二处理器 712 上运行。具体地，第一处理器 711 为 ASIC（专用集成电路）芯片，从而提高了该学习网络的效率，并降低功率消耗。第一处理器 711 执行计算机程序 73 时实现上述强化学习网络的训练方法实施例中的步骤，例如图 1 所示的步骤 S101 至 S103，第二处理器 712 执行计算机程序 73 时实现上述强化学习网络的训练方法实施例中的步骤，例如图 1 所示的步骤 S104 至 S105。或者，第一处理器 711 执行计算机程序 73 时实现上述各个强化学习网络的训练装置实施例中各单元的功能，例如图 4 所示单元 41 至 43 以及图 5 所示单元 51 至 53 的功能，第二处理器 712 执行计算机程序 73 时实现上述各个强化学习网络的训练装置实施例中各单元的功能，例如图 4 所示单元 44 至 45 以及图 5 所示单元 54 至 56 的功能。

在本发明实施例中，该处理器执行计算机程序时，当接收到训练强化学习网络的请求时，设置强化学习网络的网络参数，以进行权重配置，获取强化学习网络的当前状态，以及当前状态的奖励值和贡献值，通过遍历动作库的动作组合，获取当前状态下的动作组合的最大 Q 值，根据当前状态的最大 Q 值获取当前动作并执行，通过得到下一状态的最大 Q 值，获取当前状态的目标 Q 值，生成强化学习网络的损失函数，通过预设调整算法调整网络参数，以继续对强

化学习网络进行训练，直到损失函数收敛，从而降低了训练强化学习网络的计算量，进而加快了强化学习网络的训练速度、提高了训练效率。

该处理器执行计算机程序时实现上述强化学习网络的训练方法实施例中的步骤可参考实施例一的描述，在此不再赘述。

实施例五：

在本发明实施例中，提供了一种计算机可读存储介质，该计算机可读存储介质存储有计算机程序，该计算机程序被处理器执行时实现上述强化学习网络的训练方法实施例中的步骤，例如，图 1 所示的步骤 S101 至 S105。或者，该计算机程序被处理器执行时实现上述各个强化学习网络的训练装置实施例中各单元的功能，例如图 4 所示单元 41 至 45 以及图 5 所示单元 51 至 56 的功能。

在本发明实施例中，在计算机程序被处理器执行后，当接收到训练强化学习网络的请求时，设置强化学习网络的网络参数，以进行权重配置，获取强化学习网络的当前状态，以及当前状态的奖励值和贡献值，通过遍历动作库的动作组合，获取当前状态下的动作组合的最大 Q 值，根据当前状态的最大 Q 值获取当前动作并执行，通过得到下一状态的最大 Q 值，获取当前状态的目标 Q 值，生成强化学习网络的损失函数，通过预设调整算法调整网络参数，以继续对强化学习网络进行训练，直到损失函数收敛，从而降低了训练强化学习网络的计算量，进而加快了强化学习网络的训练速度、提高了训练效率。

本发明实施例的计算机可读存储介质可以包括能够携带计算机程序代码的任何实体或装置、存储介质，例如，ROM/RAM、磁盘、光盘、闪存等存储器。

以上所述仅为本发明的较佳实施例而已，并不用以限制本发明，凡在本发明的精神和原则之内所作的任何修改、等同替换和改进等，均应包含在本发明的保护范围之内。

权 利 要 求 书

1、一种强化学习网络的训练方法，其特征在于，所述方法包括下述步骤：

当接收到训练强化学习网络的请求时，设置所述强化学习网络的网络参数，以对所述强化学习网络进行权重配置；

获取所述强化学习网络的当前状态，在预先构建的状态奖励库中对所述当前状态进行匹配，获取所述当前状态的奖励值和贡献值；

在预先构建的动作库中遍历所述动作库的动作组合，获取所述动作组合的贡献值，并根据所述当前状态的贡献值和所述动作组合的贡献值，获取所述强化学习网络的当前状态的最大 Q 值；

根据所述当前状态的最大 Q 值获取所述强化学习网络的当前动作并执行，以使所述强化学习网络进入下一状态，获取所述下一状态的最大 Q 值，并通过所述下一状态的最大 Q 值、所述当前状态的奖励值和预设目标值公式，获取所述当前状态的目标 Q 值；

根据所述当前状态的目标 Q 值生成所述强化学习网络的损失函数，通过预设调整算法调整所述网络参数，以继续对所述强化学习网络进行训练，直到所述损失函数收敛。

2、如权利要求 1 所述的方法，其特征在于，在预先构建的状态奖励库中对强化学习网络的当前状态进行匹配的步骤，包括：

将所述当前状态与所述状态奖励库中的预设数量个奖励组对应的所有状态节点进行匹配；

当所述当前状态位于所述预设数量个奖励组中预设状态节点中时，将所述预设状态奖励组的奖励值设置为所述当前状态的奖励值，否则将所述当前状态的奖励值设置为预设一般状态奖励值。

3、如权利要求 1 所述的方法，其特征在于，在预先构建的动作库中遍历所述动作库的动作组合的步骤，包括：

将所述动作库中预设动作列表上的预设数量维动作的起始值，依次设置为

所述动作库中预设实时动作表上的预设数量个实时动作值；

获取所述预设动作列表上的预设第一维动作的步长值，并将所述预设第一维动作的步长值逐次累加到所述预设第一维动作对应的所述实时动作值；

当所述对应的所述实时动作值逐次累加到所述预设第一维动作对应的范围之外时，获取所述预设动作列表上的预设第二维动作的步长值，并将所述预设第二维动作的步长值逐次累加到所述预设第二维动作对应的所述实时动作值。

4、如权利要求 1 所述的方法，其特征在于，获取所述当前状态的目标 Q 值的步骤之后，所述方法还包括：

将所述当前状态、所述当前动作、所述当前状态的奖励值和所述下一状态作为训练样本进行存储。

5、一种强化学习网络的训练装置，其特征在于，所述装置包括：

参数设置单元，用于当接收到训练强化学习网络的请求时，设置所述强化学习网络的网络参数，以对所述强化学习网络进行权重配置；

匹配获取单元，用于获取所述强化学习网络的当前状态，在预先构建的状态奖励库中对所述当前状态进行匹配，获取所述当前状态的奖励值和贡献值；

遍历获取单元，用于在预先构建的动作库中遍历所述动作库的动作组合，获取所述动作组合的贡献值，并根据所述当前状态的贡献值和所述动作组合的贡献值，获取所述强化学习网络的当前状态的最大 Q 值；

执行获取单元，用于根据所述当前状态的最大 Q 值获取所述强化学习网络的当前动作并执行，以使所述强化学习网络进入下一状态，获取所述下一状态的最大 Q 值，并通过所述下一状态的最大 Q 值、所述当前状态的奖励值和预设目标值公式，获取所述当前状态的目标 Q 值；以及

生成调整单元，用于根据所述强化学习网络的目标 Q 值生成所述强化学习网络的损失函数，通过预设调整算法调整所述强化学习网络的网络参数，以继续对所述强化学习网络进行训练，直到所述损失函数收敛。

6、如权利要求 5 所述的装置，其特征在于，所述匹配获取单元包括：

匹配子单元，用于将所述当前状态与所述状态奖励库中的预设数量个奖励组对应的所有状态节点进行匹配；以及

状态值设置单元，用于当所述当前状态位于所述预设数量个奖励组中预设状态节点中时，将所述预设状态奖励组的奖励值设置为所述当前状态的奖励值，否则将所述当前状态的奖励值设置为预设一般状态奖励值。

7、如权利要求 5 所述的装置，其特征在于，所述遍历获取单元包括：

起始值设置单元，用于将所述动作库中预设动作列表上的预设数量维动作的起始值，依次设置为所述动作库中预设实时动作表上的预设数量个实时动作值；

第一累加单元，用于获取所述预设动作列表上的预设第一维动作的步长值，并将所述预设第一维动作的步长值逐次累加到所述预设第一维动作对应的所述实时动作值；以及

第二累加单元，用于当所述对应的所述实时动作值逐次累加到所述预设第一维动作对应的范围之外时，获取所述预设动作列表上的预设第二维动作的步长值，并将所述预设第二维动作的步长值逐次累加到所述预设第二维动作对应的所述实时动作值。

8、如权利要求 5 所述的装置，其特征在于，所述装置还包括：

经验存储单元，用于将所述当前状态、所述当前动作、所述当前状态的奖励值和所述下一状态作为训练样本进行存储。

9、一种强化学习网络训练设备，包括存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机程序，其特征在于，所述处理器执行所述计算机程序时实现如权利要求 1 至 4 项所述方法的步骤。

10、一种计算机可读存储介质，所述计算机可读存储介质存储有计算机程序，其特征在于，所述计算机程序被处理器执行时实现如权利要求 1 至 4 项所述方法的步骤。

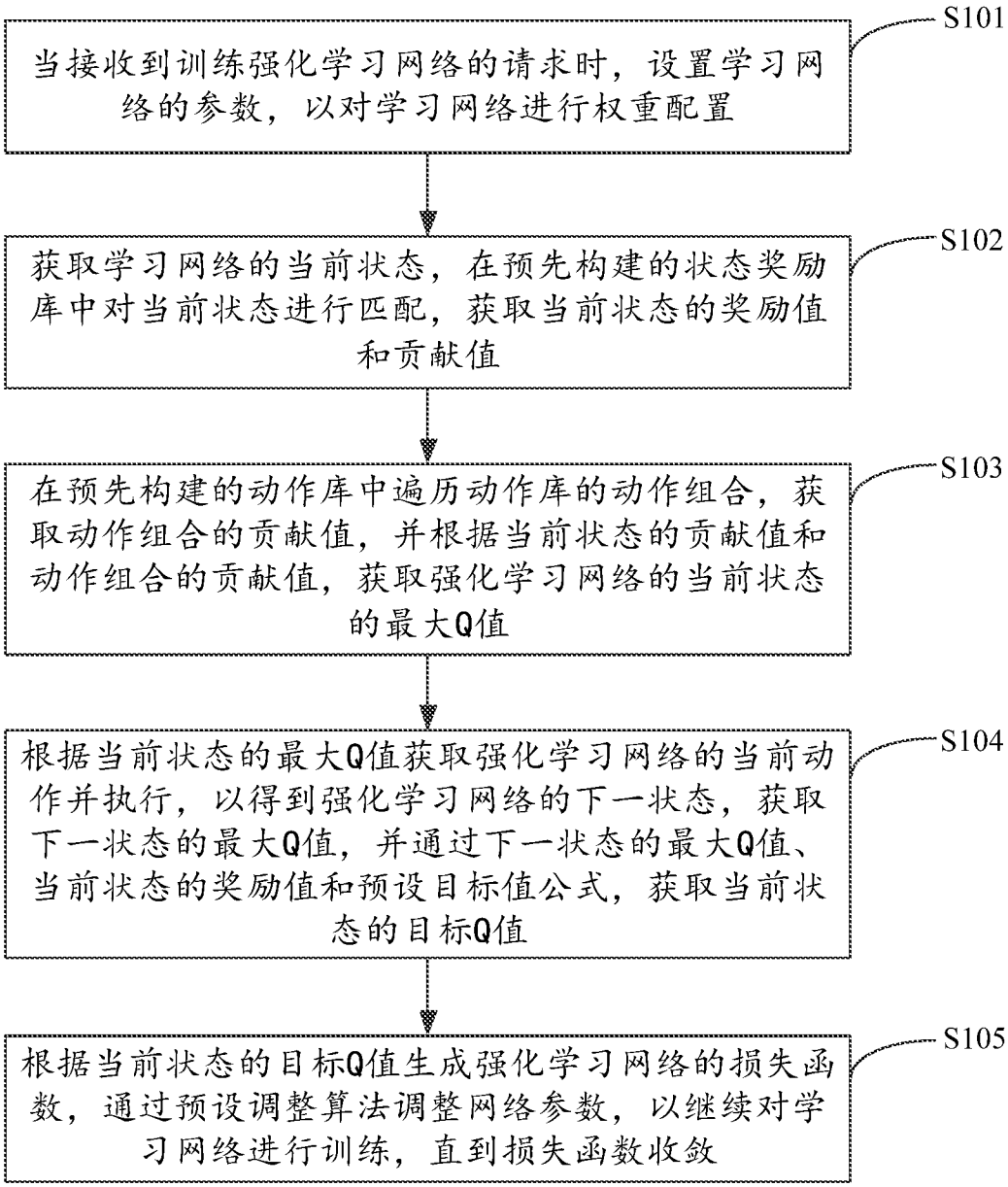
S101
当接收到训练强化学习网络的请求时，设置学习网络的参数，以对学习网络进行权重配置
S102
获取学习网络的当前状态，在预先构建的状态奖励库中对当前状态进行匹配，获取当前状态的奖励值和贡献值
S103
在预先构建的动作库中遍历动作库的动作组合，获取动作组合的贡献值，并根据当前状态的贡献值和动作组合的贡献值，获取强化学习网络的当前状态的最大Q值
S104
根据当前状态的最大Q值获取强化学习网络的当前动作并执行，以得到强化学习网络的下一状态，获取下一状态的最大Q值，并通过下一状态的最大Q值、当前状态的奖励值和预设目标值公式，获取当前状态的目标Q值
S105
根据当前状态的目标Q值生成强化学习网络的损失函数，通过预设调整算法调整网络参数，以继续对学习网络进行训练，直到损失函数收敛

图 1

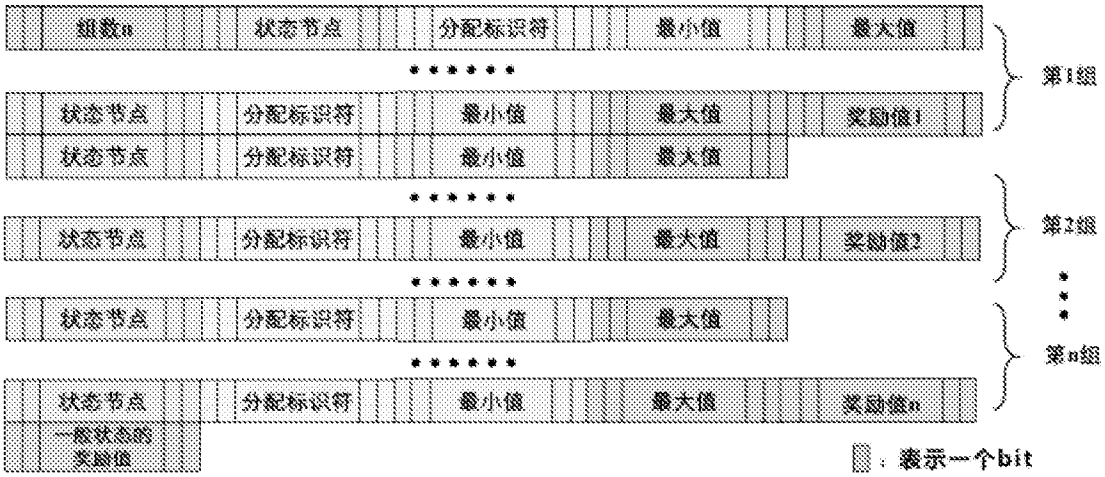
组数n
状态节点
分配标识符
最小值
最大值
第1组
状态节点
分配标识符
最小值
最大值
奖励值1
状态节点
分配标识符
最小值
最大值
第2组
状态节点
分配标识符
最小值
最大值
奖励值2
状态节点
分配标识符
最小值
最大值
第n组
状态节点
分配标识符
最小值
最大值
奖励值n
一般状态的奖励值
：表示一个bit

图 2

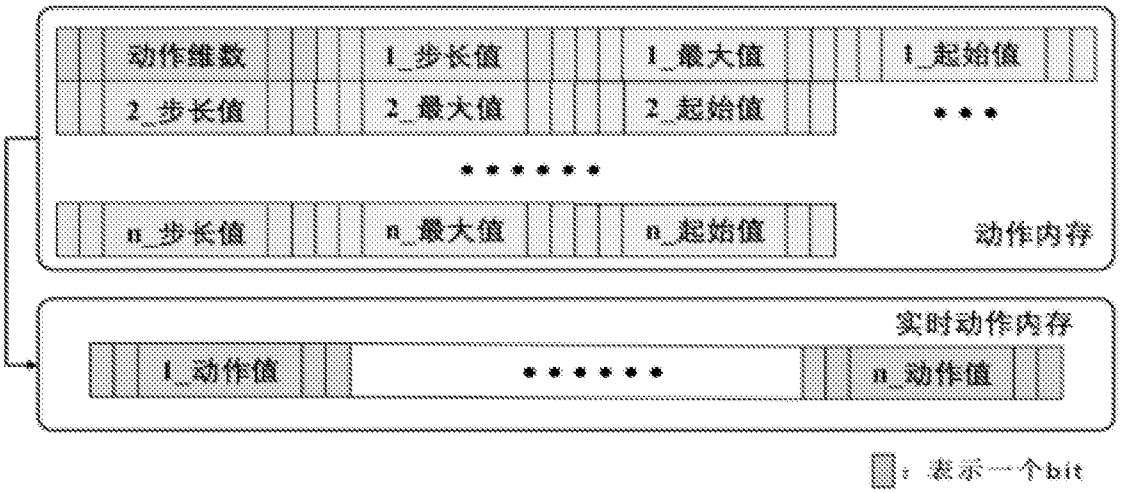
动作维数
1_步长值
1_最大值
1_起始值
2_步长值
2_最大值
2_起始值
n_步长值
n_最大值
n_起始值
动作内存
实时动作内存
1_动作值
n_动作值
：表示一个bit

图 3

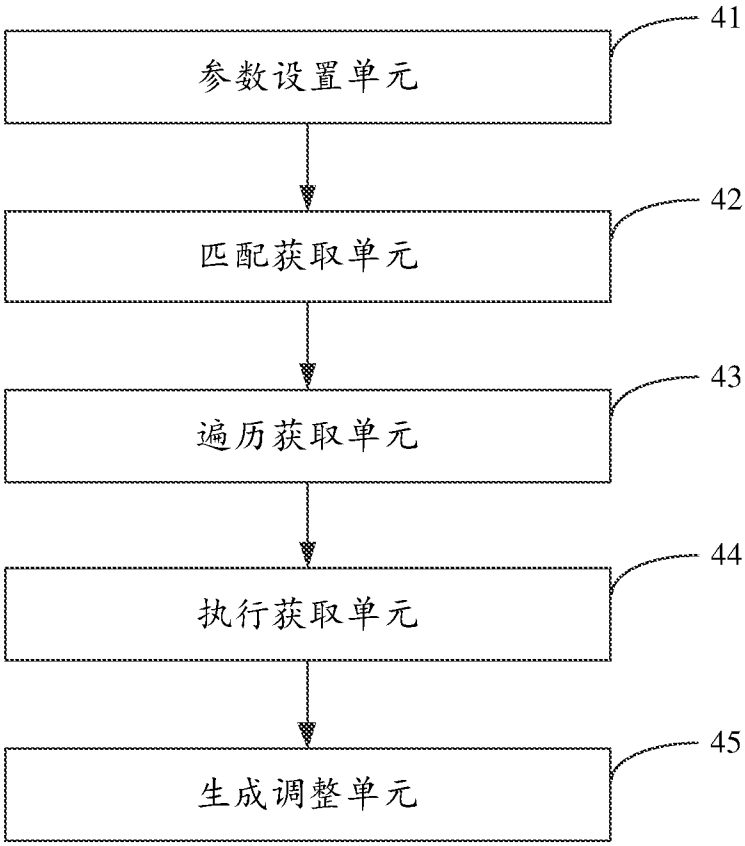

图 4

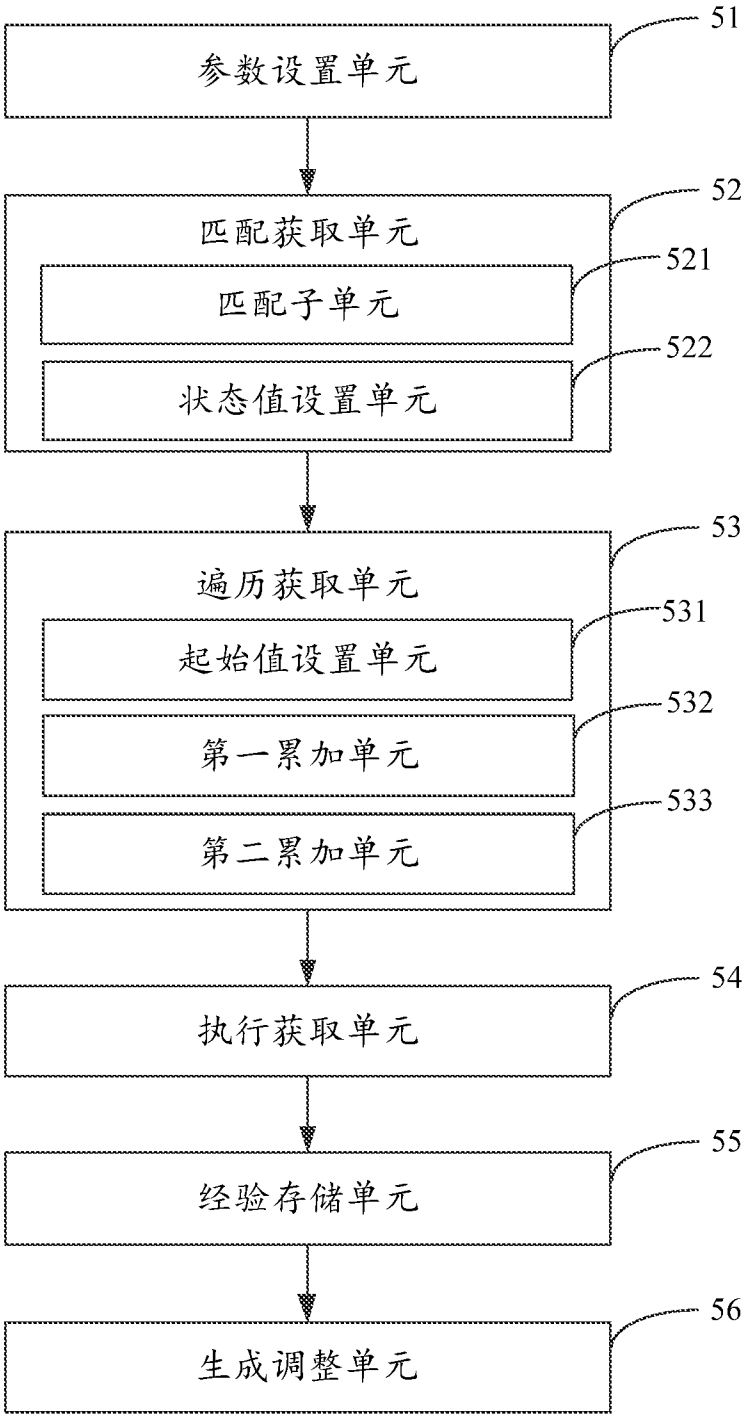
51
参数设置单元
52
匹配获取单元
521
匹配子单元
522
状态值设置单元
53
遍历获取单元
531
起始值设置单元
532
第一累加单元
533
第二累加单元
54
执行获取单元
55
经验存储单元
56
生成调整单元

图 5

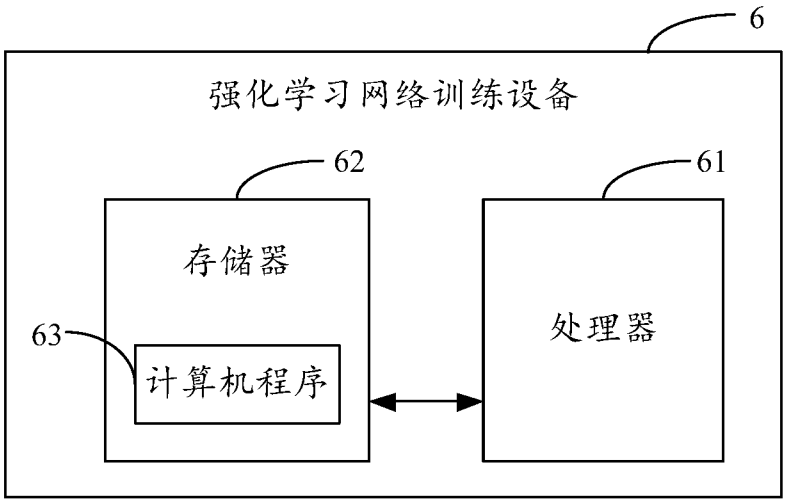
6
强化学习网络训练设备
62
61
存储器
处理器
63
计算机程序

图 6

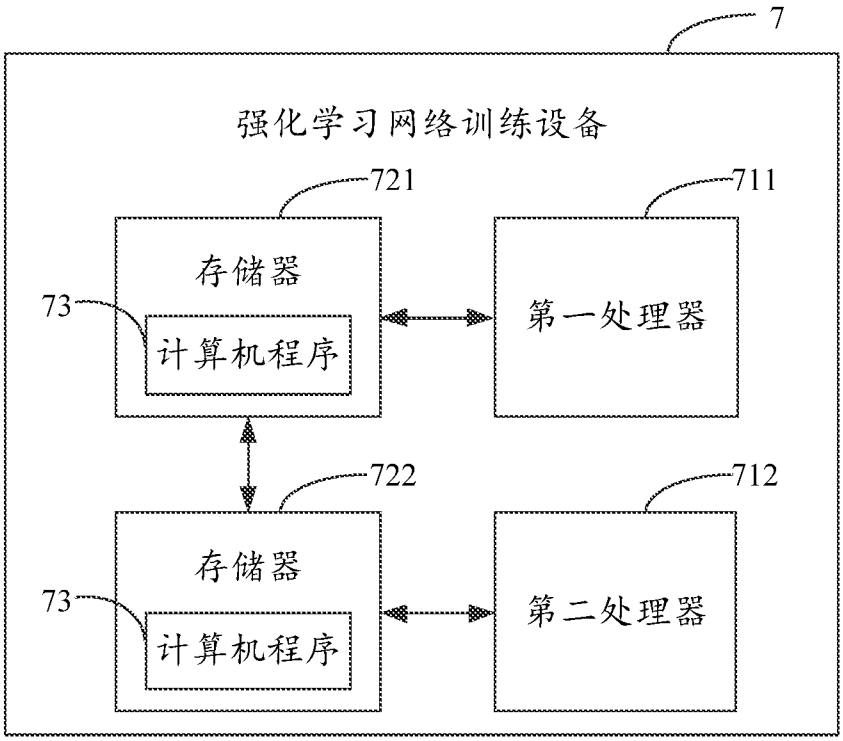
7
强化学习网络训练设备
721
711
存储器
第一处理器
73
计算机程序
722
712
存储器
第二处理器
73
计算机程序

图 7

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2018/099256

A. CLASSIFICATION OF SUBJECT MATTER

G06N 3/08(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06N3/-

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

USTXT; EPTXT; CNTXT; CNABS; WOTXT; VEN; CNKI: 动作, 强化学习, 网络, 训练, 状态, 权重, 最大Q值, 贡献值, 奖励值, 匹配, action, reinforced learning, network, train, state, weight, maximum Q value, contribution, reward value, match

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CN 105637540 A (GOOGLE INC.) 01 June 2016 (2016-06-01) description, paragraphs 11-38, and figures 2 and 5b	1-10
Y	林红等 (LIN, Hong et al.). "基于局部语义的人体动作识别方法 (Local Semantic Concept-based Human Action Recognition)" *信息技术 (Information Technology)*, No. no. 12, 31 December 2015 (2015-12-31), ISSN: 1009-2552, abstract	1-10
A	CN 105955930 A (TIANJIN UNIVERSITY OF SCIENCE & TECHNOLOGY) 21 September 2016 (2016-09-21) entire document	1-10
A	CN 108288094 A (TSINGHUA UNIVERSITY) 17 July 2018 (2018-07-17) entire document	1-10
A	WO 2018053187 A1 (GOOGLE INC.) 22 March 2018 (2018-03-22) entire document	1-10

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
16 April 2019	**09 May 2019**

Name and mailing address of the ISA/CN	Authorized officer
State Intellectual Property Office of the P. R. China **No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088** **China**	
Facsimile No. **(86-10)62019451**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2018/099256

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	105637540	A	01 June 2016	US	9679258	B2	13 June 2017
				US	2015100530	A1	09 April 2015
				US	2017278018	A1	28 September 2017
				WO	2015054264	A1	16 April 2015
				EP	3055813	A1	17 August 2016
				IN	201647015710	A	31 August 2016
CN	105955930	A	21 September 2016	None			
CN	108288094	A	17 July 2018	None			
WO	2018053187	A1	22 March 2018	DE	202017105598	U1	24 May 2018

国际检索报告

国际申请号
PCT/CN2018/099256

A. 主题的分类

G06N 3/08(2006.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06N3/-

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用））

USTXT;EPTXT;CNTXT;CNABS;WOTXT;VEN;CNKI:动作，强化学习，网络，训练，状态，权重，最大Q值，贡献值，奖励值，匹配，action, reinforced learning, network, train, state, weight, maximum Q value, contribution, reward value, match

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
Y	CN 105637540 A (谷歌公司) 2016年 6月 1日 (2016-06-01) 说明书第11-38段及图2、5b	1-10
Y	林红等. "基于局部语义的人体动作识别方法" 信息技术, 第12期, 2015年 12月 31日 (2015-12-31), ISSN：1009-2552, 摘要	1-10
A	CN 105955930 A (天津科技大学) 2016年 9月 21日 (2016-09-21) 全文	1-10
A	CN 108288094 A (清华大学) 2018年 7月 17日 (2018-07-17) 全文	1-10
A	WO 2018053187 A1 (GOOGLE INC.) 2018年 3月 22日 (2018-03-22) 全文	1-10

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2019年 4月 16日	2019年 5月 9日
ISA/CN的名称和邮寄地址 中国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451	受权官员 夏雪 电话号码 86-(20)-28950718

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2018/099256

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
CN	105637540	A	2016年 6月 1日	US	9679258	B2	2017年 6月 13日
				US	2015100530	A1	2015年 4月 9日
				US	2017278018	A1	2017年 9月 28日
				WO	2015054264	A1	2015年 4月 16日
				EP	3055813	A1	2016年 8月 17日
				IN	201647015710	A	2016年 8月 31日
CN	105955930	A	2016年 9月 21日		无		
CN	108288094	A	2018年 7月 17日		无		
WO	2018053187	A1	2018年 3月 22日	DE	202017105598	U1	2018年 5月 24日