



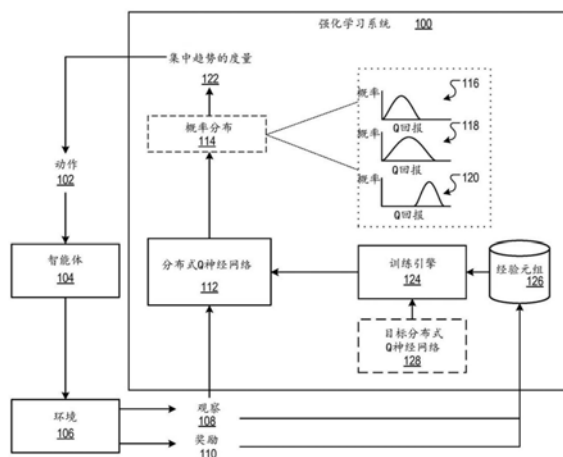
(43)申请公布日 2019.11.29

(72)发明人 W.C.达布尼 M.金德罗-贝尔-梅尔

权利要求书4页 说明书13页 附图4页

分布式强化学习

方法、系统和设备,包括在计算机存储介质上编码的计算机程序,以选择由与环境交互的强化学习智能体要执行的动作。接收表征环境的当前状态的当前观察。对于可以由与环境交互的智能体执行的多个动作的集合中的每个动作,确定针对动作-当前观察对的可能的Q回报之上的概率分布。对于每个动作,确定相对于针对动作-当前观察对的概率分布的可能的Q回报的集中趋势的度量。使用集中趋势的度量,选择响应于当前观察将由智能体执行的动作。



1. 一种选择将由与环境交互的强化学习智能体执行的动作的方法,所述方法包括:
接收表征所述环境的当前状态的当前观察;
对于可由与所述环境交互的所述智能体执行的多个动作中的每个动作,
使用具有多个网络参数的分布式Q网络处理所述动作和所述当前观察,
其中所述分布式Q网络是深度神经网络,其被配置为根据所述网络参数的当前值来处理所述动作和所述当前观察,以生成网络输出,所述网络输出限定针对所述动作-当前观察对的可能的Q回报之上的概率分布,和

其中每个可能的Q回报是从响应于所述当前观察执行所述动作的所述智能体得到的回报的估计,并且

相对于针对所述动作-当前观察对的概率分布,确定所述可能的Q回报的集中趋势的度量;以及

使用所述动作的所述集中趋势的度量,从响应于所述当前观察将由所述智能体执行的多个可能动作中选择动作。

2. 根据权利要求1所述的方法,其中选择将由所述智能体执行的动作包括:

选择具有最高集中趋势的度量的动作。

3. 根据权利要求1所述的方法,其中选择将由所述智能体执行的动作包括:

选择概率为 $1-\epsilon$ 的具有最高集中趋势的度量的动作,并且从多个动作中随机地选择选择概率为 ϵ 的动作。

4. 根据权利要求1-3中任一项中所述的方法,其中所述网络输出包含针对所述动作-当前观察对的多个可能的Q回报中的每一个的相应输出值。

5. 根据权利要求4所述的方法,其中所述集中趋势的度量是所述可能的Q回报的平均值。

6. 根据权利要求5所述的方法,其中相对于所述概率分布确定所述可能的Q回报的平均值包括:

从所述输出值确定所述多个可能的Q回报中的每一个的相应概率;

用所述可能的Q回报的概率对每个可能的Q回报进行加权;以及

通过对所述加权的可能的Q回报求和来确定所述平均值。

7. 一种训练根据权利要求1-6中任一项所述的分布式Q网络的方法,所述方法包括:

获得经验元组包含 (i) 当前训练观察, (ii) 响应于所述当前训练观察由所述智能体执行的当前动作, (iii) 响应于所述智能体执行所述当前动作而接收的当前奖励, 以及 (iv) 下一个训练观察, 其表征由于所述智能体执行所述当前动作而使所述环境所转换成的状态;

确定所述多个可能的Q回报的每个Q回报的相应当前概率, 包括:

使用所述分布式Q网络并根据所述网络参数的当前值处理所述当前训练观察和所述当前动作, 以生成当前网络输出, 所述当前网络输出限定针对所述当前动作-当前训练观察对的可能的Q回报之上的当前概率分布;

对于所述多个动作中的每个动作:

使用目标分布式Q网络并根据所述分布式Q网络的目标网络参数的当前值处理所述动作和所述下一个训练观察, 以生成所述动作-下一个训练观察对的下一个网络输出, 所述下一个网络输出限定针对所述动作-下一个训练观察对的可能的Q回报之上的下一个概率分

布,其中所述目标分布式Q网络具有与所述分布式Q网络相同的神经网络架构,但是所述目标网络参数的当前值不同于所述网络参数的当前值;并且

相对于针对所述动作-下一个训练观察对的相应的下一个概率分布确定所述可能的Q回报的集中趋势的度量;

确定argmax动作,其中所述argmax动作是来自所述多个动作中的所述可能的Q回报的集中趋势的度量最高的一个动作;

使用所述当前奖励和所述argmax动作,确定所述可能的Q回报中的每一个的相应投影样本更新;

相对于损失函数的所述网络参数确定梯度,所述损失函数取决于所述可能的Q回报的投影样本更新和所述可能的Q回报的当前概率;以及

使用所述梯度更新所述网络参数的当前值。

8. 根据权利要求7所述的方法,其中使用所述当前奖励和所述argmax动作,确定所述可能的Q回报中的每一个的相应投影样本更新包括:

从所述当前奖励,确定所述可能的Q回报中的每一个的相应样本更新;以及

从所述相应样本更新和针对所述argmax动作-下一个训练观察对的下一个概率分布中的概率,确定所述可能的Q回报中的每一个的所述相应投影样本更新。

9. 根据权利要求8所述的方法,其中所述可能的Q回报中的每一个的所述相应投影样本更新等于折扣因子和可能的Q回报的乘积加上所述当前奖励,其受限于以下约束条件:所述相应样本更新不小于所述多个可能的Q回报的最小可能的Q回报,并且不大于所述多个可能的Q回报的最大可能的Q回报。

10. 根据权利要求8或9的任一项所述的方法,其中从所述相应样本更新和针对所述argmax动作-下一个训练观察对的下一个概率分布中的概率,确定所述可能的Q回报中的每一个的所述相应投影样本更新包括,对于每个可能的Q回报:

将针对argmax动作-下一个训练观察对的下一个概率分布中的Q回报的概率按强度分配到所述投影样本更新中的至少一些,该强度对于每个投影样本更新基于所述可能的Q回报的样本更新和所述投影样本更新的所述对应Q回报之间的距离。

11. 根据权利要求9或10的任一项所述的方法,其中所述损失函数是在(i)所述相应投影样本更新和(ii)所述当前概率分布之间的Kullback-Leibler散度。

12. 根据权利要求9或10的任一项所述的方法,其中所述可能的Q回报从0至N-1被索引,并且其中确定所述相应投影样本更新包括,对于所述多个可能的Q回报的每个特定可能的Q回报:

将所述特定可能的Q回报的样本更新进行重新映射以落入0至N-1的范围中;

确定所述重新映射的样本更新的下限和上限;

基于针对所述argmax动作-下一个训练观察对的下一个概率分布中的所述特定可能的Q回报的概率,并基于所述重新映射的样本更新和所述上限之间的距离,来更新具有与下限匹配的索引的所述可能的Q回报的投影样本更新;并且

基于针对所述argmax动作-下一个训练观察对的下一个概率分布中的所述特定可能的Q回报的概率,并基于所述重新映射的样本更新和所述下限之间的距离,来更新具有与上限匹配的索引的所述可能的Q回报的投影样本更新。

13. 根据权利要求12所述的方法, 其中所述损失函数是对于每个可能的Q回报的对所述可能的Q回报的投影样本更新和所述可能的Q回报的当前概率的对数之和的负数。

14. 根据权利要求7-13中任一项的所述方法, 还包括:

周期性更新所述目标网络参数的值以匹配所述网络参数的值。

15. 根据权利要求7-13中任一项的所述方法, 还包括:

通过在所述目标网络参数的当前值和所述网络参数的更新值之间插值, 更新目标网络参数的当前值。

16. 一种根据权利要求1-6中任一项所述的训练分布式Q网络的方法, 所述方法包括:

获得经验元组, 该经验元组包含 (i) 当前训练观察, (ii) 响应于所述当前训练观察由所述智能体执行的当前动作, (iii) 响应于所述智能体执行所述当前动作而接收的当前奖励, 以及 (iv) 下一个训练观察, 其表征由于所述智能体执行所述当前动作而使所述环境所转换成的状态;

确定所述多个可能的Q回报的每个Q回报的相应当前概率, 包括:

使用所述分布式Q网络并根据所述网络参数的当前值处理所述当前训练观察和当前动作, 以生成当前网络输出, 所述当前网络输出限定针对所述当前动作-当前训练观察对的可能的Q回报之上的当前概率分布;

对于所述多个动作中的每个动作:

使用分布式Q网络并根据所述网络参数的当前值处理所述动作和所述下一个训练观察, 以生成针对所述动作-下一个训练观察对的下一个网络输出, 所述下一个网络输出限定针对所述动作-下一个训练观察对的可能的Q回报之上的下一个概率分布; 并且

相对于针对所述动作-下一个训练观察对的相应的下一个概率分布确定所述可能的Q回报的集中趋势的度量; 确定argmax动作, 其中所述argmax动作来自所述多个动作中的所述可能的Q回报的集中趋势的度量最高的一个动作;

使用所述当前奖励和所述argmax动作, 确定所述可能的Q回报中的每一个的相应投影样本更新; 相对于损失函数的所述网络参数确定梯度, 所述损失函数取决于所述可能的Q回报的投影样本更新和所述可能的Q回报的当前概率; 以及

使用所述梯度更新所述网络参数的当前值。

17. 根据权利要求16所述的方法, 其中使用所述当前奖励和所述argmax动作, 确定所述可能的Q回报中的每一个的相应投影样本更新包括:

从所述当前奖励, 确定所述可能的Q回报中的每一个的相应样本更新; 以及

从所述相应样本更新和针对所述argmax动作-下一个训练观察对的下一个概率分布中的概率, 确定所述可能的Q回报中的每一个的所述相应投影样本更新。

18. 根据权利要求17所述的方法, 其中所述可能的Q回报中的每一个的所述相应投影样本更新等于所述当前奖励加上折扣因子和可能的Q回报的乘积, 其受限于以下约束条件: 所述相应样本更新不小于所述多个可能的Q回报的最小可能的Q回报, 并且不大于所述多个可能的Q回报的最大可能的Q回报。

19. 根据权利要求17或18的任一项所述的方法, 其中从所述相应样本更新和针对所述argmax动作-下一个训练观察对的下一个概率分布中的概率, 确定所述可能的Q回报中的每一个的所述相应投影样本更新包括, 对于每个可能的Q回报:

将针对所述argmax动作-下一个训练观察对的下一个概率分布中的所述Q回报的概率按强度分配到所述投影样本更新中的至少一些,所述强度对于每个投影样本更新基于所述可能的Q回报的样本更新和所述投影样本更新的所述对应Q回报之间的距离。

20.根据权利要求18或19的任一项所述的方法,其中所述损失函数是在(i)所述相应投影样本更新和(ii)所述当前概率分布之间的Kullback-Leibler散度。

21.根据权利要求17或18的任一项所述的方法,其中所述可能的Q回报从0至N-1被索引,并且其中确定所述相应投影样本更新包括,对于所述多个可能的Q回报的每个特定可能的Q回报:

将所述特定可能的Q回报的样本更新进行重新映射以落入0至N-1的范围中;

确定所述重新映射的样本更新的下限和上限;

基于针对所述argmax动作-下一个训练观察对的下一个概率分布中的所述特定可能的Q回报的概率,并基于所述重新映射的样本更新和所述上限之间的距离,来更新具有与下限匹配的索引的所述可能的Q回报的投影样本更新;并且

基于针对所述argmax动作-下一个训练观察对的下一个概率分布中的所述特定可能的Q回报的概率,并基于所述重新映射的样本更新和所述下限之间的距离,来更新具有与上限匹配的索引的所述可能的Q回报的投影样本更新。

22.根据权利要求21所述的方法,其中所述损失函数是对于每个可能的Q回报的对所述可能的Q回报的投影样本更新和所述可能的Q回报的当前概率的对数之和的负数。

23.一种系统,包括一个或多个计算机和存储指令的一个或多个存储装置,所述指令在由所述一个或多个计算机执行所述指令时使得所述一个或多个计算机执行权利要求1-22中任一项所述的方法的相应操作。

24.一种存储指令的计算机程序产品,所述指令在由一个或多个计算机执行所述指令时使得所述一个或多个计算机执行权利要求1-22中任一项所述的方法的相应操作。

分布式强化学习

背景技术

[0001] 本说明书涉及强化学习。

[0002] 在强化学习系统中,智能体(agent)通过响应于接收对表征环境的当前状态的观察而执行由强化学习系统选择的动作来与环境交互。

[0003] 一些强化学习系统根据神经网络的输出,响应于接收给定观察来选择将由智能体执行的动作。

[0004] 神经网络是机器学习模型,其采用非线性单元的一个或多个层以为接收的输入预测输出。一些神经网络是深度神经网络,其包括除了输出层之外的一个或多个隐藏层。每个隐藏层的输出用作向网络中的下一层(即,下一个隐藏层或输出层)的输入。网络的每个层根据参数的相应集合的当前值从接收的输入生成输出。

发明内容

[0005] 本说明书描述了强化学习系统,其实现为在一个或多个位置中一个或多个计算机上的计算机程序。

[0006] 根据第一方面,提供一种方法,以选择由与环境交互的强化学习智能体要执行的动作。方法包括接收表征环境的当前状态的当前观察。对于与环境交互的智能体要执行的多个动作的集合中的每个动作,使用具有多个网络参数的分布式Q网络处理动作和当前观察。分布式Q网络是配置为根据网络参数的当前值处理动作和当前观察的深度神经网络,以生成网络输出,该网络输出限定针对动作-当前观察对的可能的Q回报之上的概率分布。每个可能的Q回报是从响应于当前观察执行动作的智能体得到的回报的估计。对于每个动作,确定相对于针对动作-当前观察对的概率分布的可能的Q回报的集中趋势的度量。使用动作的集中趋势的度量,选择动作以响应于当前观察由智能体执行。

[0007] 在一些实现方式中,选择将由智能体执行的动作包括选择具有最高集中趋势的度量的动作。

[0008] 在一些实现方式中,选择将由智能体执行的动作包括,选择概率为 $1-\epsilon$ 的具有最高集中趋势的度量的动作,并且从多个动作中随机地选择概率为 ϵ 的动作。

[0009] 在一些实现方式中,网络输出包括针对动作-当前观察对的多个可能的Q回报中的每一个的相应输出值。

[0010] 在一些实现方式中,集中趋势的度量是可能的Q回报的平均值。

[0011] 在一些实现方式中,相对于概率分布确定可能的Q回报的平均值包括从输出值确定多个可能的Q回报中的每一个的相应概率。用可能的Q回报的概率对每个可能的Q回报进行加权。通过对加权的可能的Q回报求和确定平均值。

[0012] 根据第二方面,提供一种训练分布式Q网络的方法。该方法包括获得经验元组,其包含(i)当前训练观察,(ii)响应于当前训练观察由智能体执行的当前动作,(iii)响应于智能体执行当前动作而接收的当前奖励,以及(iv)下一个训练观察,其表征由于智能体执行当前动作而使环境所转换成的状态。确定多个可能的Q回报的每个Q回报的相应当前概

率。更具体地,使用分布式Q网络并根据网络参数的当前值来处理当前训练观察和当前动作,以生成当前网络输出,当前网络输出限定针对当前动作-当前训练观察对的可能的Q回报之上的当前概率分布。对于每个动作,使用目标分布式Q网络(或在一些情况下,分布式Q网络)并根据分布式Q网络的目标网络参数的当前值来处理动作和下一个训练观察,以生成针对动作-下一个训练观察对的下一个网络输出。下一个网络输出限定针对动作-下一个训练观察对的可能的Q回报之上的下一个概率分布。目标分布式Q网络具有与分布式Q网络相同的神经网络架构,但是目标网络参数的当前值不同于网络参数的当前值。对于每个动作,确定相对于针对动作-下一个训练观察对的相应的下一个概率分布的可能的Q回报的集中趋势的度量。确定argmax动作,其中argmax动作是来自多个动作中的可能的Q回报的集中趋势的度量最高的一个动作。使用当前奖励和argmax动作,确定可能的Q回报中的每一个的相应投影样本更新。相对于损失函数的网络参数确定梯度,该损失函数取决于可能的Q回报的投影样本更新和可能的Q回报的当前概率。使用梯度来更新网络参数的当前值。

[0013] 在一些实现方式中,使用当前奖励和argmax动作,确定可能的Q回报中的每一个的相应投影样本更新包括从当前奖励,确定可能的Q回报中的每一个的相应样本更新。从相应样本更新和针对argmax动作-下一个训练观察对的下一个概率分布中的概率,确定可能的Q回报中的每一个的相应投影样本更新。

[0014] 在一些实现方式中,可能的Q回报中的每一个的相应投影样本更新等于折扣因子和可能的Q回报的乘积加上当前奖励。相应样本更新可以确定为受限于以下约束条件:相应样本更新不小于多个可能的Q回报的最小可能的Q回报。并且不大于多个可能的Q回报的最大可能的Q回报。

[0015] 在一些实现方式中,从相应样本更新和针对argmax动作-下一个训练观察对的下一个概率分布中的概率,确定可能的Q回报中的每一个的相应投影样本更新包括,对于每个可能的Q回报:将针对argmax动作-下一个训练观察对的下一个概率分布中的Q回报的概率分配到投影样本更新中的至少一些。概率按强度分布,该强度对于每个投影样本更新基于可能的Q回报的样本更新和投影样本更新的对应Q回报之间的距离。

[0016] 在一些实现方式中,损失函数是在(i)相应投影样本更新和(ii)当前概率分布之间的Kullback-Leibler散度。

[0017] 在一些实现方式中,可能的Q回报从0至N-1被索引。确定相应投影样本更新包括——对于多个可能的Q回报的每个特定可能的Q回报——将特定可能的Q回报的样本更新进行重新映射以落入0至N-1的范围中。确定重新映射的样本更新的下限和上限基于针对argmax动作-下一个训练观察对的下一个概率分布中的特定可能的Q回报的概率,并基于重新映射的样本更新和上限之间的距离,来更新具有与下限匹配的索引的可能的Q回报的投影样本更新。基于针对argmax动作-下一个训练观察对的下一个概率分布中的特定可能的Q回报的概率,并基于重新映射的样本更新和上限之间的距离,来更新具有与上限匹配的索引的可能的Q回报的投影样本更新。

[0018] 在一些实现方式中,损失函数是对于每个可能的Q回报的对可能的Q回报的投影样本更新和可能的Q回报的当前概率的对数之和的负数。

[0019] 在一些实现方式中,方法还包括周期性更新目标网络参数的值以匹配网络参数的值。

[0020] 在一些实现方式中,方法还包括通过在目标网络参数的当前值和网络参数的更新值之间插值,更新目标网络参数的当前值。

[0021] 根据第三方面,提供一种系统,包括一个或多个计算机和存储指令的一个或多个存储装置,该指令在由一个或多个计算机执行指令时使得一个或多个计算机执行先前所描述的方法。

[0022] 根据第四方面,提供一种存储指令的计算机程序产品,该指令在由一个或多个计算机执行指令时使得一个或多个计算机执行先前所描述的方法。

[0023] 可以实现本说明书中所描述的主题的特定实施例,以便实现以下优点中的一个或多个。

[0024] 如本说明书中所描述的强化学习系统使用分布式Q神经网络选择将由智能体执行的动作。对于给定的动作-观察对,分布式Q网络生成限定可能的Q回报的集合之上的概率分布的输出(即,对由响应于观察执行动作的智能体产生的结果进行估计的数值)。可以基于多个数值限定概率分布,该多个数值中的至少两个用于选择将由智能体执行的动作。相比之下,一些常规系统使用常规Q神经网络选择动作,该常规Q神经网络对于给定的动作-观察对而言生成该对的单一期望Q回报。通过对针对动作-观察对的Q回报之上的概率分布进行建模,而不是单一期望Q回报(即,如在一些常规系统中),如本说明书中所描述的系统可以学习更丰富(并且由此更高效)的动作选择模型。例如,如本说明书中所描述的系统可以比一些常规系统更加高效地选择将由与环境交互的智能体执行的动作(例如,允许智能体更高效地执行特定任务)。

[0025] 此外,如本说明书中所描述的系统可以基于损失函数训练,该损失函数取决于全Q回报概率分布,而不是单一期望Q回报(即,如在一些常规系统中),由此产生更加稳定的参数值更新。如果参数值更新一贯地使得系统选择让智能体能够高效执行任务的动作,则这些参数值更新被称为稳定。通过使能更加稳定参数值更新,如本说明书中所描述的系统可以比一些常规系统在训练期间消耗更少的计算资源(例如,计算电力和存储器)。例如,如本说明书中所描述的系统可以比一些常规系统需要的训练过程的更少的迭代以及更少的训练数据以实现可接受的性能水平。

[0026] 在一些情况下,如本说明书中所描述的系统可以用于选择将由与真实世界环境交互的机器人智能体执行的动作。在这些情况下,如本说明书中所描述的系统与在由常规系统选择将由机器人智能体执行的动作情况相比可以使机器人智能体能够更加快速地完成可接受的性能,能够执行更加有效完成任务的动作,以及能够更易于适配于先前未知的环境。例如,智能体可以是执行以下任务的机器人智能体:诸如在位置之间移动物体(如,在运送仓库中),组装组件(如,制造环境中的电子组件),在位置之间导航(如,作为自动化或半自动化运载工具)。通过基于全Q回报分布而不是单一期望Q回报来选择将由智能体执行的动作,如本说明书中所描述的系统可以使机器人智能体能够更加快速且更加安全地执行任务,甚至在先前未知的环境条件下(例如,如果智能体是在重度堵塞中或下雪天进行操作的自动化运载工具)。

[0027] 下面在所附附图和描述中提出在本说明书中的主题的一个或多个实现方式的细节。通过说明书、附图和权利要求书,主题的其他特征、方面和潜在优势将变得显而易见。主体可以表达为计算机实现的方法、配置为执行方法的计算机系统、或计算机程序产品(例

如,多个计算机存储介质中的一个,或者,在另一个示例中,如在通信网络上传输的软件),该计算机程序产品存储程序指令以使得计算机系统执行方法中的一个。

附图说明

[0028] 图1是示例强化学习系统的框图。

[0029] 图2是使用强化学习系统选择将由智能体执行的动作的示例过程的流程图。

[0030] 图3是训练强化学习系统的示例过程的流程图。

[0031] 图4是在强化学习系统的训练期间确定Q回报的投影样本更新的示例过程的流程图。

[0032] 在各附图中的相同的参考数字和命名指示相同的元件。

具体实施方式

[0033] 图1是示例强化学习系统100的框图。强化学习系统100是实现为在一个或多个位置中的一个或多个计算机上的计算机程序的系统的示例,其中实现下文所描述的系统、组件和技术。

[0034] 强化学习系统100选择将由智能体(agent) 104执行的动作102,该智能体104在多个时间步骤中的每一个与环境106交互。在每个时间步骤,系统100接收表征环境106的当前状态的数据,例如环境106的图像,并且响应于接收的数据来选择将由智能体104执行的动作102。表征环境106状态的数据在本说明书中将被称为观察108。

[0035] 本文中所描述的系统100可广泛地应用,并不限于一种具体的实现方式。然而,出于说明的目的,下面描述少量的示例实现方式。

[0036] 在一些实现方式中,环境106是真实世界环境,并且智能体104是与真实世界环境进行交互的机械式智能体。例如,智能体104可以是与环境106进行交互的机器人以完成具体任务。作为另一个示例,智能体104可以是导航通过环境106的自动化或半自动化运载工具。在这些实现方式中,动作102可以控制输入以控制机器人或自动化运载工具。在这些实现方式中的一些中,观察108可以由智能体104的传感器生成或者从智能体104的传感器导出。例如,观察108可以由智能体104的相机捕获。作为另一个示例,观察108可以从智能体104的激光传感器捕获的数据导出。作为另一个示例,观察108可以是由智能体104的高光谱传感器捕获的高光谱图像。

[0037] 在一些实现方式中,环境106可以是模拟的环境,并且智能体104可以实现为与模拟的环境进行交互的一个或多个计算机程序。例如,模拟的环境106可以是视频游戏,并且智能体104可以是玩视频游戏的模拟的用户。作为另一个示例,模拟的环境106可以是例如驾驶模拟或飞行模拟的运动模拟的环境,并且智能体104是导航通过运动模拟环境的模拟的运载工具。在这些实现方式中,动作102可以控制输入以控制模拟的用户或模拟的运载工具。

[0038] 在每个时间步骤中,该时间步骤的环境106的状态(如由观察108表征)取决于先前时间步骤的环境106的状态以及先前时间步骤由智能体104执行的动作102。

[0039] 在每个时间步骤中,系统100可以基于该时间步骤的环境106的当前状态以及智能体104的动作102接收奖励110。总体上,奖励110为数值。奖励110可以基于环境106的任一事

件或方面。例如,奖励110可以指示智能体104是否已经完成任务(例如在环境106中导航到目标位置),或者指示智能体104向完成任务的进度。

[0040] 系统100包括分布式Q神经网络112,其配置为处理包括以下的输入(在本说明书中称为动作-观察对):(i)来自可以由智能体104执行的动作用的集合中的动作102,和(ii)观察108。分布式Q网络112配置为根据分布式Q网络参数的当前值处理动作-观察对,以生成输出,该输出限定针对动作-观察对的可能的Q回报的集合之上的概率分布114。对于给定的动作-观察对而言,Q回报(其还可以被称为Q值)是对由响应于观察执行动作的智能体104产生的回报的估计的数值。回报指代在智能体104在多个时间步骤之上与环境106交互时,由系统100接收的奖励的累积度量。例如,回报可以指代由系统100接收的长期时间-折扣奖励。

[0041] 针对动作-观察对的可能的Q回报的集合是数值的集合。可能的Q回报的集合可以是连续集合或离散集合。例如,针对动作-观察对的可能的Q回报可以是连续的范围:[-2, 2],或离散集合:{-2,-1,0,1,2}。

[0042] 分布式Q网络112的输出是数值(例如向量或矩阵)的有序收集。在一些情况下,特别是当可能的Q回报的集合是离散集合时,分布式Q网络112的输出包括可能的Q回报的集合中的每个Q回报的相应输出。在这些情况下,如果每个Q回报的相应输出值例如由soft-max函数进行处理,则分布式Q网络的输出可以限定可能的Q回报的集合之上的概率分布114,从而确定每个Q回报的相应概率值。在一些情况下,分布式Q网络112的输出包括限定在可能的Q回报的集合之上的参数概率分布的相应输出值。例如,分布式Q网络112的输出可以包括限定在可能的Q回报的集合之上的正态分布的平均值和标准偏差的相应输出值。

[0043] 分布式Q网络112可以实现为任何合适的深度神经网络模型。例如,分布式Q网络112可以实现为前馈神经网络(例如,多层感知器或卷积神经网络)或者递归神经网络(例如,长短期记忆网络)。

[0044] 系统100响应于在每个时间步骤的观察108,将分布式Q网络112用于对将由智能体104执行的动作102进行选择。特别地,在每个时间步骤处,针对来自可以由智能体执行的动作用的集合的每个动作,系统100提供动作-当前观察对作为向分布式Q网络112的输入。分布式Q网络112处理每个动作-当前观察对以生成输出,该输出限定针对每个动作-当前观察对的可能的Q回报的集合之上的相应概率分布114。

[0045] 例如,分布式Q网络112的输出可以限定:(i)针对第一动作-当前观察对的可能的Q回报的集合之上的第一概率分布116,(ii)针对第二动作-当前观察对的可能的Q回报的集合之上的第二概率分布118,以及(iii)针对第三动作-当前观察对的可能的Q回报的集合之上的第三概率分布120。

[0046] 对于每个动作,系统100相对于由针对动作-当前观察对的分布式Q网络112的输出而限定的概率分布,确定可能的Q回报的集合的对应的集中趋势的度量122(即,中心值或典型值)。例如,如进一步将参考图2描述的,集中趋势的度量可以为平均值、中位数、或众数。

[0047] 系统100基于与动作对应的集中趋势的度量122选择在时间步骤将由智能体104执行的动作102。在一些实现方式中,系统100从可以由智能体104执行的动作用的集合中的全部动作之中选择具有最高对应的集中趋势的度量122的动作。在一些实现方式中,系统100根据探索策略选择动作。例如,系统100可以使用 ϵ -贪婪探索策略。在该示例中,系统100可以用概率1选择具有最高对应的集中趋势的度量的动作,并且用概率 ϵ 随机选择动作,其中 ϵ

是0和1之间的数字。

[0048] 系统100包括训练引擎124,其配置为使用强化学习技术在多次训练迭代之上训练分布式Q网络112。训练引擎124通过迭代地(即以每次训练迭代)调整分布式Q网络参数的当前值来训练分布式Q网络112。通过训练分布式Q网络112,训练引擎124可以例如使分布式Q网络112生成输出,该输出导致对将由智能体104执行的动作102进行选择,这提高了由系统100接收的奖励的累积度量。通过由系统100接收的奖励的累积度量(例如长期时间-折扣奖励),训练引擎124可以(通过扩展)使智能体104更有效地执行给定的任务。

[0049] 训练引擎124基于包括多个经验元组126的集合的训练数据来训练分布式Q网络112。每个经验元组包括指示以下的数据:(i)训练观察,(ii)响应于训练观察由智能体执行的动作,(iii)响应于执行动作的智能体而接收的奖励,以及(iv)下一个训练观察,其表征由于执行动作的智能体而环境转换成的状态。经验元组126的集合可以包括在线经验元组、离线经验元组或者两者。在线经验元组指代经验元组,其中根据分布式Q网络参数的当前值,基于由分布式Q网络112生成的输出来选择包括在经验元组中的动作。离线经验元组指代基于任何合适的动作选择策略(例如,随机动作选择策略)来选择包括在经验元组中的动作的经验元组。

[0050] 如进一步将参考图3所描述的,在每次训练迭代时,训练引擎124可以获得并处理经验元组,而对于每个可能的Q回报确定在本说明书中被称为Q回报的投影样本更新的数值。训练引擎124可以相对于分布式Q网络参数确定损失函数的梯度(例如,其取决于投影样本更新),并且使用梯度来更新分布式Q网络参数的当前值。损失函数可以例如鼓励分布式Q网络112生成限定概率分布的输出(即,响应于处理动作-训练观察对),其中每个Q回报的概率值与Q回报的投影样本更新相似。换言之,损失函数可以鼓励分布式Q网络112生成限定可能的Q回报之上的概率分布的输出,其中提高了概率分布和投影样本更新之间的相似度的至少一个数值度量(例如,Kullback-Leibler散度度量)。因此,在一些情况下,投影样本更新的集合可以解释为由分布式Q网络112的输出限定的概率分布的目标值。

[0051] 在一些实现方式中,系统100包括目标分布式Q网络128。总体上,目标分布式Q网络128具有与分布式Q网络112相同的神经网络架构(例如层数、神经拓扑术等)。此外,目标分布式Q网络128(类似于分布式Q网络112)还配置为处理动作-观察对,以生成限定可能的Q回报的集合之上的概率分布的输出。然而,目标分布式Q网络参数的值可以与对应的分布式Q网络参数的值不同。如进一步参考图3所描述的,训练引擎124可以将目标分布式Q网络128用于在训练期间确定Q回报的投影样本更新。

[0052] 系统100可以基于分布式Q网络参数的当前值周期地(例如在训练迭代的给定数目之后)更新目标分布式Q网络参数的值。例如,系统100可以更新目标分布式Q网络参数的值以匹配分布式Q网络参数的当前值。作为另一个示例,系统100可以通过在目标分布式Q网络参数的当前值和分布式Q网络参数的当前值之间插值,来更新目标分布式Q网络参数的值。例如,系统100可以将每个目标分布式Q网络参数的值设定为目标分布式Q网络参数的当前值和对应的分布式Q网络参数的值的平均值。

[0053] 图2使用强化学习系统选择将由智能体执行的动作的示例过程的流程图。为了方便起见,过程200将被描述为由位于一个或多个位置上的一个或多个计算机的系统执行。例如,根据本说明书适当编程的强化学习系统——例如图1的强化学习系统100——可以执行

过程200。

[0054] 系统接收表征环境的当前状态的当前观察(202)。当前观察可以由智能体的传感器生成或者从智能体的传感器导出。例如,当前观察可以由智能体的相机捕获。作为另一个示例,当前观察可以由从智能体的激光传感器捕获的数据导出。作为另一个示例,当前观察可以由智能体的高光谱传感器捕获的高光谱图像。

[0055] 对于来自智能体执行的动作的集合中的每个动作,系统确定可能的Q回报的集合之上的对应概率分布(204)。更具体地,对于每个动作,系统提供动作-当前观察对作为向分布式Q神经网络的输入。分布式Q网络根据分布式Q网络参数的当前值处理每个动作-当前观察对,以生成相应输出。对于由分布式Q网络处理的每个动作-当前观察对,由分布式Q网络生成的输出限定可能的Q回报的集合之上的相应概率分布。

[0056] 针对动作-观察对的可能的Q回报的集合是数值的集合。可能的Q回报的集合可以是连续集合或离散集合。例如,针对动作-观察对的可能的Q回报可以是连续的范围: $[-2, 2]$, 或离散集合: $\{-2, -1, 0, 1, 2\}$ 。

[0057] 对于给定的动作-观察对,分布式Q网络的输出是数值(例如向量或矩阵)的有序收集。在一些情况下,特别是当可能的Q回报的集合是离散集合时,分布式Q网络的输出包括针对可能的Q回报的集合中的每个Q回报的相应输出。在这些情况下,如果每个Q回报的相应输出值例如由soft-max函数进行处理,则分布式Q网络的输出可以限定可能的Q回报的集合之上的概率分布,从而确定每个Q回报的相应概率值。在一些情况下,分布式Q网络的输出包括限定可能的Q回报的集合之上的参数概率分布的相应输出值。例如,分布式Q网络的输出可以包括相应输出值,该相应输出值限定可能的Q回报的集合之上的正态分布的平均值和标准偏差。

[0058] 分布式Q网络可以实现为任何合适的深度神经网络模型。例如,分布式Q网络可以实现为前馈神经网络(例如,多层感知器或卷积神经网络)或者递归神经网络(例如,长短期记忆网络)。

[0059] 对于每个动作,系统相对于由针对动作-当前观察对的分布式Q网络的输出而限定的概率分布,来确定可能的Q回报的集合的对应的集中趋势的度量(即,中心值或典型值)(206)。集中趋势的度量可以例如是平均值、中位数或者众数。在特殊的示例中,集中趋势的度量是平均值,可能的Q回报的集合是离散集合。在这个示例中,对于每个动作,系统可以根据由针对动作-当前观察对的分布式Q网络的输出所限定的概率分布,以Q回报的概率加权可能的Q回报的集合中的每个Q回报。接着系统可以通过对加权的Q回报进行求和来确定平均值。换言之,系统可以由下式确定动作的集中趋势的度量 m :

$$[0060] \quad m = \sum_i p_i \cdot q_i \quad (1)$$

[0061] 其中 i 索引可能的Q回报 $\{q_i\}$,并且根据由针对动作-当前观察对的分布式Q网络的输出所限定的概率分布, p_i 是针对Q回报 q_i 的概率值。

[0062] 系统基于与动作对应的集中趋势的度量选择将由智能体执行的动作(208)。在一些实现方式中,系统从可以由智能体104执行的动作的集合中的全部动作之中选择具有最高对应的集中趋势的度量的动作。在一些实现方式中,系统根据探索策略选择动作。例如,系统可以使用 ϵ -贪婪探索策略。在这个示例中,系统可以选择概率为1的具有最高对应的

集中趋势的度量的动作,并且选择概率为 ϵ 的动作,其中 ϵ 是0和1之间的数字。

[0063] 图3是训练强化学习系统的示例过程的流程图。为了方便起见,过程300将被描述为由位于一个或多个位置上的一个或多个计算机的系统执行。例如,根据本说明书适当编程的强化学习系统——例如图1的强化学习系统100——可以执行过程300。

[0064] 系统获得经验元组 (302)。经验元组包括指示以下的数据: (i) 当前训练观察, (ii) 响应于训练观察由智能体执行的当前动作, (iii) 响应于智能体执行动作而接收的当前奖励,以及 (iv) 下一个训练观察,其表征作为智能体执行动作的结果而使环境转换成的状态。经验元组可以是在线经验元组或离线经验元组。在线经验元组指代其中根据分布式Q网络参数的当前值,基于由分布式Q网络生成的输出来选择包括在经验元组中的动作的经验元组(例如,参考图2所描述的)。离线经验元组指代其中基于任何合适的动作选择策略(例如,随机动作选择策略)来选择包括在经验元组中的动作的经验元组。

[0065] 系统确定可能的Q回报的集合中的每个Q回报的相应当前概率 (304)。更具体地,系统提供当前动作-当前训练观察对(即,来自302中获得的经验元组)作为向分布式Q网络的输入。分布式Q网络根据分布式Q网络参数的当前值处理当前动作-当前训练观察对,以生成输出。由分布式Q网络生成的输出限定了可能的Q回报的集合之上的概率分布,并且特别地,限定每个Q回报的相应当前概率。例如,如果分布式Q网络的输出包括每个Q回报的相应输出值,则系统可以通过使用soft-max函数处理相应输出值来确定每个Q回报的当前概率。

[0066] 对于可以由智能体执行的每个动作,系统确定可能的Q回报的集合之上的对应的下一个概率分布 (306)。更具体地,对于每个动作,系统提供动作-下一个训练观察对作为向分布式Q网络的输入。分布式Q网络根据分布式Q网络参数的当前值处理每个动作-下一个训练观察对,以生成相应输出。对于由分布式Q网络处理的每个动作-下一个训练观察对,由分布式Q网络生成的输出限定可能的Q回报的集合之上的下一个概率分布。

[0067] 在一些实现方式中,为了确定对应的下一个概率分布,系统使用目标分布式Q网络处理每个动作-下一个训练观察对,而不是使用分布式Q网络处理每个动作-下一个训练观察对。总体上,目标分布式Q网络具有与分布式Q网络相同的神经网络架构(例如层数、神经拓扑术等)。此外,目标分布式Q网络还(与分布式Q网络相似)配置为处理动作-观察对,以生成限定可能的Q回报的集合之上的概率分布的输出。然而,目标分布式Q网络参数的值可以与对应的分布式Q网络参数的值不同。

[0068] 对于每个动作,系统相对于由针对动作-下一次训练观察对的分布式Q网络(或目标分布式Q网络)的输出所限定的概率分布,来确定可能的Q回报的集合的对应的集中趋势的度量 (308)。集中趋势的度量可以例如是平均值、中位数或者众数。参考206描述了确定相对于概率分布的可能的Q回报的集合的集中趋势的度量的示例。

[0069] 系统基于与动作对应的集中趋势的度量(例如308中所确定的)来确定动作,其被称为argmax动作 (310)。例如,系统从可以由智能体104执行的动作的集合中的全部动作之中确定argmax动作,以作为具有最高对应的集中趋势的度量的动作。

[0070] 对于可能的Q回报的集合中的每个Q回报,系统确定了在本说明书被称为Q回报的投影样本更新的相应数值 (312)。总体上,系统基于当前奖励(即,来自302中获得的经验元组)和argmax动作(即,310中所确定的)来确定Q回报的投影样本更新。参考图4描述了确定Q回报的投影样本更新的示例过程。

[0071] 系统确定相对于分布式Q网络参数的损失函数的梯度(314)。总体上,损失函数取决于Q回报的投影样本更新(即,312中所确定的)和Q回报的当前概率(即,304中所确定的)。在一些情况下,损失函数鼓励分布式Q网络生成限定概率分布的输出(即响应于处理当前动作-当前训练观察对),其中每个Q回报的概率值与Q回报的投影样本更新相似。例如,损失函数可以是(i)Q回报的投影样本更新和(ii)Q回报的当前概率之间的散度(例如Kullback-Leibler散度)。换言之,损失函数 $\mathcal{L}$ 可以由下式给定:

$$[0072] \quad \mathcal{L} = KL(M||P) \quad (2)$$

[0073] 其中M表示Q回报的投影样本更新,并且P表示Q回报的当前概率。作为另一个示例,当可能的Q回报的集合是离散集合时,损失函数可以是对于每一个Q回报的Q回报的投影样本更新和Q回报的当前概率的对数之和的负数。换言之,损失函数 $\mathcal{L}$ 可以由下式给定:

$$[0074] \quad \mathcal{L} = -\sum_i m_i \log p_i \quad (3)$$

[0075] 其中i索引可能的Q回报的集合,并且对于每个i, m_i 是第i个Q回报的投影样本更新,且 p_i 是第i个Q回报的当前概率值。

[0076] 总体上,系统可以由任何合适的方法来确定相对于分布式Q网络参数的损失函数的梯度。例如,系统可以由反向传播过程确定梯度。

[0077] 系统使用损失函数的梯度更新分布式Q网络参数的当前值(316)。对于分布式Q网络的给定的参数 θ ,系统可以由下式更新参数的当前值:

$$[0078] \quad \theta \leftarrow \theta - r \cdot \nabla_{\theta} \mathcal{L} \quad (4)$$

[0079] 其中r是正学习率超参数(即,管控参数值在训练期间可以多么快速地改变的系统超参数)并且 $\nabla_{\theta} \mathcal{L}$ 是相对于参数 θ 的损失函数的梯度(例如,如在314中所确定的)。

[0080] 图4是在强化学习系统的训练期间确定Q回报的投影样本更新的示例过程的流程图。为了方便起见,过程400将被描述为由位于一个或多个位置上的一个或多个计算机的系统执行。例如,根据本说明书适当编程的强化学习系统——如图1的强化学习系统100——可以执行过程400。

[0081] 随后的过程400的描述参考特定Q回报进行描述。为了确定投影样本更新,系统可以对于可能的Q回报的集合中的每个Q回报执行过程400的步骤。过程400的描述假设可能的Q回报的集合是离散集合。

[0082] 系统确定Q回报的样本更新(402)。Q回报的样本更新是数值。系统可以基于当前奖励(即,如302中所描述的经验元组的当前奖励)、折扣因子(即,在0和1之间的数值),和Q回报自身来确定Q回报的样本更新。例如,系统可以确定Q回报q的样本更新 $\mathcal{T}q$ 为:

$$[0083] \quad \mathcal{T}q = r + \gamma \cdot q \quad (5)$$

[0084] 其中r是当前奖励,并且 γ 是折扣因子。在一些情况下,系统可以受限于以下约束条件来确定Q回报的样本更新:(i)样本更新大于可能的Q回报的集合中的最小Q回报,并且(ii)样本小于可能的Q回报的集合中的最大Q回报。例如,系统可以将样本更新剪切成在[a, b]范围中,其中a是可能的Q回报的集合中的最小Q回报,并且b是可能的Q回报的集合中的最大Q回报。

[0085] 系统将样本更新进行重新映射(即,变换)以落入0至N-1的范围中,其中从0到N-1

索引Q回报,且N是正整数(404)。例如,系统可以由以下关系式确定重新映射的样本更新b:

$$[0086] \quad b = \frac{Tq - V_{MIN}}{\Delta q} \quad (6)$$

[0087] 其中 Tq 样本更新, V_{MIN} 是可能的Q回报的集合中的最小Q回报,并且 Δq 是分离Q回报的增量(即,对于每个 $i \in \{1, \dots, N-1\}$, $q_i - q_{i-1} = \Delta q$)。

[0088] 系统确定重新映射的样本更新的下限和上限(406)。重新映射的样本更新的下限限定为小于重新映射的样本更新的最大整数值。重新映射的样本更新的上限限定为大于重新映射的样本更新的最小整数值。

[0089] 系统更新投影样本更新(408)。如先前所描述,为了确定投影样本更新,系统可以对于可能的Q回报的集合中的每个Q回报执行过程400的步骤。投影样本更新的值可以被初始化(即,在系统对于可能的Q回报的集合中的任何Q回报执行过程400的步骤之前),例如取值为0。每当系统执行过程400的步骤,特别是步骤408,一个或多个Q回报的投影样本更新的值被更新。

[0090] 系统可以通过将针对 $\arg\max$ 动作-下一个训练观察对的下一个概率分布中的Q回报的概率(即,如306中所确定的)分配至一个或多个投影样本更新,来更新投影样本更新。换言之,系统可以基于针对 $\arg\max$ 动作-下一个训练观察对的下一个概率分布中的Q回报的概率来增加一个或多个投影样本更新的值。在一些情况下,系统可以按强度分配针对 $\arg\max$ 动作-下一个训练观察对的下一个概率分布中的Q回报的概率,该强度对于每个投影样本更新基于Q回报的样本更新(即,在402中所确定)和投影样本更新的对应Q回报之间的距离。分配针对 $\arg\max$ 动作-下一个训练观察对的下一个概率分布中的Q回报的概率是指基于Q回报的概率增加一个或多个投影样本更新的值。

[0091] 例如,系统可以更新Q回报的投影样本更新,其具有与重新映射的样本更新的下限匹配(即,如406所确定的)的索引(即,在 $\{0, \dots, N-1\}$ 的范围中)。在这个示例中,系统可以基于以下来更新具有与重新映射的样本更新的下限匹配的索引的Q回报的投影样本更新:(i) 针对 $\arg\max$ 动作-下一个训练观察对的下一个概率分布中的Q回报的概率(即,如306中所确定的),以及(ii) 重新映射的样本更新和重新映射的样本更新的上限之间的距离。例如,在这个示例中,系统可以基于如下关系式来更新具有与重新映射的样本更新的下限匹配的索引的Q回报的投影样本更新:

$$[0092] \quad m_1 \leftarrow m_1 + p \cdot (u - b) \quad (7)$$

[0093] 其中 m_1 是具有与重新映射的样本更新的下限匹配的索引的Q回报的投影样本更新, p 是针对 $\arg\max$ 动作-下一个训练观察对的下一个概率分布中的Q回报的概率, u 是重新映射的样本更新的上限,并且 b 是重新映射的样本更新。

[0094] 作为另一个示例,系统可以更新Q回报的投影样本更新,该Q回报具有与重新映射的样本更新的上限(即,如406所确定的)匹配的索引(即,在 $\{0, \dots, N-1\}$ 的范围中)。在这个示例中,系统可以基于以下来更新具有与重新映射的样本更新的上限匹配的索引的Q回报的投影样本更新:(i) 针对 $\arg\max$ 动作-下一个训练观察对的下一个概率分布中的Q回报的概率(即,如306中所确定的),以及(ii) 重新映射的样本更新和重新映射的样本更新的下限之间的距离。例如,在这个示例中,系统可以基于如下关系式来更新具有与重新映射的样本

更新的上限匹配的索引的Q回报的投影样本更新:

$$[0095] \quad m_u \leftarrow m_u + p \cdot (b-1) \quad (8)$$

[0096] 本说明书结合系统和计算机程序组件使用术语“配置为”。对于要“配置为”执行特定操作或动作的一个或多个计算机的系统,意味着系统已经在其上安装了软件、固件、硬件或它们的组合,其在运行时使得系统执行操作或动作。对于要配置为执行特定操作或动作的一个或多个计算机程序,意味着一个或多个程序包括指令,当由数据处理设备执行所述指令时,使设备执行操作或动作。

[0097] 本说明书中描述的主题和功能操作的实施例可以以数字电子电路,以有形实施的计算机软件或固件、以包括本说明书中公开的结构及其结构等价物的计算机硬件、或者以它们中的一个或多个的组合来实现。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,编码在有形非暂时性存储介质上的计算机程序指令的一个或多个模块,以由数据处理设备执行或控制数据处理设备的操作。计算机存储介质可以是机器可读存储装置、机器可读存储基板、随机或串行存取存储器装置、或它们中的一个或多个的组合。替代地或附加地,程序指令可以被编码在人工生成的传播信号上,例如机器生成的电信号、光信号或电磁信号,其被生成以对信息进行编码,以发送到合适的接收器设备以供数据处理设备执行。

[0098] 术语“数据处理设备”指的是数据处理硬件并且包括用于处理数据的所有类型的设备、装置和机器,包括例如可编程处理器、计算机或多个处理器或计算机。设备可以是或还包括专用逻辑电路,例如FPGA(现场可编程门阵列)或ASIC(专用集成电路)。除硬件之外,设备可以可选地包括为计算机程序创建运行环境的代码,例如代码,其构成处理器固件、协议栈、数据库管理系统,操作系统、或它们中的一个或多个的组合。

[0099] 计算机程序,其也可以被称为或描述为程序、软件、软件应用程序、app、模块、软件模块、脚本或代码,可以用任何形式的编程语言编写,包括编译或解释语言,或者声明或程序语言;且它可以以任何形式部署,包括作为独立程序或作为模块、组件、子程序、或适用于计算环境的其它单元。程序可以但不必对应于文件系统中的文件。程序可以存储在保存其它程序或数据(例如,存储在标记语言文档中的一个或多个脚本)的文件的一部分中、存储在专用于所涉及的程序的单个文件中、或者存储在多个协调文件中(例如,存储一个或多个模块、子程序或代码的部分的文件)。计算机程序可以部署为在一台计算机上或在多台计算机上运行,所述多台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连。

[0100] 在本说明书中,术语“引擎”广泛用于指代被编程为执行一个或多个特定功能的基于软件的系统、子系统或过程。通常,引擎将被实现为一个或多个软件模块或组件,安装在一个或多个位置中的一个或多个计算机上。在某些情况下,一个或多个计算机将专用于特定引擎;在其它情况下,可以在相同的一个或多个计算机上安装和运行多个引擎。

[0101] 本说明书中描述的过程和逻辑流程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,以通过对输入数据进行操作并生成输出来执行功能。过程和逻辑流程也可以由专用逻辑电路(例如FPGA或ASIC)或专用逻辑电路和一个或多个编程的计算机的组合来执行。

[0102] 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两者,或任何其它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者两者

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,以及用于存储指令和数字和/或量子数据的一个或多个存储器装置。中央处理单元和存储器可以由专用逻辑电路补充或并入其中。通常,计算机还将包括或可操作地联接以从和/或向用于存储数据的一个或多个海量存储装置接收和/或传输数据,例如磁盘、磁光盘或光盘。然而,计算机不需要具有这样的装置。此外,计算机可以嵌入在另一个装置中,例如移动电话、个人数字助理(PDA)、移动音频或视频播放器、游戏控制器、全球定位系统(GPS)接收器或便携式存储装置(例如,通用串行总线(USB)闪存驱动器),等等。

[0103] 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、介质和存储器装置,包括例如半导体存储器装置,比如EPROM、EEPROM和闪存装置;磁盘,例如内部硬盘或可移动磁盘;磁光盘;以及CD-ROM和DVD-ROM盘。

[0104] 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在计算机上实现,该计算机具有显示装置(例如CRT(阴极射线管)或LCD(液晶显示器)监视器)以向用户显示信息,该计算机还具有键盘和指点装置,例如鼠标或轨迹球,用户可以通过其向计算机提供输入。其它类型的装置也可用于提供与用户的交互;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。另外,计算机可以通过向用户使用的装置发送文档和从用户使用的装置接收文档来与用户交互;例如,通过响应于从web浏览器接收的请求将网页发送到用户装置上的web浏览器。此外,计算机可以通过向个人装置(例如,正在运行消息收发应用程序的智能电话)发送文本消息或其它形式的消息来与用户交互,以及从用户接收响应消息作为交换。

[0105] 用于实现机器学习模型的数据处理设备还可以包括,例如,专用硬件加速器单元,其用于处理机器学习训练或生产(即推断)工作负载的公共和计算密集部分。

[0106] 可以使用机器学习框架(例如,TensorFlow框架,Microsoft认知工具包框架,Apache Singa框架或Apache MXNet框架)来实现和部署机器学习模型。

[0107] 可以在计算系统中实现本说明书中描述的主题的实施例,该计算系统包括后端组件——例如作为数据服务器;或者包括中间件组件——例如应用程序服务器;或者包括前端组件——例如具有图形用户接口、web浏览器或app的客户端计算机,用户可以通过该图形用户接口或web浏览器或app与本说明书中描述的主题的实现方式进行交互;或者一个或多个这样的后端组件、中间件组件或前端组件的任何组合。系统的组件可以通过数字数据通信的任何形式或介质(例如通信网络)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。

[0108] 计算系统可以包括客户端和服务端。客户端和服务端通常彼此远离并且通常通过通信网络进行交互。客户端和服务端之间的关系是通过运行在相应的计算机上的计算机程序产生的,并且彼此具有客户端-服务器关系。在一些实施例中,服务器将数据(例如,HTML页面)发送到用户装置,例如,用于向与作为客户端的装置交互的用户显示数据和从其接收用户输入的目的。可以在服务器处从装置接收在用户装置处生成的数据,例如,用户交互的结果。

[0109] 虽然本说明书包含许多具体的实现方式细节,但是这些不应该被解释为对任何发明的范围或可以要求保护的范围的限制,而是作为针对于特定发明的特定实施例的特征的

描述。本说明书中在单独实施例的上下文中描述的某些特征也可以组合或在单个实施例中实现。相反,在单个实施例的上下文中描述的各种特征也可以在多个实施例中单独地或以任何合适的子组合来实现。此外,尽管特征可以在上文描述为以某些组合起作用并且甚至最初如此主张,但是来自所要求保护的组合的一个或多个特征可以在一些情况下从该组合中删除,并且所要求保护的组合可以涉及子组合或子组合的变型。

[0110] 类似地,尽管以特定顺序描述在附图中描绘和在权利要求中引述了操作,但这不应被理解为要求以所示的特定顺序或先后顺序执行这样的操作,或者执行所有示出的操作以实现期望的结果。在某些情况下,多任务和并行处理可能是有利的。此外,上文所述的实施例中的各种系统模块和部件的分离不应该被理解为在所有的实施例中都需要这种分离,并且应该理解,所描述的程序组件和系统通常可以一起集成在单个软件产品中或封装成多个软件产品。

[0111] 已经描述了主题的特定实施例。其它实施例在以下权利要求的范围内。例如,权利要求中记载的动作可以以不同的顺序执行并且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的特定顺序或先后顺序,以实现期望的结果。在某些情况下,多任务和并行处理可能是有利的。

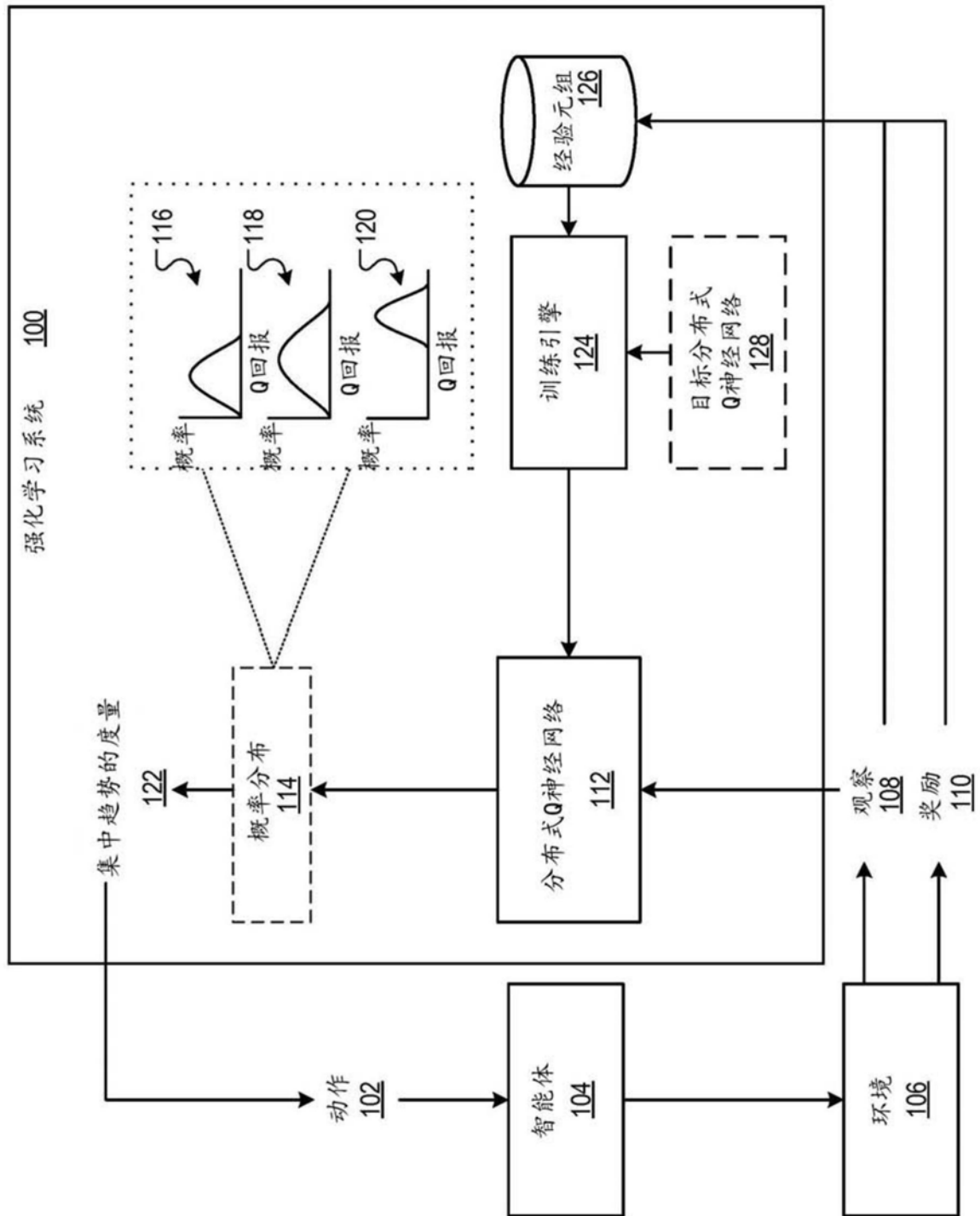


图1

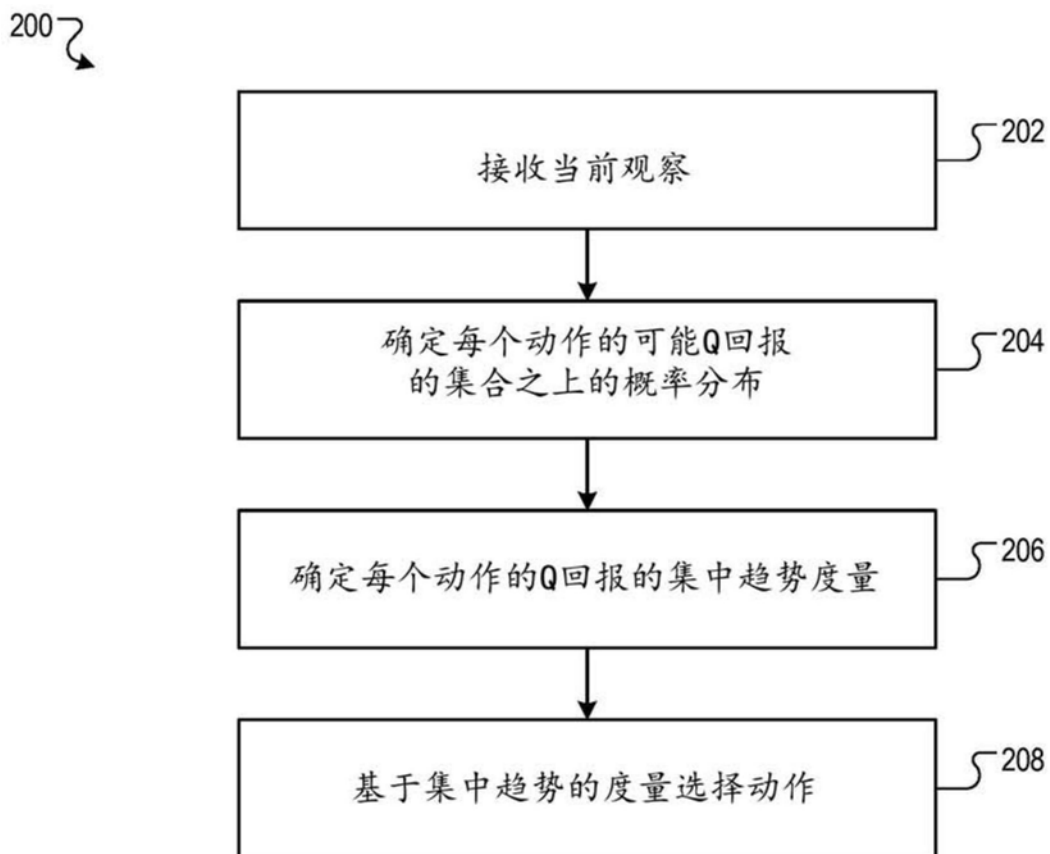


图2

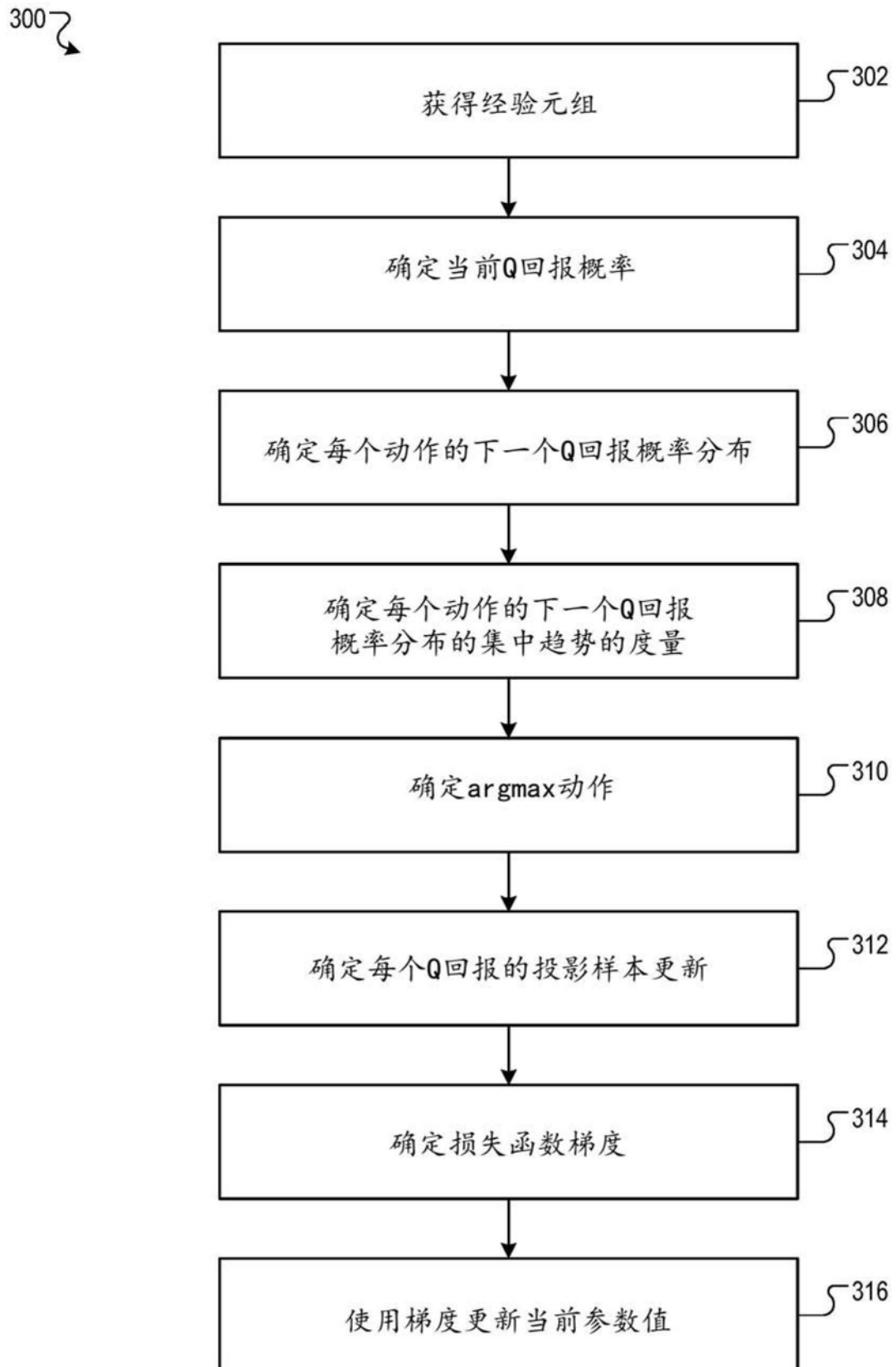


图3

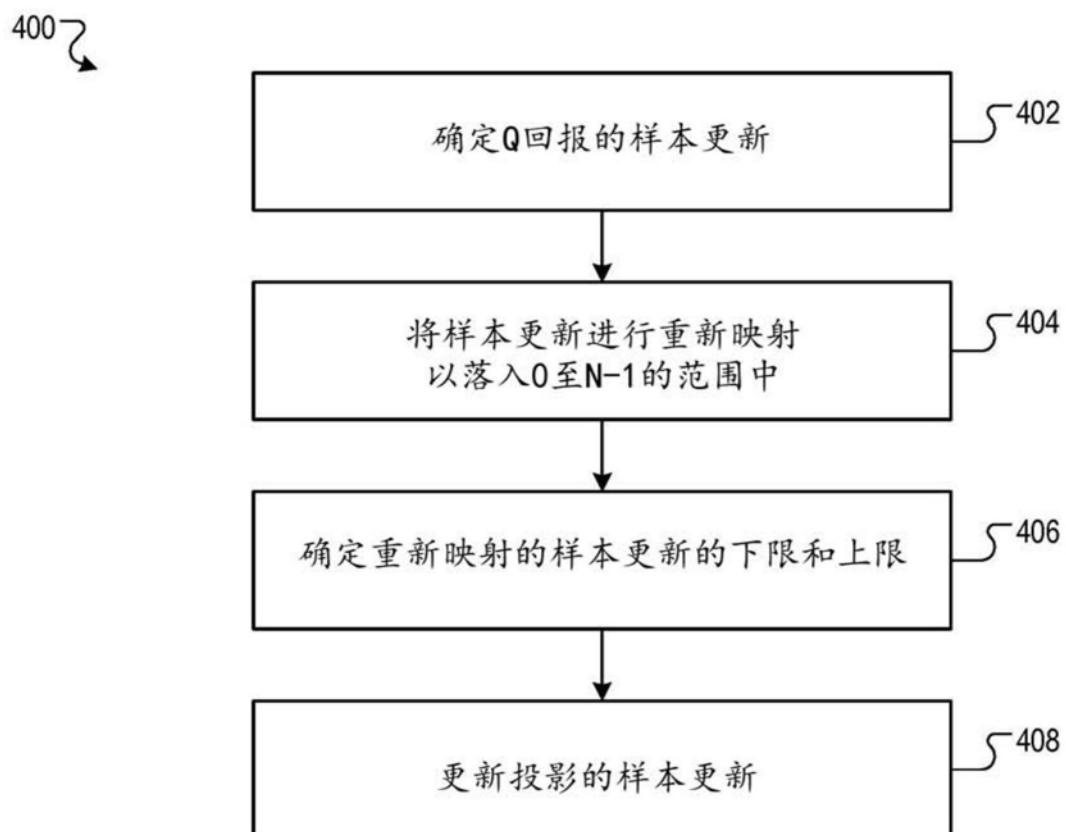


图4