



(12)发明专利申请

(10)申请公布号 CN 110383298 A

(43)申请公布日 2019.10.25

(21)申请号 201880014744.5

N.M.O. 希斯

(22)申请日 2018.01.31

(74)专利代理机构 北京市柳沈律师事务所

11105

(30)优先权数据

代理人 金玉洁

62/452,930 2017.01.31 US

(85)PCT国际申请进入国家阶段日

(51)Int.Cl.

2019.08.29

G06N 3/04(2006.01)

(86)PCT国际申请的申请数据

G06N 3/08(2006.01)

PCT/IB2018/000051 2018.01.31

(87)PCT国际申请的公布数据

W02018/142212 EN 2018.08.09

(71)申请人 渊慧科技有限公司

地址 英国伦敦

(72)发明人 M.里德米勒 R.哈夫纳

M.维切里克 T.P.利利克拉普

T.兰珀 I.波波夫 G.巴斯-马龙

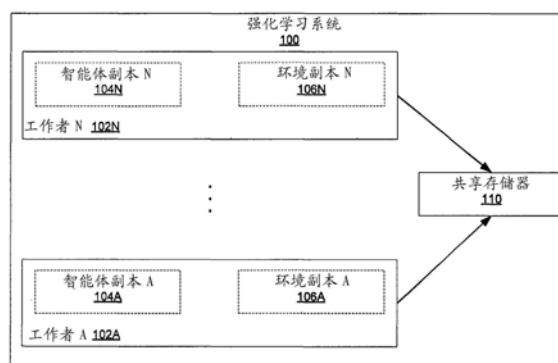
权利要求书3页 说明书12页 附图5页

(54)发明名称

用于连续控制任务的数据高效强化学习

(57)摘要

用于数据高效的强化学习的方法、系统和装置,包括编码在计算机存储介质上的计算机程序。所述系统之一是一种用于训练行动者神经网络的系统,所述行动者神经网络用于选择要由通过接收表征环境的状态的观察并且响应于每个观察执行从可能的动作的连续空间选择的动作而与所述环境交互的智能体执行的动作,其中,所述行动者神经网络按照所述行动者神经网络的参数的值将观察映射到下一些动作,并且其中,所述系统包括:多个工作者,其中,每个工作者被配置为独立于每个其他工作者进行操作,其中,每个工作者与在训练所述行动者神经网络期间与所述环境的相应副本交互的相应智能体副本关联。



1. 一种用于训练行动者神经网络的系统,所述行动者神经网络用于选择要由通过接收表征环境的状态的观察并且响应于每个观察执行从可能的动作的连续空间选择的动作而与所述环境交互的智能体执行的动作,其中,所述行动者神经网络按照所述行动者神经网络的参数的值将观察映射到下一一些动作,并且其中,所述系统包括:

多个工作者,

其中,每个工作者被配置为独立于每个其他工作者进行操作,

其中,每个工作者与在训练所述行动者神经网络期间和所述环境的相应副本交互的相应智能体副本关联,并且,

其中,每个工作者还被配置为重复地执行操作,所述操作包括:

从能够由所述多个工作者中的每一个访问的共享存储器确定所述行动者神经网络和评价神经网络的参数的当前值;

从回放存储器获得经验元组的小批次,每个经验元组包括表征所述环境的训练状态的训练观察、来自响应于所述训练观察由所述智能体副本之一执行的动作的连续空间的训练动作、由用于执行所述训练动作的智能体副本接收的训练反馈值和表征所述环境的下一训练状态的下一训练观察;

确定所述行动者神经网络的参数的当前值和所述评价神经网络的参数的当前值的更新,包括对于所述小批次中的每个经验元组:

使用所述评价神经网络处理所述经验元组中的所述训练观察和所述训练动作以按照所述评价神经网络的参数的当前值确定对于所述经验元组的神经网络输出,

从所述经验元组中的所述训练反馈值和所述经验元组中的所述下一训练观察确定对于所述经验元组的目标神经网络输出,

使用对于所述小批次中的经验元组的所述神经网络输出和目标神经网络输出之间的误差确定所述评价神经网络的参数的当前值的更新,以及

使用所述评价神经网络确定所述行动者神经网络的参数的当前值的更新;以及

将所述行动者神经网络的参数的当前值的更新和所述评价神经网络的参数的当前值的更新写入到所述共享存储器。

2. 根据权利要求1所述的系统,所述操作还包括:

接收表征由与所述工作者关联的所述智能体副本交互的所述环境副本的当前状态的当前观察;

使用行动者神经网络并且按照所述参数的当前值选择响应于所述当前观察要由与所述工作者关联的所述智能体副本执行的当前动作;

识别由于所述智能体副本响应于所述当前观察执行所述当前动作而产生的实际反馈值;

接收表征由所述行动者交互的所述环境副本的下一状态的下一观察,其中,所述环境副本响应于所述智能体副本执行所述当前动作从所述当前状态转换到所述下一状态;

生成包括所述当前观察、所述当前动作、所述实际反馈值和所述下一观察的新经验元组;以及

将所述新经验元组存储在所述回放存储器中。

3. 根据权利要求2所述的系统,所述操作还包括:

在选择响应于所述下一观察要执行的新动作之前执行权利要求1所述的确定、获得、确定和写入的多个迭代。

4. 根据权利要求3所述的系统,其中,在选择所述新动作之前执行的迭代的数量是大于1的预定数量。

5. 根据权利要求3或4中任一项所述的系统,其中,选择响应于所述下一观察要执行的新动作包括:

从所述共享存储器确定所述行动者神经网络的参数的新值;

按照所述参数的新值使用所述行动者神经网络处理所述新观察以将所述新观察映射到新的下一动作;以及

基于所述新的下一动作选择要执行的新动作。

6. 根据权利要求1-5中任一项所述的系统,其中,确定对于所述经验元组的目标神经网络输出包括:

使用目标行动者神经网络处理所述下一训练观察以按照所述目标行动者神经网络的参数的当前值确定对于所述经验元组的预测的下一动作,其中,所述目标行动者神经网络与所述行动者神经网络相同,但是所述目标行动者神经网络的参数的当前值与所述行动者神经网络的参数的当前值不同;

使用目标评价神经网络处理所述下一训练观察和对于所述经验元组的预测的下一动作以按照所述目标评价神经网络的参数的当前值生成预测的下一神经网络输出,其中,所述目标评价神经网络与所述评价神经网络相同,但是所述目标评价神经网络的参数的当前值与所述评价神经网络的参数的当前值不同;以及

从对于所述经验元组的所述训练反馈值和预测的下一神经网络输出确定对于所述经验元组的目标神经网络输出。

7. 根据权利要求6所述的系统,所述操作还包括:

使用所述行动者神经网络的参数的经更新值来更新所述目标行动者神经网络的参数的当前值;以及

使用所述评价神经网络的参数的经更新值来更新所述目标评价神经网络的参数的当前值。

8. 根据权利要求7所述的系统,其中,在训练所述行动者神经网络期间,所述目标行动者神经网络和所述目标评价神经网络的参数的当前值被约束为缓慢改变。

9. 根据权利要求7所述的系统,所述操作还包括:

确定自从所述目标行动者神经网络和所述目标评价神经网络的参数的值的在前更新以来已经发生到所述共享存储器的阈值数量的写入;以及

作为响应,更新所述目标行动者神经网络的参数的当前值和所述目标评价神经网络的参数的当前值。

10. 根据权利要求1-9中任一项所述的系统,其中,使用所述评价神经网络确定所述行动者神经网络的参数的当前值的更新包括:

按照所述行动者神经网络的参数的当前值使用所述行动者神经网络处理所述经验元组中的所述训练观察以生成对于所述训练观察的下一动作;以及

基于(i)所述评价神经网络关于在所述训练观察-下一动作输入对处采取的下一动作

并且按照所述评价神经网络的参数的当前值的梯度和(ii)所述行动者神经网络关于所述行动者神经网络在所述训练观察处采取的参数并且按照所述行动者神经网络的参数的当前值的梯度来确定对于所述行动者神经网络的参数更新。

11. 根据权利要求1-10中任一项所述的系统,其中,每个工作者在相同计算机上独立于每个其他工作者执行。

12. 根据权利要求1-10中任一项所述的系统,其中,每个工作者在不同计算机上独立于每个其他工作者执行。

13. 当从属于权利要求2时,根据权利要求2-12中任一项所述的系统,其中,所述操作还包括:

确定所述环境转换到的下一状态是已经完成由所述智能体副本正在执行的任务的子任务的状态;以及

通过复合反馈值函数将所述实际反馈值赋予为与所述子任务的完成关联的反馈值。

14. 根据权利要求1-13中任一项所述的系统,其中,所述智能体副本与所述环境副本交互以完成任务,并且其中,所述操作还包括:

识别导致所述智能体副本或另一行动者对所述任务的成功完成的动作和对应状态的轨迹;以及

将所述智能体副本和所述环境的初始状态初始化为匹配轨迹中的状态之一的状态。

15. 根据权利要求14所述的系统,其中,将所述智能体副本和所述环境的初始状态初始化为匹配轨迹中的状态之一的状态包括:

随机选择所述轨迹中的状态之一;以及

将所述智能体副本和所述环境的初始状态初始化为匹配选择的的状态的状态。

16. 根据权利要求1-15中任一项所述的系统,其中,每个工作者与每个其他工作者异步地操作,并且其中,每个工作者与每个其他工作者异步地写入所述共享存储器和从所述共享存储器读取。

17. 一种方法,包括由权利要求1-16中任一项所述的工作者之一执行的操作。

18. 一个或多个计算机可读存储介质,所述一个或多个计算机可读存储介质存储指令,所述指令在由一个或多个计算机运行时使所述一个或多个计算机实现根据权利要求1-16中任一项所述的系统。

19. 一个或多个计算机可读存储介质,所述一个或多个计算机可读存储介质存储指令,所述指令在由一个或多个计算机运行时使所述一个或多个计算机执行由权利要求1-16中任一项所述的工作者之一执行的操作。

用于连续控制任务的数据高效强化学习

技术领域

[0001] 本说明书涉及强化学习。

背景技术

[0002] 在强化学习系统中,智能体(agent)通过响应于接收表征环境的当前状态的观察执行由强化学习系统选择的动作来与环境交互。一些强化学习系统按照神经网络的输出选择响应于接收给定观察要由智能体执行的动作。

[0003] 神经网络是利用非线性单元的一个或多个层来为接收的输入预测输出的机器学习模型。一些神经网络是除了输出层之外还包括一个或多个隐藏层的深度神经网络。每个隐藏层的输出被用作到网络中的下一层(即,下一隐藏层或输出层)的输入。网络的每个层按照相应一组参数的当前值从接收的输入生成输出。

发明内容

[0004] 本说明书总体上描述强化学习系统,其训练用于从连续动作空间选择要由与环境交互的强化学习智能体执行的动作的行动者(actor)神经网络。

[0005] 在一个方面,提供了一种用于训练行动者神经网络的系统,所述行动者神经网络用于选择要由通过接收表征环境的状态的观察并且响应于每个观察执行从可能的动作的连续空间选择的动作而与所述环境交互的智能体执行的动作。所述行动者神经网络按照所述行动者神经网络的参数的值将观察映射到下一些动作。所述系统包括多个工作者。每个工作者可以被配置为独立于每个其他工作者进行操作。每个工作者可以与在训练所述行动者神经网络期间和所述环境的相应副本交互的相应智能体副本关联。每个工作者还可以被配置为重复地执行操作。所述操作可以包括从能够由所述多个工作者中的每一个访问的共享存储器确定所述行动者神经网络和评价神经网络的参数的当前值。所述操作还可以包括:从回放存储器获得经验元组的小批次,每个经验元组包括表征所述环境的训练状态的训练观察、来自响应于所述训练观察由所述智能体副本之一执行的动作的连续空间的训练动作、由用于执行所述训练动作的智能体副本接收的训练反馈值(reward)和表征所述环境的下一训练状态的下一训练观察。所述操作还可以包括确定所述行动者神经网络的参数的当前值和所述评价神经网络的参数的当前值的更新。确定所述更新可以包括对于所述小批次中的每个经验元组:使用所述评价神经网络处理所述经验元组中的所述训练观察和所述训练动作以按照所述评价神经网络的参数的当前值确定对于所述经验元组的神经网络输出,从所述经验元组中的所述训练反馈值和所述经验元组中的所述下一训练观察确定对于所述经验元组的目标神经网络输出,使用对于所述小批次中的经验元组的神经网络输出和目标神经网络输出之间的误差确定所述评价神经网络的参数的当前值的更新,以及使用所述评价神经网络确定所述行动者神经网络的参数的当前值的更新。所述操作还可以包括将所述行动者神经网络的参数的当前值的更新和所述评价神经网络的参数的当前值的更新写入到所述共享存储器。

[0006] 在一些实施方式中,神经网络参数的更新相对独立于与环境的交互,从而提高数据使用的效率。在一些实施方式中,使用多个工作者允许数据收集和网络训练分散在多个计算机和智能体上。

[0007] 一般地,智能体可以是软件智能体或现实世界智能体。如之后所描述,智能体可以是机械智能体,诸如机器人,在这种情况下智能体副本可以是与副本环境交互的副本机器人;并且/或者智能体副本可以是与模拟环境交互的软件智能体。

[0008] 在一些实施方式中,智能体副本可以通过与其相应的环境副本交互来为回放存储器生成新经验元组。例如,这可以包括接收表征由与工作者关联的智能体副本交互的环境副本的当前状态的当前观察。该过程还可以包括使用行动者神经网络并且按照参数的当前值选择响应于当前观察要由与工作者关联的智能体副本执行的当前动作。该过程然后还可以包括识别由于智能体副本响应于当前观察执行当前动作所产生的实际反馈值。该过程还可以包括接收表征所述行动者交互的环境副本的下一状态的下一观察,其中,所述环境副本响应于所述智能体副本执行当前动作而从当前状态转换到下一状态。该过程还可以包括生成包括当前观察、当前动作、实际反馈值和下一反馈值的新经验元组并且将新经验元组存储在回放存储器中。

[0009] 所述系统可以在选择响应于下一观察要执行的新动作之前执行先前描述的确定当前参数值、获得小批次以及确定和写入参数值更新的操作的多个迭代。在选择所述新动作之前执行的迭代的数量可以是大于1的预定数量。选择响应于所述下一观察要执行的新动作可以包括:从所述共享存储器确定所述行动者神经网络的参数的新值;按照所述参数的新值使用所述行动者神经网络处理所述新观察以将所述新观察映射到新的下一动作;以及基于所述新的下一动作选择要执行的新动作。

[0010] 在一些实施方式中,确定对于经验元组的目标神经网络输出可以利用目标行动者神经网络和目标评价神经网络以增加稳定性。因此,确定对于经验元组的目标神经网络输出可以包括:使用目标行动者神经网络处理所述下一训练观察以按照所述目标行动者神经网络的参数的当前值确定对于所述经验元组的预测的下一动作,其中,所述目标行动者神经网络与所述行动者神经网络相同,但是所述目标行动者神经网络的参数的当前值与所述行动者神经网络的参数的当前值不同;使用目标评价神经网络处理所述下一训练观察和对于所述经验元组的预测的下一动作以按照所述目标评价神经网络的参数的当前值生成预测的下一神经网络输出,其中,所述目标评价神经网络与所述评价神经网络相同,但是所述目标评价神经网络的参数的当前值与所述评价神经网络的参数的当前值不同;以及从对于所述经验元组的所述训练反馈值和预测的下一神经网络输出确定对于所述经验元组的目标神经网络输出。所述操作然后还可以包括使用所述行动者神经网络的参数的经更新值来更新所述目标行动者神经网络的参数的当前值;以及使用所述评价神经网络的参数的经更新值来更新所述目标评价神经网络的参数的当前值。在一些实施方式中,目标行动者神经网络和目标评价神经网络的参数的当前值可以被约束为在训练行动者神经网络期间仅缓慢改变,例如,被更新的频率低于行动者和评价神经网络。例如,系统可以被配置为确定自从所述目标行动者神经网络和所述目标评价神经网络的参数的值的在前更新以来已经发生到所述共享存储器的阈值数量的写入,作为响应,更新所述目标行动者神经网络和所述目标评价神经网络的参数的当前值。

[0011] 在一些实施方式中,使用评价神经网络确定行动者神经网络的参数的当前值的更新可以利用深度确定性策略梯度(DDPG)强化学习程序。在所描述的系统的背景中,这可以包括按照所述行动者神经网络的参数的当前值使用所述行动者神经网络处理所述经验元组中的所述训练观察以生成对于所述训练观察的下一动作;以及基于(i)所述评价神经网络关于在所述训练观察-下一动作输入对处采取的下一动作并且按照所述评价神经网络的参数的当前值的梯度和(ii)所述行动者神经网络关于所述行动者神经网络在所述训练观察处采取的参数并且按照所述行动者神经网络的参数的当前值的梯度(例如取决于上述(i)和(ii)的乘积)来确定对于所述行动者神经网络的参数更新。

[0012] 在系统的一些实施方式中,每个工作者可以在相同计算机或在不同计算机上独立于每个其他工作者执行。

[0013] 在一些实施方式中,系统可以被配置为确定环境转换到的下一状态是已经完成由智能体副本正在执行的任务的子任务的状态。系统然后可以通过复合反馈值函数将实际反馈值赋予为与子任务的完成关联的反馈值。这可以有助于系统可靠地找到复杂问题的解并且可以减少需要的数据量/训练时间。

[0014] 如前所述,智能体副本可以与相应环境副本交互以完成任务。系统可以被配置为识别导致智能体副本或另一行动者对任务的成功完成的动作和对应状态的轨迹或序列。系统然后将所述智能体副本和所述环境的初始状态初始化为匹配轨迹中的状态之一的状态。以这种方式初始化智能体可以促进快速学习。可以例如通过先前训练的策略或通过人类演示者来提供成功轨迹。将所述智能体副本和所述环境的初始状态初始化为匹配轨迹中的状态之一的状态可以包括:随机选择所述轨迹中的状态之一;以及将所述智能体副本和所述环境的初始状态初始化为匹配选择的的状态的状态。

[0015] 在系统的一些实施方式中,每个工作者与每个其他工作者异步地操作,并且,每个工作者与每个其他工作者异步地写入所述共享存储器和从所述共享存储器读取。

[0016] 本说明书中描述的主题的特定实施例可以被实施以实现下述优点中的一个或多个。

[0017] 强化学习系统可以有效且直接地高维、连续的动作空间中学习智能体的有效动作选择策略,即,通过如本说明所述地训练行动者神经网络。特别地,通过如本说明所述地训练行动者神经网络,即便对于需要精细的动作控制并且当动作空间对于离散化并且然后有效地进行探索来说棘手时,强化学习系统也能够有效地学习有效的动作选择策略。

[0018] 更特别地,通过使用独立且异步地操作的多个工作者以分布式方式训练行动者神经网络,相对于训练这样的行动者神经网络的常规方法可以减少训练行动者神经网络所需的时间。通过对于每次工作者选择动作并且生成经验元组采样多个小批次并且写入多个参数值更新,行动者神经网络的训练更加数据高效,即,与常规方法相比需要智能体与环境的较少的实际交互。因此,行动者神经网络可以被更快地训练同时对各种连续控制任务实现卓越的性能。对于一些任务,行动者神经网络可以被训练为允许智能体完成之前不成功的连续控制任务,即,因为使用常规方式对任务训练行动者神经网络将需要比可行收集的情况更多的数据。

[0019] 本说明书中描述的主题的一个或多个实施例的细节在附图和下面的描述中阐述。根据描述、附图和权利要求书,所述主题的其他特征、方面和优点将变得明了。

附图说明

- [0020] 图1示出示例强化学习系统。
- [0021] 图2是用于训练行动者神经网络的示例过程的流程图。
- [0022] 图3是用于执行环境交互步骤的示例过程的流程图。
- [0023] 图4是用于执行网络训练步骤的示例过程的流程图。
- [0024] 图5是用于使用DDPG技术生成参数值更新的示例过程的流程图。
- [0025] 各附图中同样的参考数字和标记指示同样的元素。

具体实施方式

[0026] 本说明书总体上描述强化学习系统,其训练用于选择要由与环境交互的强化学习智能体执行的动作的行动者神经网络以执行连续控制任务。也就是说,为了与环境交互,智能体接收表征环境的当前状态的数据并且响应于接收的数据执行来自连续动作空间的动作以试图在环境中执行任务。表征环境的状态的数据在本说明书中将被称为观察。

[0027] 在一些实施方式中,环境是模拟环境并且智能体被实现为与模拟环境交互的一个或多个计算机程序。例如,模拟环境可以是视频游戏并且智能体可以是玩视频游戏的模拟用户。作为另一示例,模拟环境可以是机器人或运动模拟环境,例如,驾驶模拟或飞行模拟,并且智能体是通过运动模拟导航以执行某导航任务(例如,在不违反安全约束的情况下导航到环境中特定点)的模拟交通工具。在这些实施方式中,动作可以是控制模拟用户或模拟交通工具的可能控制输入的空间中的点。

[0028] 在一些实施方式中,环境是现实世界环境并且智能体是与现实世界环境交互的机械智能体。例如,智能体可以是与环境交互以完成特定任务(例如,运动到特定位置或以某种可取的方式与环境中的对象交互,例如执行够取和/或抓取和/或放置动作)的机器人。作为另一示例,智能体可以是导航通过环境的自主或半自主交通工具。在这些实施方式中,动作可以是控制机器人或自主交通工具的可能控制输入的空间中的点。

[0029] 在一些情况下,观察使用表征环境的状态的低维特征向量来表征环境的状态。在这些情况下,低维特征向量的不同维度的值可以具有变化的范围。例如,观察可以包括识别机械智能体的关节的当前位置(例如角度)并且可选地运动(例如角速度)的信息。观察还可以包括识别环境中对象的位置、从机械智能体到这些对象的距离或这二者的信息。

[0030] 在一些其他情况下,观察使用来自表征环境的状态的一个或多个图像(例如,模拟环境的图像和/或在机械智能体与现实世界环境交互时由机械智能体的传感器捕获的图像)的高维像素输入表征环境的状态。

[0031] 图1示出示例强化学习系统100。强化学习系统100是被实现为一个或多个位置的一个或多个计算机上的计算机程序的系统的示例,其中可以实现下面描述的系统、组件和技术。

[0032] 强化学习系统100被配置为训练用于选择要由与环境交互的智能体执行的动作的行动者神经网络。

[0033] 特别地,行动者神经网络按照行动者神经网络的参数值(在本说明书中称为行动者网络参数)将观察映射到下一些动作。换言之,行动者神经网络被配置为接收观察作为输入并且处理观察以生成来自连续动作空间的动作作为输出。例如,输出可以是用于机械智

能体的关节的控制输入,例如,设置每个关节的相应速度的向量。特别地,动作空间是连续的而不是离散的,这增加了选择要由智能体执行的动作并且因此训练行动者神经网络的复杂度。

[0034] 特别地,强化学习系统100训练行动者神经网络以与用于训练行动者神经网络的常规方法(即,生成来自连续而不是离散的动作空间的动作作为输出的神经网络)相比数据更高效得多并且计算上更高效得多的方式确定动作选择参数的经训练值。

[0035] 通常,为了训练行动者神经网络,系统重复地执行两个训练步骤:环境交互步骤和神经网络训练步骤。

[0036] 在环境交互步骤中,系统使用行动者神经网络在智能体正在与环境交互时选择要由智能体执行的动作以便生成经验元组。如将会在下面更详细地描述的,经验元组均一般地包括识别被接收的观察、响应于观察被执行的动作和对环境执行动作的效应的数据。环境交互步骤因此可能是时间密集的,因为生成每个经验元组需要智能体执行动作并且获得指定执行动作的效应的数据。

[0037] 在神经网络训练步骤中,系统在已经被生成的经验元组的小批次(mini-batch)上更新行动者网络参数的值。神经网络训练步骤因此可能是计算上密集的,因为它至少需要对于小批次中的每个经验元组确定和反向传播梯度。

[0038] 在训练期间,神经网络训练步骤与交互经验的获取(即,环境交互步骤)交错,并且后者的性质因此受前者的状态影响。

[0039] 更特别地,在交互步骤期间,常规方法使用单个行动者神经网络来选择由与单个版本的环境交互的单个智能体执行的动作。

[0040] 另外,常规方法在每次动作被选择为要由智能体执行之后,在经验元组的小批次上执行单个神经网络训练步骤。然而,这样的方案需要在训练行动者神经网络的过程期间生成大量经验元组,从而使得训练过程相对数据低效,原因是其需要收集大量训练数据以便使训练有效。

[0041] 另一方面,强化学习系统100在每个动作选择之间执行多个神经网络训练步骤。也就是说,在每个动作选择之间,系统100在经验元组的多个小批次上训练行动者神经网络。要在每个动作选择之间执行的训练步骤的数量一般固定为大于二的数量,例如固定为每个动作选择之间五个、十个、二十个或四十个训练步骤。

[0042] 在每个动作选择之间执行多个训练步骤增加训练过程的数据效率。特别地,与使用常规方法相比,需要生成少得多的经验元组来训练行动者神经网络以具有可接受的性能水平。对于一些任务,在每个动作选择之间执行多个训练步骤将训练过程的数据效率增加到这样的程度——智能体能够有效地学习以执行任务,即便是使用常规技术先前无法这样的情况下。

[0043] 然而,由于训练步骤是在计算上密集的,因此在每个动作选择之间执行多个训练步骤可能增加训练过程的计算时间。

[0044] 为了顾及并减轻计算时间的这种增加,强化学习系统100可以包括多个工作者102A-N,每个工作者独立地进行操作以通过更新存储在工作者102A-N中的每一个能够访问的共享存储器110中的行动者网络参数的值来训练行动者神经网络。

[0045] 通常,每个工作者102A-N是能够独立于其他工作者执行该工作者的操作的硬件或

在硬件中实现的软件。在一些实施方式中,在不同计算机上实现工作者102A-N中的每一个。在其他实施方式中,在相同计算机上实现本地工作者102A-N中的两个或更多个,但是这些工作者中的每一个在能够独立于由每个其他工作者执行的计算执行该工作者的计算的计算机内的单独的线程、进行或其他硬件或软件中执行。

[0046] 每个工作者102A-N与在训练行动者神经网络期间与环境的相应副本(replica) 106A-N交互的相应智能体副本104A-N关联,即,与相应智能体副本104A-N通信并向其提供指令。当环境是现实世界环境时,每个工作者102A-N可以与具有相同动力学(dynamics)并与现实世界环境交互以执行相同的连续控制任务的不同机械智能体关联。当环境是模拟环境时,每个工作者102A-N可以与具有相同模拟动力学并与模拟环境的副本交互以执行模拟的连续控制任务的不同模拟智能体关联。

[0047] 在训练期间,每个工作者102A-N对于连续控制任务的许多不同场景片段(episode)异步并且独立地执行与神经网络训练步骤的迭代交错的环境交互步骤的迭代,以重复地将行动者网络参数的当前值的更新写入共享存储器110。下面将参考图2-图5更详细地描述执行这些迭代。

[0048] 一旦已经完成训练,例如,在已经执行阈值数量的训练迭代之后、在阈值时间量已经消逝之后或者在行动者神经网络取得对于任务的令人满意的性能水平之后,强化学习系统100可以使用存储在共享存储器110中的行动者网络参数的值作为行动者网络参数的经训练值。特别地,强化学习系统100可以使用经训练值以选择要由智能体执行的动作、存储经训练值供以后使用或将经训练值提供给另一系统以选择在执行连续控制任务时要由智能体执行的动作时使用。

[0049] 图2是用于训练行动者神经网络的示例过程200的流程图。出于方便,过程200将被描述为由位于一个或多个位置的一个或多个计算机的系统来执行。例如,被适当地编程的强化学习系统中的工作者(例如,图1的强化学习系统100中的工作者之一)可以执行过程200。

[0050] 工作者从共享存储器确定行动者网络参数的当前值(步骤202)。也就是说,工作者访问共享存储器以获得当前存储在存储器中的参数的值。

[0051] 工作者按照行动者网络参数的当前值执行环境交互步骤的迭代以生成经验元组(步骤204)。作为执行环境交互步骤的迭代的一部分,工作者响应于当前观察选择要由智能体副本执行的动作,这导致工作者接收表征环境副本的新状态的新观察。下面参考图3更详细地描述执行环境交互步骤的迭代。

[0052] 在响应于新观察选择要由智能体副本执行的动作之前,工作者执行网络训练步骤的固定数量的迭代(步骤206)。通常,固定数量为大于一的数量,例如五个、十个、二十个或四十个迭代。下面参考图4更详细地描述执行网络训练步骤的迭代。

[0053] 工作者从共享存储器确定行动者网络参数的新值(步骤208)并且按照新值响应于新观察选择要由智能体副本执行的动作(步骤210)。然后,工作者可以基于智能体响应于新观察执行新动作的效应来生成另一经验元组。

[0054] 为了训练行动者神经网络,每个工作者对于连续控制任务的许多不同场景片段并且与其他工作者异步地执行过程200许多不同次数。

[0055] 在一些情况下,每个工作者在环境内的随机状态或在环境内的一组预定状态之一

初始化连续控制任务的每个场景片段。

[0056] 然而,在一些其他情况下,工作者可以利用识别导致智能体副本或另一行动者(例如,另一机械智能体或人)对任务的成功完成的动作和对应状态的一个或多个轨迹的数据。

[0057] 在这些情况下,工作者可以将一些或所有场景片段的初始状态初始化为匹配轨迹中的一个轨迹中的状态之一的状态。例如,工作者可以将来自一个或多个轨迹的状态随机地选择为给定场景片段的初始状态。

[0058] 图3是用于执行环境交互步骤的迭代的示例过程300的流程图。出于方便,过程300将被描述为由位于一个或多个位置的一个或多个计算机的系统来执行。例如,被适当地编程的强化学习系统中的工作者(例如,图1的强化学习系统100中的工作者之一)可以执行过程300。

[0059] 工作者从共享存储器确定行动者网络参数的当前值(步骤302)。

[0060] 工作者接收表征由与工作者关联的智能体副本交互的环境副本的当前状态的当前观察(步骤304)。

[0061] 工作者使用行动者神经网络并且按照行动者网络参数的当前值选择响应于当前观察要由与工作者关联的智能体副本执行的当前动作(步骤306)。特别地,工作者使用行动者神经网络并且按照行动者网络参数的当前值来处理当前观察以将当前观察映射到输出动作。在一些实施方式中,工作者选择输出动作作为要执行的当前动作。在一些其他实施方式中,为了增加探索,工作者将从先前分布采样的噪声添加到输出动作以生成当前动作。

[0062] 工作者识别由于智能体副本响应于当前观察执行当前动作所产生的实际反馈值(reward)(步骤308)。

[0063] 在一些情况下,反馈值是稀疏反馈值,例如作为正值的反馈值,例如,当作为智能体执行当前动作的结果连续控制任务已经被完成时为一,并且当任务未被完成时为零。

[0064] 在其他情况下,为了进一步提高训练行动者神经网络的数据效率,工作者使用复合反馈值函数来生成与常规反馈值相比较不稀疏的反馈值并且因此改善行动者神经网络的训练。

[0065] 特别地,工作者保持识别表征已经完成由智能体副本正在执行的连续控制任务的子任务的状态的观察的数据。

[0066] 在一些情况下,工作者还保持识别与完成每个子任务关联的反馈值的数据。在这些情况下,工作者从下一观察和保持的数据确定环境转换到的下一状态是否是已经完成由智能体副本正在执行的任务的给定子任务的状态,并且如果是,则将实际反馈值赋予为与给定子任务的完成关联的反馈值,即,而不是常规反馈值方案中的零。这样,用指示智能体对任务执行得多么好的附加反馈来增强训练。

[0067] 在一些其他情况下,工作者还保持平滑变化的复合反馈值函数,其在智能体副本执行不同子任务时不同地赋予反馈值。例如,该函数可以赋予等于上面赋予的子任务完成反馈值加上基于从智能体到环境中的与当前子任务关联的对象的距离的反馈值的反馈值。

[0068] 工作者接收表征由智能体副本交互的环境副本的下一状态的下一观察(步骤310)。特别地,下一状态是作为智能体副本执行当前动作的结果环境副本转换到的状态。

[0069] 工作者生成包括当前观察、当前动作、实际反馈值和下一观察的新经验元组(步骤312)并且将新经验元组存储在回放存储器(replay memory)中(步骤314)。

[0070] 在一些情况下,回放存储在所有工作者之间共享并且可以由所有工作者访问。

[0071] 在其他情况下,回放存储器仅特定于该工作者并且每个工作者与唯一回放存储器关联。

[0072] 图4是用于执行网络训练步骤的示例过程400的流程图。出于方便,过程400将被描述为由位于一个或多个位置的一个或多个计算机的系统来执行。例如,被适当地编程的强化学习系统中的工作者(例如,被适当地编程的图1的强化学习系统100中的工作者之一)可以执行过程400。

[0073] 工作者从共享存储器确定行动者网络参数和评价神经网络的参数的当前值(步骤402)。评价神经网络是这样的神经网络,其被配置为按照评价神经网络的参数(在本说明书中称为“评价网络参数”)的当前值处理输入观察和输入动作,以生成作为由于智能体副本响应于输入观察执行输入动作而将产生的返回值的估计(即,将来的实际反馈值的不计时间和(time-discounted sum))的神经网络输出。

[0074] 工作者从回放存储器获得经验元组的小批次,即,从在所有工作者之间共享的存储器或特定于该工作者的存储器采样小批次(步骤404)。每个经验元组包括表征环境的训练状态的训练观察、来自智能体副本之一响应于训练观察而执行的动作的连续空间的训练动作、由智能体副本对于执行训练动作接收的训练反馈值和表征环境的下一训练状态的下一训练观察。

[0075] 工作者使用小批次确定行动者网络参数和评价网络参数的当前值的更新(步骤406)。一般地,工作者通过使用深度确定性策略梯度(DDPG)强化学习技术在小批次上训练行动者神经网络和评价神经网络来确定更新。使用DDPG技术确定这些更新在下面参考图5在高层次进行描述,并且更详细地在Timothy P.Lillicrap、Jonathan J.Hunt、Alexander Pritzel、Nicolas Heess、Tom Erez、Yuval Tassa、David Silver、Daan Wierstra的“CONTINUOUS CONTROL WITH DEEP REINFORCEMENT LEARNING”(arXiv:1509.02971),以及于2016年7月22日提交并且标题为“CONTINUOUS CONTROL WITH DEEP REINFORCEMENT LEARNING”的美国专利申请第15/217,758号中进行描述,上述两篇文件都通过引用整体并入本文。

[0076] 工作者将行动者神经网络的参数的当前值和评价神经网络的参数的当前值的更新写入共享存储器(步骤408)。特别地,工作者使用DDPG技术确定参数的新值并且然后将新值写入共享存储器,即,通过覆盖当前存储在存储器中的值。

[0077] 图5是用于使用DDPG技术确定参数值更新的示例过程500的流程图。出于方便,过程500将被描述为由位于一个或多个位置的一个或多个计算机的系统来执行。例如,被适当地编程的强化学习系统中的工作者(例如,图1的强化学习系统100中的工作者之一)可以执行过程500。

[0078] 对于小批次中的每个经验元组,工作者使用评价神经网络处理经验元组中的训练观察和训练动作以按照评价神经网络的参数的当前值确定对于经验元组的神经网络输出(步骤502)。

[0079] 对于小批次中的每个经验元组,工作者从经验元组中的训练反馈值和经验元组中的下一训练观察确定对于经验元组的目标神经网络输出(步骤504)。

[0080] 特别地,为了确定对于经验元组的目标神经网络输出,工作者使用目标行动者神

神经网络处理下一训练观察,以按照目标行动者神经网络的参数的当前值确定对于经验元组的预测的下一动作。目标行动者神经网络与所述行动者神经网络相同,但是目标行动者神经网络的参数的当前值一般与所述行动者神经网络的参数的当前值不同。

[0081] 工作者然后使用目标评价神经网络处理下一训练观察和对于经验元组的预测的下一动作以按照目标评价神经网络的参数的当前值生成预测的下一神经网络输出。目标评价神经网络与所述评价神经网络相同,但是目标评价神经网络的参数的当前值与所述评价神经网络的参数的当前值不同。

[0082] 工作者然后从对于经验元组的训练反馈值和预测的下一神经网络输出确定对于经验元组的目标神经网络输出,例如,通过计算训练反馈值和预测的下一神经网络输出的不计时间和。

[0083] 工作者使用对于小批次中的经验元组的神经网络输出和目标神经网络输出之间的误差确定评价神经网络的参数的当前值的更新(步骤506)。

[0084] 工作者使用评价神经网络确定行动者神经网络的参数的当前值的更新(步骤508)。

[0085] 一般地,为了更新行动者网络参数的当前值,工作者按照行动者神经网络的参数的当前值使用行动者神经网络处理经验元组中的训练观察以生成对于训练观察的下一动作,并且基于(i)评价神经网络关于在训练观察-下一动作输入对处采取的下一动作并且按照评价神经网络的参数的当前值的梯度以及(ii)行动者神经网络关于行动者神经网络在训练观察处采取的参数并且按照行动者神经网络的参数的当前值的梯度来确定对于行动者神经网络的参数更新。

[0086] 在Timothy P.Lillicrap、Jonathan J.Hunt、Alexander Pritzel、Nicolas Heess、Tom Erez、Yuval Tassa、David Silver、Daan Wierstra的“CONTINUOUS CONTROL WITH DEEP REINFORCEMENT LEARNING”以及于2016年7月22日提交并且标题为“CONTINUOUS CONTROL WITH DEEP REINFORCEMENT LEARNING”的美国专利申请第15/217,758号中更详细地描述了使用DDPG技术来确定这些更新(即,使用目标神经网络输出和对于元组的神经网络输出确定行动者和评价网络参数的新值)。这两篇文件都通过引用整体并入本文。

[0087] 在一些情况下,在训练期间,目标行动者神经网络和目标评价神经网络的参数的当前值被约束为缓慢改变。在这些情况下,工作者可以通过对于目标网络中的每一个确定目标网络参数的当前值与对应的经更新值之间的加权和,使用行动者神经网络的参数的经更新值更新目标行动者神经网络的参数的当前值,并且使用评价神经网络的参数的经更新值更新目标评价神经网络的参数的当前值。

[0088] 在其他情况下,工作者确定自从目标行动者神经网络和目标评价神经网络的参数的值的在前更新以来是否已经发生到共享存储器的阈值数量的写入,并且只有在已经发生阈值数量的写入的情况下才更新目标行动者神经网络的参数的当前值和目标评价神经网络的参数的当前值。

[0089] 本说明书与系统和计算机程序组件有关地使用术语“被配置”。一个或多个计算机的系统被配置为执行特定操作或动作的意思是系统在其上安装有软件、固件、硬件或其组合,在操作时,软件、固件、硬件或其组合使系统执行操作或动作。一个或多个计算机程序被配置为执行特定操作或动作的意思是一个或多个程序包括指令,当在由数据处理装置执行

该指令时使装置执行操作或动作。

[0090] 本说明书中描述的主题和功能操作的实施例可以在数字电子电路、有形地实现的计算机软件或固件、计算机硬件(包括本说明书中公开的结构以及其结构等效物)或它们中的一个或多个的组合中实现。本说明书中描述的主题的实施例可以被实现为一个或多个计算机程序,即编码在有形非暂态存储介质上以供数据处理装置执行或控制数据处理装置的操作的计算机程序指令的一个或多个模块。计算机存储介质可以是机器可读存储设备、机器可读存储基片、随机或串行存取存储器或它们中一个或多个的组合。替代地或额外地,程序指令可以编码在人工生成的传播的信号(例如,机器生成的电、光或电磁信号,其被生成编码信息以传输到合适的接收器装置以供数据处理装置执行)上。

[0091] 术语“数据处理装置”是指数据处理硬件并且涵盖用于处理数据的所有种类的装置、设备和机器,举例来说包括可编程处理器、计算机或多处理器或计算机。所述装置还可以是或者进一步包括专用逻辑电路,例如FPGA(现场可编程门阵列)或ASIC(专用集成电路)。除了硬件之外,所述装置可以可选地包括代码,该代码为计算机程序创建执行环境,例如,构成处理器固件、协议栈、数据库管理系统、操作系统或它们中一种或多种的组的代码。

[0092] 计算机程序(其也被称为或被描述为程序、软件、软件应用、app、模块、软件模块、脚本或代码)可以以任何编程语言编写,包括编译或解释性语言、陈述性或过程语言,并且可以以任何形式部署计算机程序,包括作为独立程序或作为模块、组件、子程序或其他适于用在计算环境中的单元。程序可以但不必与文件系统中的文件对应。程序可以存储在保持其他程序或数据的文件的一部分(例如,存储在标记语言文档中的一个或多个脚本)中、在专用于所讨论程序的单个文件中或在多个协作的文件(例如,存储一个或多个模块、子程序或代码的一些部分的文件)中。计算机程序可以部署为在一个或多个计算机上执行,这些计算机位于一个站点或跨多个站点分布并通过数据通信网络互联。

[0093] 在本说明书中,术语“数据库”被用于宽泛地指任何数据集合:数据不需要以任何特定方式结构化,或者根本不结构化,并且其可以存储在一个或多个位置的存储设备上。因此,例如,索引数据库可以包括多个数据集合,每个数据集合可以被不同地组织和访问。

[0094] 类似地,在本说明书中,术语“引擎”被用于宽泛地指基于软件的系统、子系统或被编程为执行一个或多个特定功能的过程。通常,引擎将被实现为安装在一个或多个位置的一个或多个计算机上的一个或多个软件模块或组件。在一些情况下,一个或多个计算机将专用于特定引擎;在其他情况下,多个引擎可以安装并运行在相同的计算机或多个计算机上。

[0095] 本说明书中描述的过程和逻辑流可以由执行一个或多个计算机程序的一个或多个可编程计算机执行,从而通过对输入数据进行操作并生成输出来执行功能。所述过程和逻辑流程也可以由专用逻辑电路(例如FPGA或ASIC)执行,或者由专用逻辑电路和一个或多个经编程计算机的组合来执行。

[0096] 适用于执行计算机程序的计算机可以基于通用或专用微处理器或二者,或者任何其他种类的中央处理单元。通常,中央处理单元将从只读存储器或随机存取存储器或二者接收指令和数据。计算机的基本元件是用于执行或运行指令的中央处理单元和用于存储指令和数据的一个或多个存储设备。中央处理单元和存储器可以由专用逻辑电路来补充或可

以结合到其中。一般来说,计算机将可以包括用于存储数据的一个或多个大容量存储设备(例如、磁盘、磁光盘或光盘)或可以操作性地联接一个或多个大容量存储设备以从其接收数据或将数据传送到其处或者接收和传输二者。然而,计算机不必具有这样的设备。另外,计算机可以嵌入到另一设备(例如,仅举例几个,移动电话、个人数字助理(PDA)、移动音频或视频播放器、游戏控制台、全球定位系统(GPS)接收器或便携式存储设备(例如通用串行总线(USB)快闪驱动器))中。

[0097] 适用于存储计算机程序指令和数据的计算机可读介质包括所有形式的非易失性存储器、介质和存储设备,举例来说,包括半导体存储设备,例如EPROM、EEPROM和闪存设备;磁盘,例如内部硬盘或可移动盘;磁光盘;以及CD-ROM和DVD-ROM盘。

[0098] 为了提供与用户的交互,本说明书中描述的主题的实施例可以实现在具有显示设备及键盘和指向设备的计算机上,显示设备例如用于向用户显示信息的CRT(阴极射线管)或LCD(液晶显示)监视器,用户可以通过键盘和指向装置向计算机提供输入,指向装置例如为鼠标或轨迹球。也可以使用其他种类的设备来提供与用户的交互;例如,提供给用户的反馈可以是任何形式的感觉反馈,例如视觉反馈、听觉反馈或触觉反馈;并且可以以任何形式接收来自用户的输入,包括声学、语音或触觉输入。此外,通过向由用户使用的设备发送文档和接收来自该设备的文档;例如,通过响应于从web浏览器接收的请求而将网页发送到用户的设备上的web浏览器,计算机可以与用户交互。另外,计算机可以通过将文本消息或其他形式的消息发送到个人设备(例如,运行消息应用的智能电话)并且作为回应从用户接收响应消息来与用户交互。

[0099] 用于实现机器学习模型的数据处理装置还可以包括例如用于处理机器学习训练或产生的一般和计算密集部分(即,推断、工作负荷)的专用硬件加速器单元。

[0100] 机器学习模型可以使用机器学习框架(例如,TensorFlow框架、微软认知工具包框架(Microsoft Cognitive Toolkit framework)、Apache Singa框架或Apache MXNet框架)来实现和部署。

[0101] 本说明书中描述的主题的实施例可以在计算系统中实现,该计算系统包括后端组件(例如作为数据服务器)或包括中间件组件(例如应用服务器)或包括前端组件(例如具有图形用户界面、web浏览器或app的客户端计算机,用户可通过其与本说明书中描述的主题的实施方式进行交互),或者一个或多个这样的后端、中间件或前端组件的任意组合。可以通过数字数据通信的任意形式或媒介(例如通信网络)将系统的组件互连。通信网络的示例包括局域网(LAN)和广域网(WAN),例如互联网。

[0102] 计算系统可以包括客户端和服务端。一般来说,客户端和服务端彼此远离,并且通常通过通信网络交互。通过运行在相应计算机并且彼此具有客户端-服务器关系的计算机程序而形成的客户端和服务器的关系。在一些实施例中,服务器向充当客户端的用户设备传输数据(例如,HTML页面)(例如,用于向与设备交互的用户显示数据和从其接收用户输入的目的)。在用户设备处生成的数据(例如,用户交互的结果)可以在服务器处从该设备接收。

[0103] 尽管本说明书包含许多具体的实现细节,但是这些细节不应当被解释为对任何发明的范围或对可能请求保护的范围进行限制,而应该被解释为可能特定于特定发明的特定实施例的特征的描述。在本说明书中在单独的实施例的上下文中描述的某些特征也可以以

组合实现在单个实施例中。相反,在单个实施例的上下文中描述的各种特征也单独地可以在多个实施例中实施或以任何合适的子组合实施。此外,尽管特征可以如上被描述为以某些组合起作用并且甚至最初也这样地请求,但在一些情况下,来自所请求保护的组合的一个或多个特征可以从组合中排除,并且所请求保护的组合可以涉及子组合或子组合的变体。

[0104] 类似地,尽管在附图中以特定顺序描绘了操作并且在权利要求中以特定顺序记载了操作,但是这不应当被理解为需要以所示出的特定顺序或以相继的顺序来执行这样的操作或者需要执行所有示意的操作来实现期望的结果。在某些情况下,多任务和并行处理可能是有利的。另外,上述实施例中各种系统模块和组件的分离不应当被理解为在所有实施例中要求这样的分离,而是应当理解,所描述的程序组件和系统可一般地在单个软件产品中被集成在一起或者被封装成多个软件产品。

[0105] 已经描述了主题的特定实施例。其他实施例在所附权利要求的范围内。例如,记载在权利要求中的动作可以以不同的顺序执行而仍然实现期望的结果。作为一个示例,附图中描绘的过程不一定需要所示出的特定顺序或者相继顺序来实现期望的结果。在一些情况下,多任务和并行处理可能是有利的。

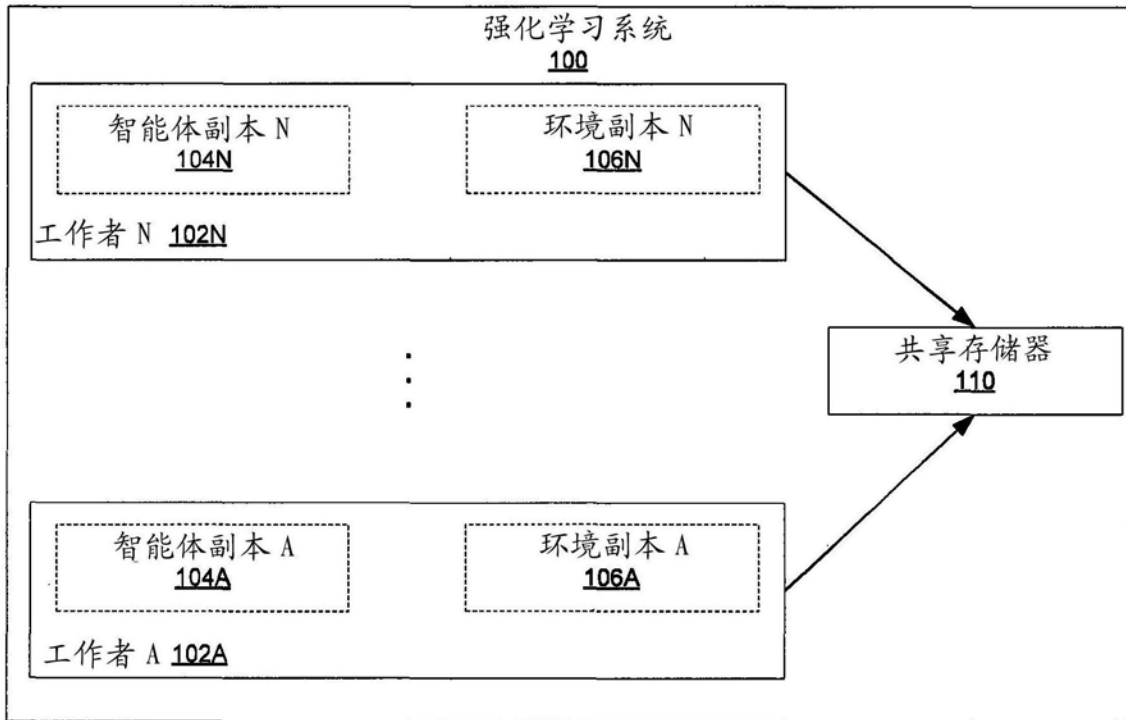


图1

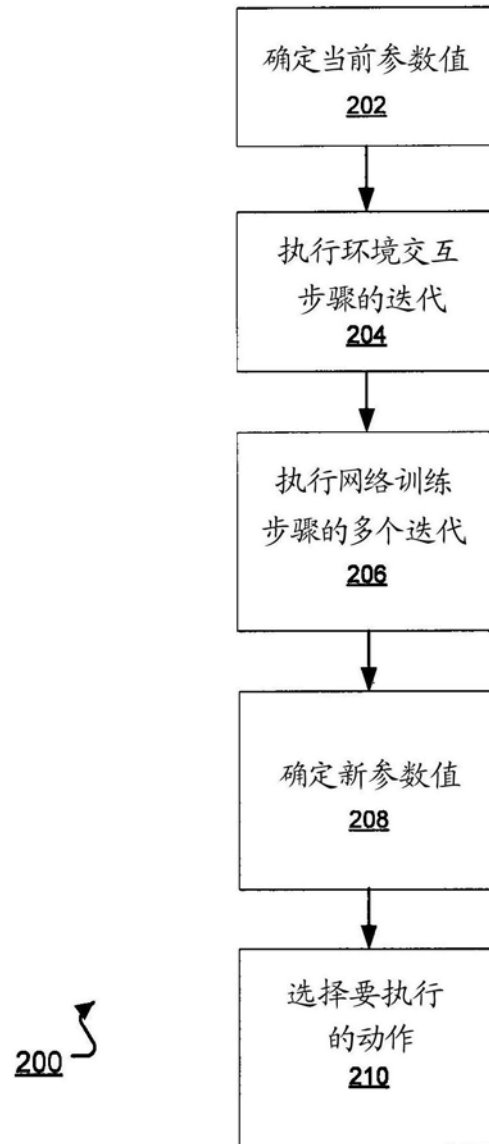


图2

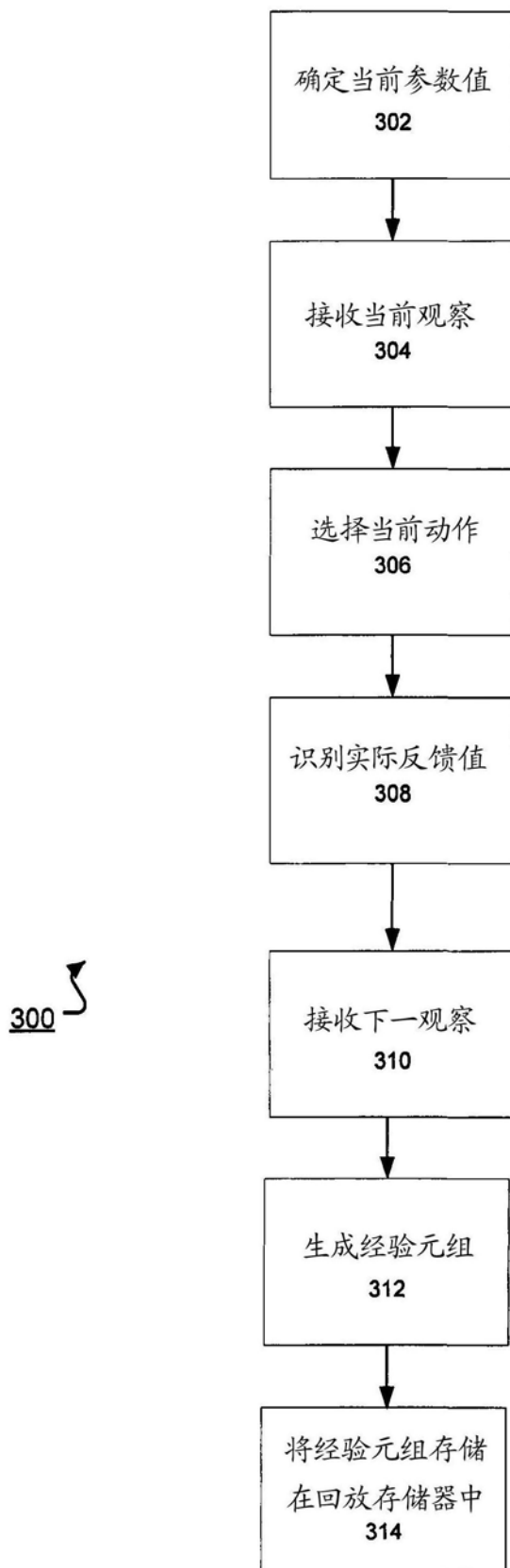


图3

400 ↗



图4

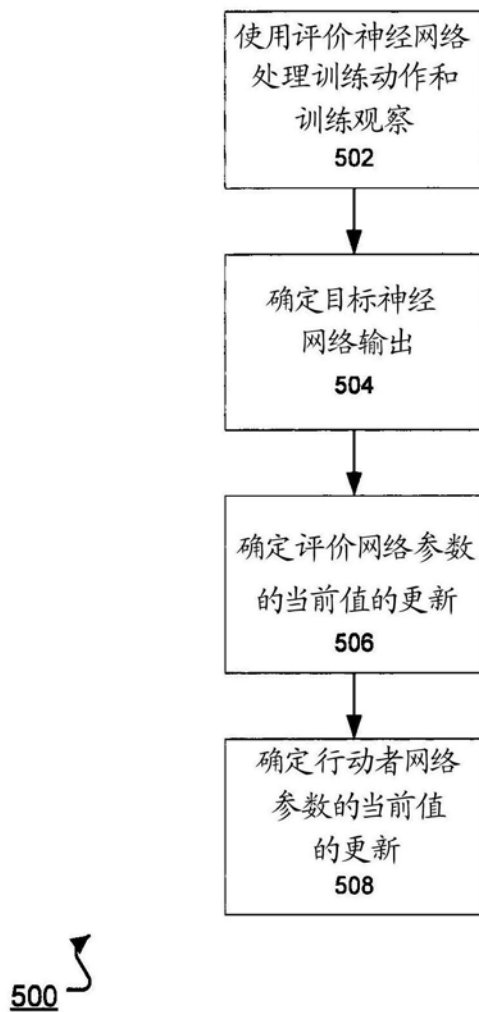


图5