



(12) 发明专利申请

(10) 申请公布号 CN 119487528 A

(43) 申请公布日 2025. 02. 18

(21) 申请号 202380050131.8

(22) 申请日 2023.06.27

(30) 优先权数据

63/356,009 2022.06.27 US

(85) PCT国际申请进入国家阶段日

2024.12.26

(86) PCT国际申请的申请数据

PCT/EP2023/067491 2023.06.27

(87) PCT国际申请的公布数据

W02024/003058 EN 2024.01.04

(71) 申请人 渊慧科技有限公司

地址 英国

(72) 发明人 朱利安·佩罗拉特

巴特·德维尔德

卡尔·保罗·图伊尔斯

(74) 专利代理机构 中原信达知识产权代理有限
责任公司 11219

专利代理师 周亚荣 李宝泉

(51) Int.Cl.

G06N 3/092 (2006.01)

G06N 3/045 (2006.01)

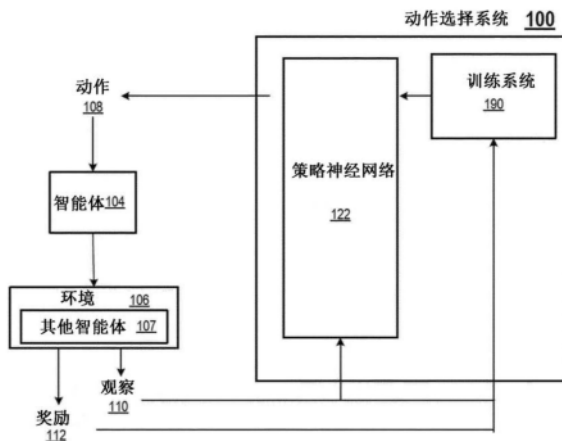
权利要求书3页 说明书19页 附图5页

(54) 发明名称

具有正则化纳什动力学的无模型强化学习

(57) 摘要

用于训练用于控制智能体的策略神经网络的方法、系统和装置,包括编码在计算机存储介质上的计算机程序。特别地,策略神经网络可以通过具有正则化纳什动力学的无模型强化学习被训练。



1. 一种训练策略神经网络的方法,所述策略神经网络用于控制与环境交互以执行任务的智能体,所述方法包括:

初始化用于控制所述智能体的正则化控制策略;以及

在多个训练阶段中的每个训练阶段,执行操作,所述操作包括:

为所述任务生成训练数据,对于执行所述任务的轨迹期间的多个步中的每个步,所述训练数据包括元组,所述元组包括标识(i)观察、(ii)响应于所述元组中的所述观察而执行的动作和(iii)所述任务的奖励的数据;

根据所述训练数据生成经修改的任务的经修改的训练数据,包括,对于每个元组:

使用截至该训练阶段的正则化控制策略,从该元组中的奖励生成变换奖励,以及

生成经修改的元组,所述经修改的元组包括标识(i)该元组中标识的观察、(ii)该元组中标识的动作和(iii)该元组的所述变换奖励的数据;以及

在所述经修改的训练数据上训练所述策略神经网络;以及

在训练所述策略神经网络之后,将所述正则化控制策略更新为使用所述策略神经网络选择动作的策略。

2. 根据权利要求1所述的方法,其中,所述环境包括一个或多个其他智能体,并且执行所述任务需要与所述一个或多个其他智能体交互。

3. 根据权利要求2所述的方法,其中:

对于所述轨迹中各步的真子集,所述动作是由所述智能体响应于所述观察而执行的动作,以及

对于所述轨迹中的所述多个步中的不在所述真子集中的每个步,所述动作是由所述一个或多个其他智能体中的一个智能体响应于所述元组中的所述观察而执行的动作。

4. 根据权利要求3所述的方法,其中,每个步处的所述观察表示(i)对所述智能体和所述一个或多个其他智能体可用的关于所述环境的状态的公共信息,以及(ii)仅对响应于所述观察而执行所述动作的所述智能体可用的关于所述环境的状态的私有信息。

5. 根据权利要求2至4中任一项所述的方法,其中,对于每个元组,该元组中标识的所述任务的所述奖励是针对响应于所述观察而执行所述动作的所述智能体的奖励,并且其中,该元组进一步包括标识针对除了响应于该元组而执行所述动作的所述智能体之外的每个智能体的相应奖励的数据。

6. 根据权利要求2至5中任一项所述的方法,其中,生成所述训练数据包括控制所述智能体使用所述策略神经网络执行所述任务的回合。

7. 根据当从属于权利要求3时的权利要求6所述的方法,其中,生成所述训练数据包括控制所述智能体使用所述策略神经网络执行所述任务的回合,同时还使用所述策略神经网络控制所述一个或多个其他智能体。

8. 根据任一前述权利要求所述的方法,其中,在所述经修改的任务上训练所述策略神经网络包括通过强化学习在测量预期的变换奖励的目标上训练所述策略神经网络。

9. 根据权利要求8所述的方法,其中,训练所述策略神经网络包括与评价者神经网络联合训练所述策略神经网络,所述评价者神经网络被配置为接收观察并处理该观察以生成该观察的预测值,并且其中,所述训练包括:

对于每个轨迹,计算所述轨迹中每个步的相应的值估计,该值估计是在所述环境处于

由该步中的所述元组表征的所述状态后接收的未来变换奖励的估计;以及

使用所述变换奖励和所述轨迹的每个轨迹中各步的相应的值估计来训练所述策略神经网络和所述评价者神经网络。

10. 根据权利要求9所述的方法, 其中, 训练所述评价者神经网络包括使用回归损失来训练所述评价者神经网络, 所述回归损失针对每个轨迹中的每个步测量该步的相应的值估计与通过使用所述评价者神经网络处理该步的元组中的所述观察而生成的预测值之间的误差。

11. 根据权利要求9或权利要求10所述的方法, 其中, 使用所述变换奖励和所述轨迹的每个轨迹中各步的相应的值估计来训练所述策略神经网络和所述评价者神经网络包括通过神经复制器动力学 (NeuRD) 损失来训练所述策略神经网络。

12. 根据任一前述权利要求所述的方法, 其中, 使用所述正则化控制策略从所述元组中的所述奖励生成变换奖励包括:

使用所述正则化控制策略生成第一变换奖励。

13. 根据权利要求12所述的方法, 其中, 使用所述正则化控制策略, 从所述元组中的所述奖励生成变换奖励包括:

使用来自先前训练阶段的先前正则化控制策略生成第二变换奖励; 以及

通过组合所述第一变换奖励和所述第二变换奖励生成所述变换奖励。

14. 根据权利要求13所述的方法, 其中, 通过组合所述第一变换奖励和所述第二变换奖励生成所述变换奖励包括:

计算所述第一变换奖励和所述第二变换奖励的加权和。

15. 权利要求12至14中任一项的方法, 其中, 使用所述正则化控制策略生成第一变换奖励包括:

基于 (i) 由所述策略神经网络指派给由所述智能体在该步处执行的所述动作的概率和 (ii) 由所述正则化策略指派给由所述智能体在该步处执行的所述动作的概率的比率来计算修正子; 以及

将所述修正子应用于所述智能体的所述奖励。

16. 根据权利要求15所述的方法, 其中, 将所述修正子应用于所述智能体的所述奖励包括:

从所述奖励中减去所述修正子。

17. 根据权利要求15或权利要求16所述的方法, 其中, 所述修正子基于正则化参数和所述比率的对数的乘积。

18. 根据当从属于权利要求9时的任一前述权利要求所述的方法, 其中, 计算所述轨迹中每个步的相应的值估计包括:

从轨迹中的最后一步开始递归计算所述相应的值估计并继续直到所述轨迹中的第一步。

19. 根据任一前述权利要求所述的方法, 其中, 所述策略神经网络被配置为接收包括表征所述环境的状态的观察的输入并且生成策略输出, 所述策略输出定义来自响应于所述观察而要执行的动作集合中的动作。

20. 根据任一前述权利要求所述的方法, 其中, 所述智能体是机械智能体, 并且所述环

境是真实世界环境。

21. 根据权利要求20所述的方法,其中,所述智能体是机器人。

22. 根据任一前述权利要求所述的方法,其中,所述环境是包括多个电子装备项目的服务设施的真实世界环境,并且所述智能体是被配置为控制所述服务设施的操作的电子智能体。

23. 根据任一前述权利要求所述的方法,其中,所述环境是用于制造产品的真实世界制造环境,并且所述智能体包括电子智能体,所述电子智能体被配置为控制制造单元或操作以制造所述产品的机器。

24. 根据权利要求1至19中任一项所述的方法,其中,所述环境是真实世界环境的模拟,并且所述智能体是真实世界智能体的模拟,并且其中,所述方法进一步包括:

在所述训练之后,提供所述策略神经网络用于控制所述真实世界环境中的所述真实世界智能体,特别是通过使用所述策略神经网络处理所述环境的观察以生成策略输出来控制所述环境中所述智能体的动作。

25. 根据权利要求1至19中任一项的方法,其中,所述环境是真实世界环境的模拟,并且所述智能体是真实世界智能体的模拟,并且其中,所述方法进一步包括:

在所述训练之后,使用所述策略神经网络控制所述真实世界环境中的所述真实世界智能体,特别是通过使用所述策略神经网络处理所述环境的观察以生成策略输出来控制所述环境中所述智能体的动作。

26. 根据权利要求24或25所述的方法,其中,所述智能体是机械智能体。

27. 根据权利要求26所述的方法,其中,所述智能体是机器人。

28. 根据权利要求24或25所述的方法,其中,所述真实世界环境是包括多个电子装备项目的服务设施的真实世界环境,并且所述真实世界智能体是被配置为控制所述服务设施的操作的电子智能体。

29. 根据权利要求24或25所述的方法,其中,所述真实世界环境是用于制造产品的真实世界制造环境,并且所述真实世界智能体包括电子智能体,所述电子智能体被配置为控制制造单元或操作以制造所述产品的机器。

30. 一种系统,包括:

一个或多个计算机;以及

与所述一个或多个计算机通信耦合的一个或多个存储设备,其中,所述一个或多个存储设备存储指令,所述指令在由所述一个或多个计算机执行时,使所述一个或多个计算机执行根据权利要求1至29中任一项所述的相应方法的操作。

31. 一个或多个非暂时性计算机存储介质,所述一个或多个非暂时性计算机存储介质存储指令,所述指令在由一个或多个计算机执行时,使所述一个或多个计算机执行根据权利要求1至29中任一项所述的相应方法的操作。

具有正则化纳什动力学的无模型强化学习

[0001] 相关申请的交叉引用

[0002] 本申请要求于2022年6月27日提交的美国临时申请第63/356,009号的优先权。在先申请的公开内容被视为本申请的一部分并通过引用并入本申请的公开内容中。

技术领域

[0003] 本说明书涉及使用机器学习模型处理数据。

背景技术

[0004] 机器学习模型接收输入并基于所接收的输入生成输出,例如预测输出。一些机器学习模型是参数模型,并基于所接收的输入和模型的参数的值生成输出。

[0005] 一些机器学习模型是深度模型,其采用多层模型来为所接收的输入生成输出。例如,深度神经网络是一种深度机器学习模型,其包括输出层和一个或多个隐藏层,每个隐藏层对所接收的输入应用非线性变换以生成输出。

发明内容

[0006] 本说明书总体上描述了一种在一个或多个位所的一个或多个计算机上实现为计算机程序的系统,其使用策略神经网络控制与环境交互的智能体在环境中执行任务。

[0007] 本说明书还描述了用于训练策略神经网络的技术,使得策略神经网络可用于有效地控制智能体。

[0008] 更具体地说,本说明书描述了使用称为“正则化纳什动力学”的技术在多个训练阶段训练策略神经网络的技术。

[0009] 描述了一种计算机实现的训练策略神经网络的方法,该策略神经网络用于控制与环境交互的智能体,以及可选地控制一个或多个其他智能体,以执行任务。

[0010] 该方法初始化正则化控制策略,例如任意地,用于控制智能体。该方法涉及多个训练阶段。正则化控制策略可以确定性地或随机地定义响应于观察而要执行的动作。正则化控制策略通常能够是用于控制智能体的任何适当的控制策略,该控制策略为集合中的至少一些动作指派非零得分,并使用得分来选择智能体将执行的动作。在来自多个训练阶段的至少一些训练阶段,并且将如下所述,正则化控制策略取决于在前一训练阶段中为策略神经网络确定的权重。

[0011] 多个训练阶段中的每个训练阶段能够涉及为任务生成训练数据,对于执行任务的轨迹(即动作序列)期间的多个步中的每个步,训练数据包括元组,该元组包括数据,所述数据标识(i)观察、(ii)响应元组中的观察而执行的动作和(iii)任务的奖励。对于由智能体执行的动作,可以通过使用策略神经网络处理观察以生成用于选择动作的策略输出来确定所执行的动作。在实施方式中,其他动作可以由一个或多个其他智能体执行。

[0012] 该方法涉及通过使用截至(当前)训练阶段的正则化控制策略为每个元组从该元组中的奖励生成变换奖励来为修改的任务生成修改的训练数据。在一些实施方式中,还能

够使用(来自先前的训练阶段的)先前的正则化控制策略。生成变换奖励能够涉及使用正则化控制策略来计算应用于奖励以获得变换奖励的修正子。修正子能够例如根据由策略神经网络指派给动作的概率与由正则化策略指派给动作的概率的比率(的对数)确定。由修正子对奖励的修改程度能够通过例如乘以修正子的(具有大于零的值的)正则化参数来确定。对于多个智能体,由策略神经网络控制的智能体的修正子可能具有与其他智能体相反的符号。

[0013] 该方法生成经修改的元组,该元组包括数据,所述数据标识(i)标识元组的观察、(ii)元组中标识的动作和(iii)元组的变换奖励,并且策略神经网络在经修改的训练数据上进行训练。例如,能够使用基于变换奖励的强化学习技术来训练策略神经网络,例如通过使用取决于变换奖励的强化学习目标、使用诸如Adam的优化器(梯度下降优化算法)来更新策略神经网络的诸如权重的参数。该方法能够在更新正则化控制策略之前的训练阶段执行生成训练数据和/或修改训练数据和训练策略神经网络的多个(交错)迭代,例如通过在训练策略神经网络之后控制智能体在环境中行动,以获得更多的训练数据。

[0014] 在训练策略神经网络之后,将正则化控制策略更新为使用策略神经网络选择动作的策略。也就是说,由(在训练阶段结束时的)经训练的策略神经网络定义的控制策略定义正则化控制策略。例如,能够维护与策略神经网络相对应的正则化策略神经网络,并且能够将其参数(例如权重)更新为与截至训练阶段完成的策略神经网络的更新权重相同。在实施方式中,正则化控制策略用于正则化奖励,并且还由奖励的正则化产生。

[0015] 能够实施本说明书中描述的主题的特定实施例,以实现以下优点中的一个或多个优点。

[0016] 通过如本说明书中所述对神经网络进行训练,系统能够训练神经网络以即使在具有非常大的状态空间、非常大的动作空间或两者兼有的任务上也能表现良好。

[0017] 此外,该系统能够训练神经网络以即使在(除了大的状态和/或动作空间之外)具有不完美信息的多智能体场景中也能表现良好,其中观察传达“私有”信息和“公共”信息两者,并且其他智能体的“私有”信息对正在行动的智能体不可用。

[0018] 特别地,该系统通过重复执行训练阶段来训练策略神经网络,在该训练阶段中系统通过在修改的任务上训练神经网络来直接修改底层学习动力学,该修改的任务由使用正则化策略计算的修改奖励来定义。

[0019] 通过修改动力学,系统修改任务,使得策略神经网络即使在存在不完美信息、非常大的状态空间、非常大的动作空间或三者的某种组合的情况下也能有效地学习高质量的控制策略。例如,经修改的奖励能够引导用于训练策略神经网络的强化学习技术朝着纳什均衡的方向发展。通过使用经训练的策略神经网络更新正则化策略并利用使用经更新的正则化策略修改的奖励来重复训练过程,系统能够迭代地接近纳什均衡,特别是无需明确建模其他智能体的私有状态。广义上讲,所描述的技术允许训练到收敛,即使在由不完美信息表征的天文状态或动作空间中也是如此。

[0020] 特别地,系统能够训练策略神经网络以学习高质量的控制策略,而无需对于复杂的真实世界任务或具有大状态空间的其他任务可能难以获得足够数量的任何演示数据,也无需对于复杂的环境可能难以生成或学习的环境模型。

[0021] 本说明书主题的一个或多个实施例的细节在附图和以下描述中阐述。主题的其他

特征、方面和优点将从描述、附图和权利要求中变得显而易见。

附图说明

- [0022] 图1示出了示例动作选择系统。
- [0023] 图2是用于训练策略神经网络的示例过程的流程图。
- [0024] 图3是用于执行训练阶段的示例过程的流程图。
- [0025] 图4是用于生成变换奖励的示例过程的流程图。
- [0026] 图5示出了策略神经网络的示例架构。
- [0027] 图6A示出了训练系统的一个示例训练架构。
- [0028] 图6B示出了训练系统的另一个示例训练架构。
- [0029] 各个图中的相同附图标记和名称指示相同元素。

具体实施方式

- [0030] 图1示出了示例动作选择系统100。动作选择系统100是作为计算机程序在一个或多个位所上的一个或多个计算机上实现的系统的示例,其中实现了下文描述的系统、组件和技术。
- [0031] 动作选择系统100通过选择要由智能体104执行的动作108然后使智能体104执行动作108来控制在环境106中交互的智能体104执行任务。
- [0032] 智能体能够在每个时间步接收相应的奖励112,其中奖励由标量数值指定并且表征例如智能体朝着完成指派的任务的进度。
- [0033] 作为一个特定示例,奖励112能够是稀疏二进制奖励,除非任务由于动作被执行而成功完成,否则该奖励为零,例如,只有当任务由于执行的动作而成功完成时,该奖励才非零,例如,等于一。
- [0034] 作为另一个特定示例,奖励112能够是密集奖励,其测量截至在尝试执行任务的回合期间被接收的单独观察的智能体朝着完成任务的进度,例如,使得非零奖励在成功完成任务之前能够被接收并且经常被接收。
- [0035] 特别地,在每个时间步处,系统100接收表征该时间步处环境状态的观察110。
- [0036] 然后,系统100使用策略神经网络122处理包括观察110的输入。
- [0037] 策略神经网络122被配置为接收包括表征环境106的状态的观察110的输入,并处理该输入以生成策略输出,该策略输出定义响应于观察而要执行的动作104。
- [0038] 例如,策略神经网络122能够生成策略输出,该策略输出为能够由智能体104执行的动作集合中的每个动作指派相应的得分。在该示例中,系统能够例如选择具有最高得分的动作或将得分映射到概率分布,并且然后从概率分布中采样动作。
- [0039] 作为另一个示例,为了选择动作104,策略神经网络122能够生成策略输出,该策略输出包括动作集合上的得分分布的参数,例如,可能动作的连续空间上的高斯分布的参数。在该示例中,系统能够例如选择表示分布的平均值的动作或从分布中采样动作。
- [0040] 作为另一个示例,动作集合能够包括多个子集,例如,每个子集中的动作是执行任务所需的不同步部分期间的有效动作。
- [0041] 在该示例中,策略神经网络122能够包括多个“策略头部”,每个“策略头部”对应于

不同的动作子集并生成相应的输出,该输出要么 (i) 为对应子集中的每个动作指派相应的得分,要么 (ii) 包括定义对应子集中动作上的得分分布的参数。

[0042] 在该示例中,为了选择动作104,系统能够选择子集之一,并且然后使用与子集相对应的头部的输出来选择动作,例如如上所述。例如,每个子集能够在执行任务的回合期间在不同的时间步集合处有效,或者能够在不同的环境状态集合处有效。然后,系统能够基于当前时间步或当前时间步处的当前环境状态来选择子集和对应的头部。

[0043] 策略神经网络122能够具有允许神经网络122将输入格式的数据映射到策略输出的任何适当架构。例如,策略神经网络122能够包括卷积层、全连接层、循环层、注意力层等中的一个或多个。

[0044] 下文参考图5更详细地描述了一个示例架构。

[0045] 在一些实施方式中,执行任务可能需要智能体104与环境106中也执行相同任务或执行一个或多个其他任务的一个、两个或更多个其他智能体107交互。

[0046] 其他智能体107的操作可以由计算机实现的方法来控制,例如,类似于控制智能体104的方法,或者一些或全部可以部分或整体由人为控制。也就是说,其他智能体也能够使用策略神经网络122的实例(或不同的学习策略)或由固定的计算机实现的策略来控制。

[0047] 通常,由智能体104和其他智能体执行的任务可以具有相同的特征,例如,它们可以都是路由任务、导航任务、组装任务等,或者它们可以都是调度任务,或者它们可以都是制造任务。

[0048] 其他任务可以与智能体104的任务相同或不同。

[0049] 由智能体执行的任务可能有助于实现共同目标,或者也可能是对抗性的,即,由其他智能体执行的任务可能会阻碍智能体104执行任务。

[0050] 通常,智能体104在环境中执行任务的能力受到在环境中执行一个或多个其他任务的一个或多个其他智能体的影响。

[0051] 在智能体104与其他智能体交互的一些实施方式中,不同的智能体能够在执行任务的给定回合期间在不同的时间步处执行动作。也就是说,对于任务回合中时间步的真子集(即小于全部的子集),在环境中执行的动作是由智能体104响应于时间步处的观察110而执行的动作。对于轨迹中不在该真子集中的每个时间步,一个或多个其他智能体中的一个智能体都会执行动作。

[0052] 作为特定示例,当智能体104与单个其他智能体交互时,智能体104和该其他智能体能够在交替的时间步处执行动作。

[0053] 替代地,不同的智能体能够在同一时间步处执行动作,即两个或更多个智能体能够在同一时间步处执行相应的动作。

[0054] 附加地,在智能体104与其他智能体交互的一些实施方式中,由每个动作在给定时间步处接收到的观察能够包括不同的信息。

[0055] 也就是说,观察110能够包括 (i) 对于智能体104和一个或多个其他智能体107可用的关于环境106的状态的公共信息,以及 (ii) 仅对于响应于观察110执行动作的智能体可用的关于环境106状态的私有信息。

[0056] 例如,关于环境106的状态的公共信息包括能够在给定时间步处由环境中所有智能体的传感器测量的信息。公共信息还能够包括被提供给所有智能体的关于环境的静态信

息,例如环境地图或表征环境静态属性的其他数据。

[0057] 私有信息能够包括例如仅能够由单个智能体的传感器测量的信息,例如本体感受传感器读数或来自智能体内部传感器的传感器读数。私有信息还能够包括表征由智能体正在执行的任务的信息,例如,指令智能体执行哪个任务的条件输入,例如,表征智能体目标的自然语言语句或表征环境目标状态的目标图像。

[0058] 在一些实施方式中,观察110包括公共信息张量以及私有信息张量,该公共信息张量包括给定时间步的公共信息,该私有信息张量包括给定时间步的私有信息。以该方式分离私有信息和公共信息能够使神经网络122学习如何有效地结合这两种不同类型的信息。

[0059] 此外,在一些实施方式中,公共信息包括关于智能体104的公共的信息,并且能够被其他智能体107观察到。也就是说,观察110中的公共信息能够为智能体104提供关于对于其他智能体107可用的关于智能体104在给定时间步处的状态的公共信息的知识。

[0060] 如下文将描述,在一些实施方式中,观察110包括来自先前时间步的历史信息,或者策略神经网络122否则以来自先前时间步的信息为条件,例如,借助包括一个或多个循环神经网络层。在这些实施方式中,对于在时间步处正在行动并由策略神经网络控制的任何给定智能体,系统仅以公共历史信息 and 给定智能体的私有历史信息为条件而不来自任何其他智能体的私有历史信息为条件,来条件化策略神经网络122。

[0061] 系统100还包括训练系统190,其训练策略神经网络122,使得策略神经网络122能够用于有效地控制智能体104。

[0062] 在实施方式中,训练系统190使用称为“正则化纳什动力学”的技术,通过无模型强化学习在多个训练阶段上训练策略神经网络122。

[0063] 由训练系统190执行的训练被称为“无模型”强化学习,因为训练系统190使用不使用环境显式模型的强化学习技术,即,对由智能体104和其他智能体107执行的动作对环境状态的影响进行建模。

[0064] 通过如本文所述训练策略神经网络122,策略神经网络122即使在存在不完善信息(即,当其他智能体的私有信息对正在行动的智能体不可观察时)、非常大的状态空间、非常大的动作空间或三者的某种组合的情况下,也能够有效地学习高质量的控制策略。

[0065] 如下文将描述,在训练期间,一个或多个其他智能体107还可以由策略神经网络122控制,以便生成用于训练的训练数据。

[0066] 在训练后,其他智能体107能够由不同的策略控制,例如由人为操作的控制器或由不同软件控制的控制器控制,或者能够继续由策略神经网络122的实例控制。

[0067] 下文参考图2至4更详细地描述训练。

[0068] 在一些实施方式中,环境是真实世界环境,智能体是与真实世界环境交互的机械智能体,例如机器人或载具,例如在环境中操作或导航通过环境的自主或半自主陆地、空中或海上载具,并且动作是由机械智能体在真实世界环境中为执行任务而采取的动作。例如,智能体可以是与环境交互以完成特定任务的机器人,例如以在环境中定位感兴趣的对象或将感兴趣的对象移动到环境中的指定位置或导航到环境中的指定目的地。

[0069] 作为另一个示例,由策略神经网络控制的智能体可以是载具,例如有或没有人类驾驶员,并且可以通过使用策略神经网络来控制智能体以控制驾驶员必须遵守的交通信号,诸如交通灯和车道控制信号。如果存在,则一个或多个其他智能体也可以是机械智能

体,诸如机器人或人类驱动的或自主或半自主的载具,例如组成载具队列。因此,该系统的一些实施方式能够用于控制交通,例如道路交通,或仓库机器人,或诸如无人机的飞行器,以促进交通的平稳流动。

[0070] 在这些实施方式中,观察可以包括例如以下中的一个或多个:图像、对象位置数据和传感器数据,以在智能体与环境交互时捕获观察,例如来自图像、距离或位置传感器或来自致动器的传感器数据。例如,在机器人的情况下,观察可以包括表征机器人当前状态的数据,例如以下中的一个或多个:关节位置、关节速度、关节力、扭矩或加速度,例如重力补偿扭矩反馈,以及由机器人所持项目的全局或相对姿势。在机器人或其他机械智能体或载具的情况下,观察可能类似地包括智能体的一个或多个部件的位置、线速度或角速度、力、扭矩或加速度以及全局或相对姿势中的一个或多个。观察可以被定义为1、2或3维,并且可以是绝对和/或相对观察。观察还可以包括例如感测到的电子信号,诸如电机电流或温度信号;和/或图像或视频数据,例如来自相机或激光雷达传感器的数据,例如来自智能体的传感器的数据或来自环境中与智能体分开定位的传感器的数据。在某些交通控制应用中,观察可以包括描绘环境中的静态对象的地图数据和/或描绘环境中的非静态对象的实时道路条件数据,该静态对象例如道路、树木、建筑物等,该非静态对象例如行人、其他载具、道路信号的状态——例如交通灯的颜色——等。

[0071] 在这些实施方式中,动作可以是用于控制机器人或其他机械智能体的控制信号或用于控制自主或半自主的陆地、空中、海上载具的控制信号,该用于控制机器人或其他机械智能体的控制信号例如是机器人关节的扭矩或更高级别的控制命令,该用于控制自主或半自主的陆地、空中、海上载具的控制信号例如是控制表面或其他控制元件的扭矩(例如,载具的转向控制元件)或更高级别的控制命令。控制信号能够包括例如机器人的一个或多个关节或另一个机械智能体的部件的位置、速度或力/扭矩/加速度数据。控制信号还可以或代替地包括电子控制数据,诸如电机控制数据,或更一般地包括用于控制环境内的一个或多个电子设备的数据,其控制对环境的观察状态有影响。例如,在自主或半自主的陆地或空中或海上载具的情况下,控制信号可以定义用于控制导航(例如转向)和移动(例如载具的制动和/或加速)的动作。在交通控制设置中,动作还可以或代替地包括物理移动,例如向前移动、左转或右转、让行等,以及发信号动作,例如向另一个智能体发送消息或信号以显示智能体的意图。在这样的设置中,处理一个或多个历史观察以及当前观察能够用于指示例如载具的智能体之间的历史关系。奖励可以指示智能体是否已完成任务,例如导航到环境中的目标位所,或者智能体朝着完成任务的进度,或者一些其他测量,诸如智能体的进度的速度、使用的能量或燃料、到达目的地的估计时间、与另一载具相比到达目的地的相对次序等。在使用时,另一个智能体的假定奖励可以对应于这些中的一个或多个。

[0072] 在一些实施方式中,环境是上述真实世界环境的模拟,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。例如,模拟环境可以是机器人或载具的模拟,并且强化学习系统可以在模拟上进行训练,然后一旦训练后在真实世界中使用。

[0073] 在一些实施方式中,环境是用于制造诸如化学、生物或机械产品或食品的产品真实世界制造环境。如本文所用,“制造”产品还包括精炼起始材料以创造产品,或处理起始材料,例如以去除污染物,以产生清洁或回收的产品。制造工厂可以包括多个制造单元,诸如用于化学或生物物质的容器,或用于处理固体或其他材料的机器,例如机器人。制造单元

被配置成使得产品的中间版本或组件在产品制造期间可在制造单元之间移动,例如,经由管道或机械输送。如本文所用,产品的制造还包括由厨房机器人制造食品。

[0074] 智能体可以包括被配置为控制制造单元的电子智能体或操作以制造产品的诸如机器人的机器。也就是说,智能体可以包括控制系统,该控制系统被配置为控制化学、生物或机械产品的制造。例如,控制系统可以被配置为控制制造单元或机器中的一个或多个,或者控制产品的中间版本或成分在制造单元或机器之间的移动。

[0075] 作为一个示例,由智能体执行的任务可以包括制造产品或其中间版本或成分的任务。作为另一个示例,由智能体执行的任务可以包括控制——例如最小化——资源使用的任务,诸如控制电力消耗或水消耗或制造过程中使用的任何材料或消耗品的消耗的任务。

[0076] 动作可以包括控制动作,以控制机器或制造单元的使用,用于处理固体或液体材料以制造产品或其中间体或成分,或控制产品的中间版本或成分在制造环境内的移动,例如,在制造单元或机器之间。通常,动作可以是对环境的观察状态有影响的任何动作,例如,被配置为调整下文所述的任何感测参数的动作。这些动作可以包括调整制造单元的物理或化学条件的动作,或控制机器的机械部件或机器人的关节的移动的动作。动作可以包括对制造单元或机器施加操作条件的动作,或导致设置改变以调整、控制或开启或关闭制造单元或机器的操作的动作。

[0077] 奖励或回报可以与任务的执行度量有关。例如,在制造产品的任务的情况下,度量可以包括制造的产品的数量、产品的质量、产品的生产速度或执行制造任务的物理成本的度量,例如,用于执行任务的能量、材料或其他资源的数量的度量。在要控制资源的使用的任务的情况下,度量可以包括资源使用的任何度量。

[0078] 通常,对环境状态的观察可以包括表示装备的电子项目和/或机械项目的运行的任何电子信号。例如,环境状态的表示可以从感测制造环境状态的传感器做出的观察中得出,该传感器例如是感测制造单元或机器的状态或配置的传感器,或感测制造单元或机器之间材料移动的传感器。作为一些示例,这样的传感器可以被配置为感测机械移动或力、压力、温度;电气条件,诸如电流、电压、频率、阻抗;一种或多种材料的数量、水平、流动/移动速率或流动/移动路径;物理或化学条件,例如物理状态、形状或配置或化学状态,诸如pH;单元或机器的配置,诸如单元或机器的机械配置或阀门配置;图像或视频传感器,用于捕获制造单元或机器或移动的图像或视频观察;或任何其他适当类型的传感器。在诸如机器人的机器的情况下,来自传感器的观察可以包括对机器的一个或多个部件的位置、线速度或角速度、力、扭矩或加速度或姿势的观察,例如表征机器或机器人的当前状态或由机器或机器人持有或处理的项目的数据。观察还可以包括例如诸如电机电流或温度信号的感测到的电子信号,或例如来自相机或激光雷达传感器的图像或视频数据。诸如这些的传感器可以是环境中智能体的一部分或与智能体分开定位。

[0079] 在一些实施方式中,环境是服务设施的真实世界环境,该服务设施包括电子装备的多个项目,诸如服务器场或数据中心,例如电信数据中心,或用于存储或处理数据的计算机数据中心,或任何服务设施。服务设施还可以包括控制装备项目的操作环境的辅助控制装备,例如诸如温度控制的环境控制装备,例如冷却装备,或气流控制或空调装备。任务可以包括控制——例如最小化——资源使用的任务,诸如控制电力消耗或水消耗的任务。智能体可以包括电子智能体,其被配置为控制装备项目的操作,或控制辅助——例如环境

的——控制装备的操作。

[0080] 通常,动作可以是对环境的观察状态有影响的任何动作,例如,被配置为调整下文所述的任何感测参数的动作。这些可以包括控制装备或辅助控制装备的项目或对装备或辅助控制装备的项目施加操作条件的动作,例如,导致设置改变以调整、控制或开启或关闭装备或辅助装备的项目的操作的动作。

[0081] 通常,对环境状态的观察可以包括表示设施或设施中装备运行的任何电子信号。例如,环境状态的表示可以从由感测设施的物理环境的状态的任何传感器做出的观察、或由感测装备的一个或多个项目或辅助控制装备的一个或多个项目的状态的任何传感器做出的观察中得出。这些包括被配置为感测以下的传感器:诸如电流、电压、功率或能量的电气条件;设施的温度;设施内或设施的冷却系统内的流体流动、温度或压力;或物理设施配置,诸如通风口是否打开。

[0082] 奖励或回报可以与任务的执行的度量有关。例如,在诸如控制电力或水的使用的任务的用于控制——例如最小化——资源使用的任务的情况下,度量可以包括资源使用的任何度量。

[0083] 在一些实施方式中,环境是发电设施的真实世界环境,例如诸如太阳能发电场或风电场的可再生发电设施。任务可以包括控制任务以控制由设施产生的电力,例如以控制电力输送到配电网,例如以满足需求或降低电网元件之间失配的风险,或最大化由设施产生的电力。智能体可以包括电子智能体,其被配置为控制设施的发电或者将产生的电力耦合到电网中。当存在多个智能体时,每个智能体可以控制由不同的相应发电实体——例如设施或设施的一部分——的发电。动作可以包括控制电力发电机的电气或机械配置的动作,该电气或机械配置诸如一个或多个可再生发电元件的电气或机械配置,例如以控制风力涡轮机或一个或多个太阳能电池板或镜子的配置,或者控制旋转发电机的电气或机械配置。机械控制动作可以例如包括控制能量输入到电能输出的转换的动作,例如转换效率或能量输入到电能输出的耦合度。电气控制动作可以例如包括控制所产生的电力的电压、电流、频率或相位中的一个或多个的动作。

[0084] 奖励或回报可以与任务的执行度量有关。例如,在控制向配电网输送电力的任务的情况下,该度量可能与传输功率的测量有关,或与发电设施和电网之间的诸如电压、电流、频率或相位失配的电气失配的测量有关,或与发电设施中的电力或能量损失的测量有关。在最大化向配电网输送电力的任务的情况下,该度量可能与传输到电网的电力或能量的测量有关,或与发电设施中的电力或能量损失的测量有关。

[0085] 一般而言,环境状态的观察可以包括表示发电设施中发电装备的电气或机械功能的任何电子信号。例如,环境状态的表示可以从任何传感器做出的观察中得出,这些传感器感测发电设施中正在发电的装备的物理或电气状态,或这样的装备的物理环境,或支持发电装备的辅助装备的条件。这样的传感器可以包括被配置为感测装备的电气条件的传感器,所述电气条件诸如电流、电压、功率或能量;物理环境的温度或冷却;流体流动;或装备的物理配置;以及例如来自本地或远程传感器的对电网的电气条件的观察。环境状态的观察还可以包括关于发电装备的操作的未来条件的一个或多个预测,诸如对未来风力级别或太阳辐照度的预测或对电网未来电气条件的预测。

[0086] 作为另一个示例,环境可以是化学合成或蛋白质折叠环境,使得每个状态是蛋白

质链或一种或多种中间体或前体化学品的相应状态,并且智能体是用于确定如何折叠蛋白质链或合成化学品的计算机系统。在该示例中,动作是用于折叠蛋白质链的可能折叠动作或用于组装前体化学品/中间体的动作,并且要实现的结果可以包括例如折叠蛋白质以使蛋白质稳定并使其实现特定的生物功能或为化学品提供有效的合成途径。作为另一个示例,智能体可以是机械智能体,其自动执行或控制由系统选择的蛋白质折叠动作或化学合成步而无需人工干预。观察可以包括对蛋白质或化学品/中间体/前体的状态的直接或间接观察和/或可以从模拟中得出。

[0087] 以类似方式,环境可以是药物设计环境,使得每个状态是潜在药物活性化合物的相应状态,并且智能体是用于确定药物活性化合物的元素和/或药物活性化合物的合成途径的计算机系统。药物/合成可以基于从药物目标得出的奖励来设计,例如在模拟中。作为另一个示例,智能体可以是执行或控制药物合成的机械智能体。

[0088] 作为进一步的示例,动作可以包括呈现广告,观察可以包括广告闪现或点击计数或点击率,并且奖励可以表征由一个或多个用户先前选择的项目或内容。

[0089] 在某些情况下,观察可以包括由第三方(例如,智能体的操作员)向智能体提供的文本或口头指令。例如,智能体可以是自主载具,并且自主载具的用户可以向智能体提供文本或口头指令(例如,导航到特定位置)。

[0090] 作为另一个示例,环境可以是电气、机械或机电设计环境,例如,模拟电气、机械或机电实体的设计的环境。模拟环境可以是实体预期在其中工作的真实世界环境的模拟。任务可以是设计实体。观察可以包括表征实体的观察,即,实体的机械形状或电气、机械或机电配置的观察,或实体的参数或属性的观察。动作可以包括修改实体例如修改一个或多个观察的动作。奖励或回报可以包括实体设计的一个或多个性能度量。例如,奖励或回报可以涉及实体的诸如重量或强度的一个或多个物理特性,或者涉及实体的诸如执行实体被设计的特定功能的效率的测量的一个或多个电气特性。设计过程可以包括输出制造设计,例如以用于制造实体的计算机可执行指令的形式。该过程可以包括根据设计制造实体。因此,可以通过例如强化学习来优化实体的设计,然后输出用于制造实体的优化设计,例如作为计算机可执行指令;然后可以制造具有优化设计的实体。作为另一个特定示例,智能体可以是配置为路由集成电路的互连线的路由智能体,观察可以是组件位置和互连,动作可以是路由动作,例如互连或组件选择和/或放置动作,并且奖励可以取决于路由度量,诸如互连长度、电阻、电容、阻抗、损耗、速度或传播延迟,和/或物理线参数,诸如宽度、厚度或几何形状以及设计规则。在路由之后,可以使用路由的互连线制造集成电路。

[0091] 在一些应用中,一个或多个智能体是软件智能体,并且环境是真实世界的计算环境。在一个示例中,智能体管理跨计算资源——例如在移动设备上和/或数据中心中——的任务分布。在这些应用中,观察可以包括对诸如计算和/或存储器容量或互联网可访问资源的计算资源的观察;并且动作可以包括将任务指派给特定计算资源。奖励可以被配置为最大化或最小化以下中的一个或多个:计算资源的利用、电力、带宽和计算速度。

[0092] 在另一个示例中,软件智能体管理例如由一个或多个真实世界的服务器对连续到达的作业队列的处理。观察可以包括对连续作业的出发时间、或连续作业出发之间的时间间隔、或服务器处理每个作业所花费的时间的观察,例如一系列时间的开始和结束,或连续作业的到达时间,或到达之间的时间间隔,或表征作业类型的数据。动作可以包括将特定作

业分配给特定计算资源的动作;奖励可以被配置为最小化总体排队或处理时间或一个或多个单独作业的排队或处理时间,或者通常基于观察来优化任何度量。

[0093] 在一些应用中,环境是真实世界的计算环境,并且软件智能体管理跨计算资源——例如在移动设备上和/或数据中心中——的任务/作业的分布。在这些实施方式中,观察可以包括与计算资源在处理任务/作业中的操作有关的观察,动作可以包括将任务/作业指派给特定计算资源,并且奖励可以与使用计算资源处理任务/作业的一个或多个度量有关,例如计算资源、带宽或电力的使用度量,或处理时间度量,或数值准确性度量,或与计算资源之间期望的负载平衡相关的一个或多个度量。

[0094] 在一些应用中,环境是数据分组通信网络环境,并且一个或多个智能体是一个或多个路由器的一部分,用于通过通信网络路由数据分组。动作可以包括数据分组路由动作,并且观察可以包括例如路由表的观察,该路由表包括路由度量,诸如路由路径长度、带宽、负载、跳数、路径成本、延迟、最大传输单元(MTU)和可靠性的度量。奖励可以与一个或多个路由度量相关地定义,即被配置为最大化一个或多个路由度量。

[0095] 如前所述,环境可以是模拟环境。通常在模拟环境的情况下,观察可以包括一个或多个先前描述的观察或观察类型的模拟版本,并且动作可以包括一个或多个先前描述的动作或动作类型的模拟版本。例如,模拟环境可以是运动模拟环境,例如驾驶模拟或飞行模拟,并且智能体可以是导航通过运动模拟的模拟载具。在这些实施方式中,动作可以是控制输入以控制模拟用户或模拟载具。通常,智能体可以被实现为与模拟环境交互的一个或多个计算机。

[0096] 模拟环境可以是特定真实世界环境和智能体的模拟。例如,该系统可用于在系统的训练或评估期间选择模拟环境中的动作,并且在训练或评估或两者完成后,可以部署用于控制作为模拟主体的特定真实世界环境中的真实世界智能体。这能够避免对真实世界环境或真实世界智能体造成不必要的磨损和损坏,并且能够允许控制神经网络在很少发生或在真实世界环境中难以或不安全来重现的情况下进行训练和评估。例如,该系统可以使用特定真实世界环境模拟中的机械智能体模拟来部分训练,并且过后被部署以控制特定真实世界环境中的真实机械智能体。因此,在这样的情况下,模拟环境的观察与真实世界环境有关,并且模拟环境中的选择的动作与由机械智能体在真实世界环境中要执行的动作有关。

[0097] 可选地,在上述任何实施方式中,任何给定步处的观察可以包括可能有利于表征环境的来自先前步的数据,例如,在先前步处执行的动作、在先前步接收到的奖励或两者。

[0098] 在一些实施方式中,执行任务可能需要智能体与环境中也执行相同任务或执行一个或多个其他任务的一个、两个或更多个其他智能体进行交互。其他智能体的操作可以由计算机实施的方法来控制,例如,类似于控制第一智能体的方法,或者一些或全部可以部分或全部由人为控制。

[0099] 通常,任务可以具有相同的特性,例如,它们可以都是路由任务、导航任务、组装任务等,或者它们可以都是调度任务,或者它们可以都是制造任务。其他任务可以与第一智能体的任务相同或不同。由智能体执行的任务可能有助于实现共同目标,或者可能是对抗性的,即,由其他智能体执行的任务可能会阻止智能体执行任务。在典型的实施方式中,第一智能体在环境中执行任务的能力受到在环境中执行一个或多个其他任务的一个或多个其他智能体的影响。

[0100] 图2是用于训练策略神经网络的示例过程200的流程图。为方便起见,过程200将被描述为由位于一个或多个位所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动作选择系统,例如图1的动作选择系统100,能够执行过程200。

[0101] 特别地,系统初始化正则化策略(步骤202)。通常,正则化策略是向动作集合中的每个动作指派相应(非零)可能性的策略。

[0102] 正则化策略能够被初始化为向动作集合中的一些或所有动作指派非零可能性的任何任意策略。例如,正则化策略能够是向每个动作指派相同得分的统一策略,也能够是向每个动作指派随机得分的随机策略。

[0103] 在初始化正则化策略之后,系统跨多个训练阶段(也称为“级(stage)”)执行训练(步骤204)。

[0104] 在每个训练阶段,系统使用截至该训练阶段的正则化策略来训练策略神经网络以更新策略神经网络的权重(步骤206)。

[0105] 例如,在系统在第一训练阶段训练策略神经网络之前,系统能够使用适当的机器学习初始化技术初始化策略神经网络的权重。

[0106] 对于每个后续训练阶段,系统能够在从前一个训练阶段处确定的权重开始训练策略神经网络,即,无需在训练阶段之间重置权重的值。

[0107] 使用正则化策略执行训练阶段将在下文参考图3进行描述。

[0108] 在每个训练阶段完成时,系统更新正则化策略(步骤208)。

[0109] 特别地,系统将正则化策略更新为等于使用截至训练阶段完成的策略神经网络控制智能体的策略,即,具有截至训练阶段完成已经确定的更新权重的策略神经网络或具有截至训练阶段完成所确定的目标权重的策略神经网络。“目标权重”是指在训练阶段期间更新权重时计算为权重的移动平均值的权重。

[0110] 使用策略神经网络控制智能体的策略能够是例如根据由策略神经网络生成的得分对动作进行采样,向由策略神经网络生成的得分添加噪声,然后使用带噪声得分对动作进行采样的一个策略等。

[0111] 图3是用于执行训练阶段的示例过程300的流程图。为方便起见,过程300将被描述为由位于一个或多个位所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动作选择系统,例如图1的动作选择系统100,能够执行过程200。

[0112] 在每个训练阶段,系统生成任务的训练数据,该数据反映了在执行任务时执行一个或多个动作轨迹的结果(步骤302)。

[0113] 动作轨迹能够包括执行任务的回合中的一些或全部步的数据。

[0114] 回合是智能体尝试执行任务的交互时段,即从任务的初始状态开始并持续直到智能体成功完成任务、与环境交互阈值步数量或满足不同终止标准的时段。

[0115] 例如,每个轨迹能够包括回合中所有步的数据,或者可以包括来自回合中多达最大数量的步。

[0116] 因此,对于执行任务的轨迹中的多个步中的每个步,训练数据包括元组,该元组包括标识(i)观察、(ii)响应于元组中的观察而执行的动作和(iii)任务的奖励的数据。

[0117] 例如,系统能够通过任务回合过程中使用策略神经网络控制智能体来生成每个轨迹。

[0118] 当任务需要多智能体交互时,系统还能够使用策略神经网络控制在环境中交互的其他智能体。

[0119] 因此,在这些情况下,对于轨迹中步的真子集,动作是由智能体响应于观察而执行的动作,并且对于不在真子集中的轨迹中的多个步中的每个步,动作是由一个或多个其他智能体之一响应于元组中的观察而执行的动作。

[0120] 因此,轨迹中不同的步处的不同元组中的动作能够是由不同智能体执行的动作。

[0121] 此外,对于每个元组,元组中标识的任务的奖励是响应于观察而执行动作的智能体的奖励。在某些情况下,元组还包括标识除响应于元组而执行动作的智能体之外的每个智能体的相应奖励或假定奖励(的值)的数据。例如,当两个智能体正在执行对抗性任务时,在给定步处未行动的智能体的奖励能够是在给定步处行动的智能体的奖励的负数,即,大小相等但符号相反。

[0122] 系统从训练数据生成用于修改的任务的经修改的训练数据(步骤304)。

[0123] 更具体地说,对于每个元组,系统使用截至训练阶段的正则化控制策略从元组中的奖励生成变换奖励,并生成修改的元组,该修改的元组包括标识(i)元组中标识的观察、(ii)元组中标识的动作和(iii)元组的变换奖励的数据。

[0124] 当每个元组有多个智能体和多个奖励时,系统能够为元组中的每个奖励——即为轨迹中每个步针对每个智能体——生成相应的变换奖励。

[0125] 因此,系统修改任务的“动力学”以生成经修改的任务,即,在环境中行动导致变换奖励(而不是原始训练数据中的实际奖励)被接收的任务。

[0126] 使用正则化控制策略修改奖励将在下文参考图4中更详细地描述。

[0127] 生成经修改的训练数据能够被称为训练阶段的“奖励变换”级。

[0128] 然后,系统通过强化学习在经修改的训练数据上,例如在测量预期变换奖励的目标上,训练策略神经网络(步骤306)。

[0129] 例如,目标能够是最大化预期回报,其中回报被计算为预期未来变换奖励的总和或时间折扣总和。当存在多个智能体时,目标能够是最大化每个智能体的预期回报,其中回报被计算为该智能体的预期未来变换奖励的总和或时间折扣总和。

[0130] 训练策略神经网络(以及可选地,评价者神经网络)能够被称为训练阶段的动力学级。

[0131] 通常,系统能够使用任何适当的强化学习技术在经修改的训练数据上训练策略神经网络。

[0132] 作为特定示例,系统能够使用演员-评价者强化学习技术联合评价者神经网络训练策略神经网络。

[0133] 评价者神经网络是被配置为接收观察并处理观察以生成观察的预测值的神经网络,其中预测值是从环境处于由观察所表征的状态开始将接收的回报(就变换奖励而言)的预测。

[0134] 例如,能够将评价者神经网络实现为策略神经网络的分开的输出头部,该分开的输出头部给定策略神经网络的输入观察生成值预测。

[0135] 具体地,在该示例中,系统针对每个轨迹计算轨迹中每个步的相应值估计,该相应值估计是环境处于由该步中的元组所表征的状态之后接收的未来变换奖励的组合——例

如总和或时间折扣总和——的估计,然后使用变换奖励和每个轨迹中步的相应值估计来训练策略神经网络和评价者神经网络。

[0136] 当存在多个智能体时,系统针对每个轨迹计算每个智能体在轨迹中每个步的相应值估计,该值估计是环境处于由该步中的元组所表征的状态之后所接收的针对智能体的未来变换奖励的估计,然后使用变换奖励和每个轨迹中步的智能体的相应值估计来训练策略神经网络和评价者神经网络。

[0137] 在一些实施方式中,系统计算每个智能体的相应值估计,例如以递归方式,从轨迹的最后一步开始,并且一直持续到轨迹的第一步,即使用由策略神经网络和评价者神经网络生成的变换奖励和输出。以该方式计算值估计能够减少偏差。例如,系统能够使用V-迹估计器计算每个智能体的相应值估计,例如如Espeholt et al.arXiv:1802.01561中所述。

[0138] 在一些实施方式中,为了进一步减少偏差,即使轨迹小于完整回合的长度,系统也能够使用来自完整回合的数据来计算值估计,即无需任何引导。下面描述了当轨迹短于完整回合时计算跨完整回合的估计的一种示例技术。

[0139] 在一些实施方式中,当使用V-迹估计器计算值估计时,系统使用策略神经网络和评价者神经网络的权重的目标值。例如,在训练阶段,目标值能够被保持为权重的当前值的移动平均值。

[0140] 例如,系统能够使用相应的值估计来训练评价者神经网络,并使用取决于由评价者神经网络生成的输出——例如取决于使用评价者神经网络生成的值估计、Q函数估计或两者——的损失来训练策略神经网络。

[0141] 作为特定示例,但并非必需,系统能够通过基于由评价者神经网络生成的输出计算的神经复制器动力学(NeuRD)损失来训练策略神经网络。D.Hennes et al.,Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS) (自主智能体和多智能体系统(AAMAS)的国际会议论文集)(2020)中更详细地描述了NeuRD损失。例如,NeuRD损失能够是取决于以下的度量的关于策略神经网络权重的梯度的可能动作的总和的损失:(i) 轨迹中步的观察的策略输出与目标值之间的差,例如,观察的策略输出的动作的平均值,以及(ii) 动作的优势估计,该优势估计使用值估计计算并测量响应于观察执行动作相对于执行来自集合的不同动作的相对优势(就从变换奖励计算出的估计回报而言)。

[0142] 当存在多个智能体时,对于每个步,系统能够仅计算在NeuRD损失上训练策略神经网络时在该步处行动的智能体的损失。例如,总体损失能够是轨迹中所有时间步的平均NeuRD损失。

[0143] 为了训练评价者神经网络,系统能够在回归损失上训练神经网络,该回归损失针对每个轨迹中每个步测量该步的相应值估计与通过使用评价者神经网络处理该步的元组中的观察而生成的预测值之间的误差。

[0144] 当存在多个智能体时,对于每个步,系统在回归损失上训练评价者神经网络时能够仅使用在该步中行动的智能体的值估计。例如,回归损失能够是轨迹中所有时间步的平均平方误差。

[0145] 通常,在给定的训练阶段内,系统能够交错训练和训练数据的生成,即,使得在多个训练步中的每个训练步处,使用截至训练步的已经收集的一些或全部训练数据来训练神

经网络。

[0146] 因此,通过在训练阶段期间交错执行步骤302至306的迭代,将根据策略神经网络的不同权重生成不同的元组。

[0147] 替代地,系统能够首先生成训练阶段的所有训练数据,并且然后在训练数据上训练策略神经网络。

[0148] 完成该阶段后,系统使用如上所述的策略神经网络更新正则化策略。

[0149] 通过重复执行其中系统通过在经修改的任务上训练神经网络来直接修改底层学习动力学的这些训练阶段,系统能够训练神经网络,以即使在具有非常大的状态空间、非常大的动作空间或两者兼有的任务上也能表现良好。此外,系统能够训练神经网络,以即使在(除了大的状态和/或动作空间之外)具有不完美信息的多智能体场景中也能表现良好,其中观察传达“私有”信息和“公共”信息两者,并且其他智能体的“私有”信息对正在行动的智能体不可用。

[0150] 因此,训练技术被称为“正则化纳什动力学”,因为在每个训练级,系统都会学习策略,该策略(近似地)收敛到由使用正则化策略生成的经修改的训练数据定义的经修改任务的固定点。

[0151] 在一些实施方式中,在对多个训练阶段中的每个训练阶段执行过程300之后,系统使用经修改的训练技术对策略神经网络进行微调。

[0152] 例如,系统能够通过计算训练中使用的值估计时对由策略神经网络生成的对数几率——即动作得分——应用不同的投影来微调策略神经网络。例如,系统能够利用不同的投影对训练进行微调,该投影使用指定的阈值水平和概率量子数对动作概率进行阈值化和离散化。该阈值水平和概率量子数能够是训练过程的超参数。执行该微调能够帮助减少由策略神经网络由于softmax投影为空间中的每个动作指派非零概率而导致的低概率错误。

[0153] 图4是使用正则化策略修改轨迹中给定步的奖励的示例过程400的流程图。为方便起见,过程400将被描述为由位于一个或多个位所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动作选择系统,例如图1的动作选择系统100,能够执行过程400。

[0154] 系统使用截至训练阶段的正则化控制策略从在给定步处的奖励生成第一变换奖励(步骤402)。

[0155] 为了生成第一变换奖励,系统计算(i)由策略神经网络指派给由智能体在该步处执行的动作的概率(根据动作被选择时的权重)与(ii)由正则化策略指派给由智能体在该步处执行的动作的概率的比率。

[0156] 然后,系统根据该比率计算修正符。

[0157] 作为一个示例,系统能够基于正则化参数和比率的对数的乘积来计算修正子。

[0158] 正则化参数能够是训练的超参数,并且管控正则化策略对实际奖励的影响强烈程度。例如,较大的正则化参数能够在单个训练阶段中在策略神经网络的训练期间产生更稳定和更快的收敛,但能够提供对正则化控制策略的偏差,这可能意味着在多个训练阶段中到收敛的较慢整体训练。

[0159] 然后,系统将修正子应用于智能体的奖励以生成第一变换奖励。例如,当接收奖励的智能体是在该步处行动的智能体时,系统能够从奖励中减去修正子以生成第一变换奖

励。当接收奖励的智能体不是在该步处行动的智能体时,系统能够将修正子添加到奖励中以生成第一变换奖励。

[0160] 在某些情况下,系统将变换奖励设置为等于第一变换奖励。

[0161] 在一些其他情况下,系统通过将来自前一训练阶段的正则化策略并入计算经修改的奖励,对训练阶段之间从一种正则化策略到另一种正则化策略的过渡进行平滑。对于第一训练阶段,即,没有前一训练阶段的情况下,系统能够将来自前一训练阶段的正则化策略设置为等于第一训练阶段的正则化策略。

[0162] 特别地,在这些情况下,系统使用来自前一训练阶段的前一正则化控制策略生成第二变换奖励(步骤404)。也就是说,系统能够以与第一变换奖励相同的方式计算第二变换奖励,但使用来自前一训练阶段的前一正则化控制策略代替当前训练阶段的正则化控制策略。

[0163] 然后,系统通过组合第一变换奖励和第二变换奖励来生成变换奖励(步骤406)。例如,系统能够计算第一变换奖励和第二变换奖励的加权和以生成变换奖励。

[0164] 在训练阶段期间执行多个训练迭代的一些实施方式中,加权和中的权重能够取决于迭代,其中在策略神经网络训练期间稍后发生的迭代将更高的权重指派给第一变换奖励。通常,每个训练迭代对应于训练阶段期间对策略神经网络权重的更新。例如,每次训练迭代能够对应于在训练迭代期间在训练数据批量(或例如在下面描述的SIMD架构中的并行的一组多个训练数据批量)上训练神经网络。

[0165] 在这些实施方式中的一些中,在阈值训练迭代之后,系统不再使用第二变换奖励,并将变换奖励设置为等于第一变换奖励。

[0166] 例如,指派给第一变换奖励的权重能够等于1和常数 a ——例如二、三或四——中的最小值乘以当前迭代的索引与总迭代次数的比率。指派给第二变换奖励的权重能够等于1减去指派给第一变换奖励的权重。因此,在该示例中,一旦在给定的训练阶段期间训练完成超过 $1/a$,系统就会停止使用第二变换奖励,并将变换奖励设置为等于第一变换奖励。

[0167] 作为某些特定示例,在一些实施方式中,能够通过添加或减去项 $\eta \log \frac{\pi(a|o)}{\pi_{reg}(a|o)}$ 来生成变换奖励,其中 η 是正则化参数, $\pi(a|o)$ 是由策略神经网络定义的动作选择策略(例如,观察的动作概率 α 或观察的历史 o), $\pi_{reg}(a|o)$ 是正则化控制策略。仅作为示例, η 可以在0到1的范围内,例如0.2。在一些实施方式中,添加到奖励或从奖励中减去的项能够是 $\eta \gamma(a|o)$,其中 $\gamma(a|o) = a \log \frac{\pi(a|o)}{\pi_{reg,curr}(a|o)} + (1-a) \log \frac{\pi(a|o)}{\pi_{reg,prev}(a|o)}$,其中 $\pi_{reg,curr}$ 是当前正则化控制策略,并且 $\pi_{reg,prev}$ 是先前的正则化控制策略,并且 a 在它们之间进行插值,例如,对于 Δ_m 次总迭代中的迭代 n 为 $a = \min(1, 2n/\Delta_m)$ 。

[0168] 图5示出了策略神经网络122的示例架构。

[0169] 在图5的示例中,策略神经网络122包括神经网络躯干520和三个策略神经网络头部540。

[0170] 此外,在图5的示例中,评价者神经网络被实现为接收神经网络躯干520的输出的值头部550。

[0171] 如图5所示,躯干520处理观察510以生成环境状态的嵌入作为输出。

[0172] 策略神经网络头部540和值头部550中的每一个被配置为接收包括环境的状态的

嵌入的输入,并处理输入以生成相应的输出,例如相应的策略输出或值估计。

[0173] 如上所述,当策略神经网络122包括多个策略头部540时,每个策略头部对应于集合中的不同动作子集。

[0174] 在此示例中,为了选择动作,系统能够选择子集中的一个,然后使用与该子集相对应的头部的输出选择动作,例如如上所述。例如,每个子集能够在任务的不同“阶段”有效,例如,在执行任务的回合期间的不同时间步的不同集合处有效,或者能够在环境状态的不同集合处有效。

[0175] 可选地,系统还能够接收特定于阶段的观察520,其也作为输入提供给头部540和550。特定于阶段的观察520能够包括针对任务的不同阶段的不同类型的数据。例如,数据能够针对不同阶段处于不同的粒度级别,例如,在任务开始时具有较高粒度级别,并且在任务结束时具有较低粒度级别。

[0176] 在图5的示例中,策略神经网络122是卷积神经网络。更具体地说,在图5的示例中,躯干520和头部540和550各自包括相应的金字塔模块560,该金字塔模块560包括卷积神经网络层的“金字塔”,例如,具有卷积残差块(“convresblocks”),后面是反卷积残差块(“deconvresblocks”)。

[0177] 更一般地,策略神经网络122能够具有不同类型的卷积架构或具有不同类型的神经网络层,例如循环层、注意力层等,的架构。

[0178] 图6A示出了用于执行训练阶段的训练系统的一个示例训练架构。

[0179] 如图6A所示,在训练阶段,多个演员602独立且异步地使用如上所述的策略神经网络122生成训练数据(“经验”),并将生成的训练数据存储在重放缓冲存储器604中。例如,演员能够使用“自对弈(self-play)”,其中环境中的所有智能体都由策略神经网络122控制。单个学习器606从重放缓冲存储器604重复采样训练数据,并使用训练数据训练如上所述的策略神经网络122。然后,学习器606定期将作为训练的结果生成的更新权重(“参数”)发送给演员602,以便演员602能够使用更新的权重来生成附加的训练数据。

[0180] 演员602和学习器606能够以多种配置实现。例如,每个演员602能够在相应的机器上实现,并且学习器606能够在不同的机器上实现。

[0181] 作为另一个示例,所有演员602和学习器606都能够在同一机器上实现。在该示例中,演员602能够被实现为主要使用CPU资源的软件,缓冲器604能够主要使用RAM或其他设备存储器,而学习器606主要使用由机器的机器学习硬件加速器提供的资源。

[0182] 可选地,训练系统190还能够包括一个或多个评估器608,该一个或多个评估器608评估策略神经网络122的性能。如图6A的示例所示,评估器608能够通过使用不同的控制策略(“机器人(bot)”)控制其他智能体来评估策略神经网络122的性能。

[0183] 图6B示出了用于执行训练级的训练系统的另一个示例训练架构。

[0184] 如图6B所示,在训练阶段期间,系统利用多个学习器机器660使用单指令多数据(SIMD)模型来缩放训练管道。

[0185] 在该示例中,若干学习器机器660参与训练,并且每个机器660实现多个演员602、缓冲器604和学习器606。每个机器660内的演员602异步生成训练数据,如上所述。附加地,每个机器660内的学习器606与其他机器660的学习器606同步执行学习步:每个学习器从对应的缓冲器604读取不同训练数据批量并且计算网络参数梯度;然后系统190计算跨所有机

器660的梯度的平均值;并且最后每个学习器606将那些平均梯度作为更新应用于网络权重。

[0186] 如上所述,在一些实施方式中,为了进一步减少偏差,即使轨迹小于完整回合的长度,系统也能够使用来自完整回合的数据来计算值估计,即,无需任何引导。现在将描述当回合是可变长度时在配置600或650中计算这些估计的一个示例技术。

[0187] 计算可变长度轨迹的精确回报需要在重放缓冲区中存储完整的回合(代替固定长度的时间步序列)。当学习器606对回合批量进行采样时,较短的回合会填充到批量中最长回合的长度。批量沿时间轴被切成可变数量的固定长度组块。然后,学习器606对该组块列表进行两次传递:a)前向传递以重新计算每个时间步的精确策略和轨迹统计数据;b)后向传递,其中每个组块再次被重新处理,但在该组块的开始时该组块被给出精确的网络状态和完整的轨迹统计数据,诸如精确的回报等。请注意,(a)发生在梯度计算之外并用于避免对于回报和对于网络状态两者的引导或任何种类的近似。而(b)发生在损失之内,并且用于计算网络权重的梯度。

[0188] 对每个组块计算的梯度在(b)期间被累积,并使用优化器在(b)传递结束时被应用于网络权重。

[0189] 本说明书与系统和计算机程序组件相结合,使用术语“被配置”。对于被配置为执行特定操作或动作的一个或多个计算机的系统,意味着该系统已在其上安装了软件、固件、硬件或它们的组合,这些软件、固件、硬件或它们的组合在操作中使系统执行操作或动作。对于要被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,所述指令在由数据处理装置执行时,使装置执行操作或动作。

[0190] 本说明书中描述的主题和功能操作的实施例能够在数字电子电路中、在有形体现的计算机软件或固件中、在计算机硬件中(包括本说明书中公开的结构及其结构等效物)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能够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一个或多个计算机程序指令模块,用于由数据处理装置执行或控制数据处理装置的操作。计算机存储介质能够是机器可读存储设备、机器可读存储基板、随机或串行访问存储器设备,或它们中的一个或多个的组合。可替代地或另外,程序指令能够被编码在人工生成的传播信号上,例如机器生成的电、光或电磁信号,该信号被生成以对信息进行编码,以便传输到合适的接收器装置,供数据处理装置执行。

[0191] 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,包括例如可编程处理器、计算机或多个处理器或计算机。该装置还能够是或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,该装置还能够可选地包括为计算机程序创建执行环境的代码,例如构成处理器固件、协议栈、数据库管理系统、操作系统或它们中的一个或多个的组的代码。

[0192] 计算机程序,也可以称为或描述为程序、软件、软件应用、app、模块、软件模块、脚本或代码,能够用任何形式的编程语言编写,包括编译或解释语言,或声明或过程语言;并且它能够以任何形式部署,包括作为独立程序或作为适合在计算环境中使用的模块、组件、子程序或其他单元。程序可以但不需要对应于文件系统中的文件。程序能够存储在保存其他程序或数据的文件的一部分中,例如,存储在标记语言文档中的一个或多个脚本,专用于

讨论中的程序的单个文件中,或多个协调文件中,例如,存储一个或多个模块、子程序或代码部分的文件。计算机程序能够被部署在一个计算机上或位于一个站点或分布在多个站点并通过数据通信网络互连的多个计算机上执行。

[0193] 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所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;在其他情况下,可以在同一个或多个计算机上安装和运行多个引擎。

[0194] 本说明书中描述的过程和逻辑流程能够由一个或多个可编程计算机实施一个或多个计算机程序以通过对输入数据进行操作并生成输出来执行功能来执行。过程和逻辑流程还能够由例如FPGA或ASIC的专用逻辑电路或由专用逻辑电路和一个或多个编程计算机的组合来执行。

[0195] 适合执行计算机程序的计算机能够基于通用或专用微处理器或两者,或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两者接收指令和数据。计算机的基本元件是用于执行或实施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器设备。中央处理单元和存储器能够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耦合以从一个或多个用于存储数据的例如磁盘、磁光盘或光盘的大容量存储设备接收数据或向其传输数据或两者都有。然而,计算机不需要具有这样的设备。此外,计算机能够被嵌入到另一设备中,例如移动电话、个人数字助理(PDA)、移动音频或视频播放器、游戏机、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)闪存驱动器,仅举几例。

[0196] 适合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、介质和存储器设备,包括例如半导体存储器设备,例如EPROM、EEPROM和闪存设备;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。

[0197] 为了提供与用户的交互,本说明书中描述的主题的实施例能够在具有用于向用户显示信息的例如CRT(阴极射线管)或LCD(液晶显示器)监视器的显示设备的计算机上实现,以及键盘和例如鼠标或轨迹球的指点设备,用户能够通过该指点设备向计算机提供输入。其他种类的设备也能够用于提供与用户的交互;例如,提供给用户的反馈能够是任何形式的感官反馈,例如,视觉反馈、听觉反馈或触觉反馈;并且来自用户的输入能够以包括声学、语音或触觉输入的任何形式接收。此外,计算机能够通过向由用户使用的设备发送文档和从由用户使用的设备接收文档来与用户交互;例如,通过响应于从Web浏览器接收到的请求将网页发送到用户设备上的Web浏览器。此外,计算机能够通过向例如运行消息传递应用的智能手机的个人设备发送文本消息或其他形式的消息并且接收来自用户的响应消息来与用户交互。

[0198] 用于实现机器学习模型的数据处理装置还能够包括例如专用硬件加速器单元,用于处理机器学习训练或生产的常见和计算密集型部分,即推理工作负载。

[0199] 能够使用例如Tensor Flow框架、Microsoft Cognitive Toolkit框架、Apache Singa框架或Apache MXNet框架的机器学习框架来实现和部署机器学习模型。

[0200] 本说明书中描述的主题的实施例能够在计算系统中实现,该计算系统包括后端组件,例如数据服务器,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的图形用户界面、Web浏览器或应用的客户端计算机,或一个或多个这样的后端、中间件或前端组件的任何组合。系统的组件能够通过例如通信网络的任何形式或介质的数字数据通信互连。通信网络的示例包括局域网 (LAN) 和广域网 (WAN), 例如互联网。

[0201] 计算系统能够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。客户端和服务端的关系由于在相应的计算机上运行并且彼此具有客户端-服务端关系的计算机程序而产生。在一些实施例中,服务端将例如HTML页面的数据传输到用户设备,例如用于向与作为客户端的设备交互的用户显示数据并接收来自该用户的用户输入。在用户设备处生成的数据,例如用户交互的结果,能够在服务端处从设备接收。

[0202] 虽然本说明书包含许多具体的实施细节,但这些细节不应被解释为对任何发明的范围或能够要求保护的范围的限制,而应被解释为对能够特定于特定发明的特定实施例的特征的描述。本说明书中在分开的实施例的上下文中描述的某些特征也能够单个实施例中组合实施。相反,在单个实施例的上下文中描述的各种特征也能够多个实施例中分开地或以任何合适的子组合来实现。此外,尽管特征能够在上面描述为以某些组合起作用并且甚至最初如此要求保护,但是在某些情况下,能够从所要求保护的组合中删除来自所要求保护的组合的一个或多个特征,并且所要求保护的组合可以针对子组合或子组合的变体。

[0203] 类似地,虽然以特定次序在图中描绘并在权利要求中叙述操作,但不应理解为要求以所示的特定次序或顺序次序执行这样的操作,或者要求执行所有所示的操作以实现期望的结果。在某些情况下,多任务和并行处理能够是有利的。此外,上述实施例中各种系统模块和组件的分离不应理解为要求在所有实施例中都进行这样的分离,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分组成多个软件产品。

[0204] 已经描述了主题的具体实施例。其他实施例在以下权利要求的范围内。例如,权利要求中所述的动作能够以不同的次序执行,并且仍然实现理想的结果。作为一个示例,附图中描绘的过程不一定需要所示的特定次序或顺序次序来实现期望的结果。在某些情况下,多任务和并行处理能够是有利的。

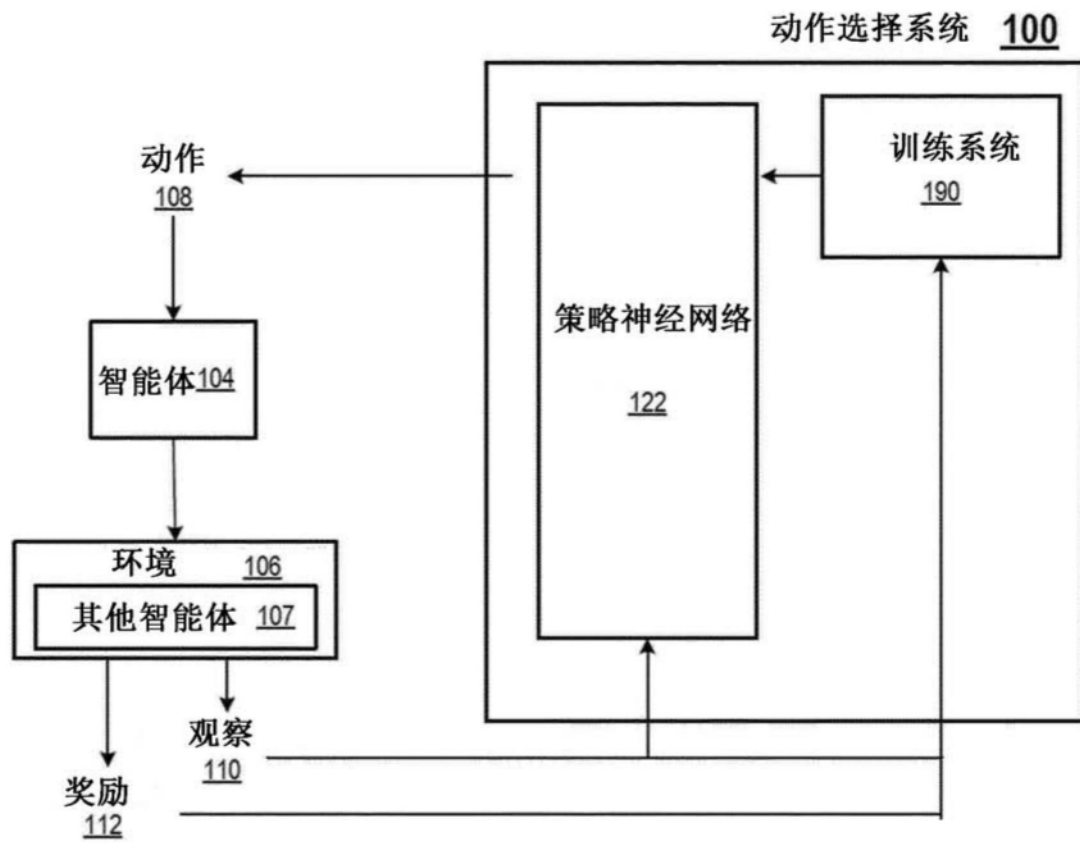


图1

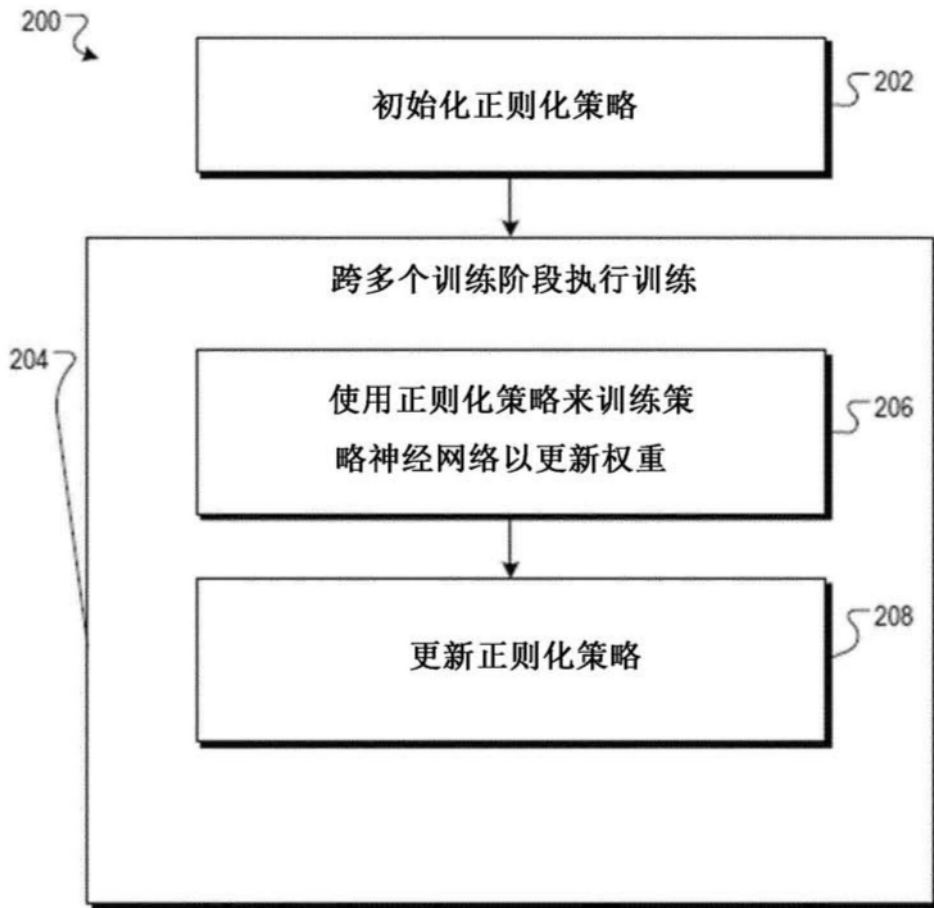


图2

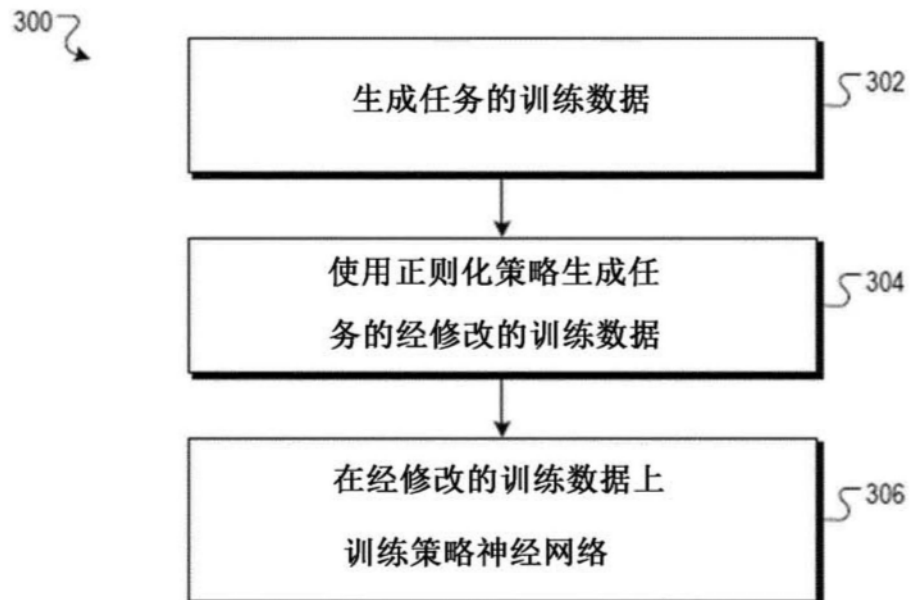


图3

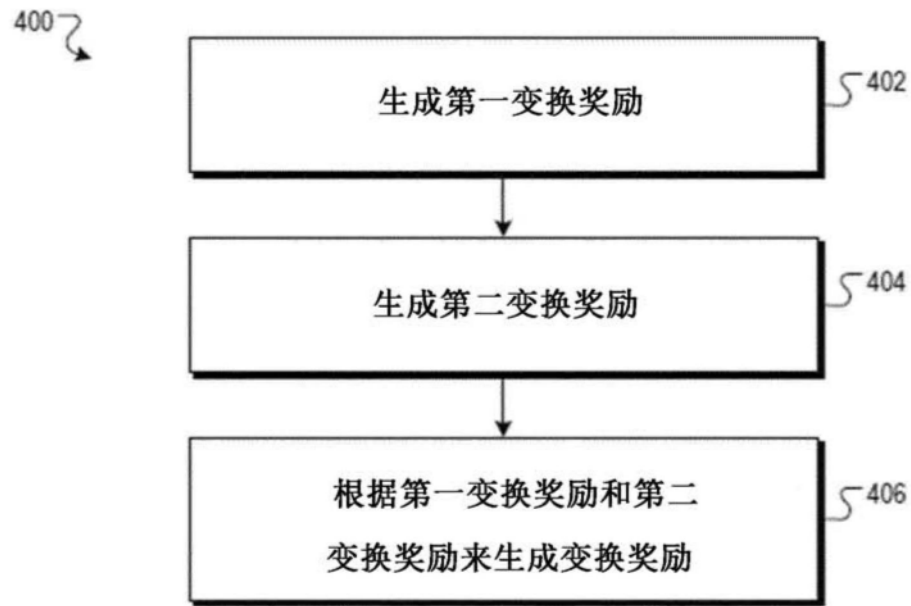


图4

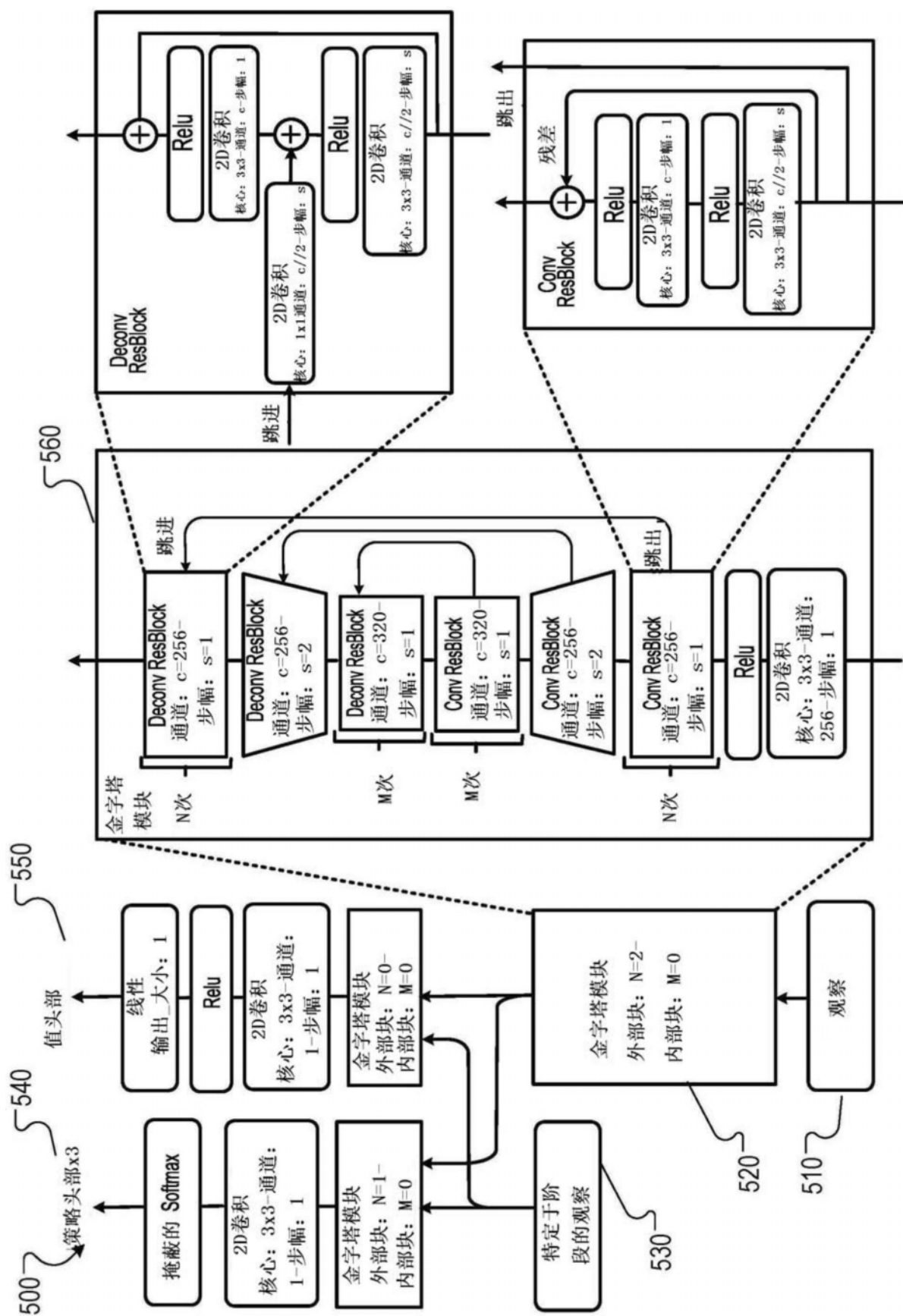


图5

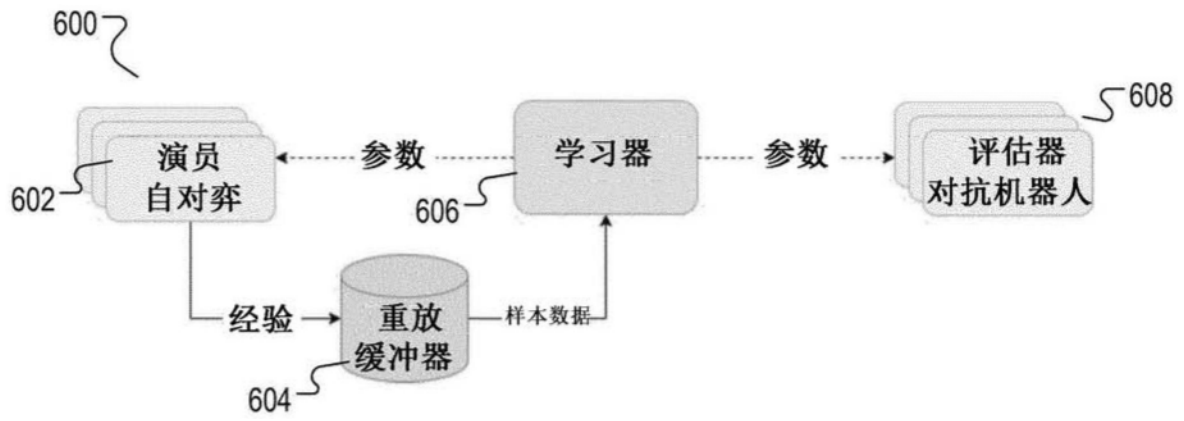


图6A

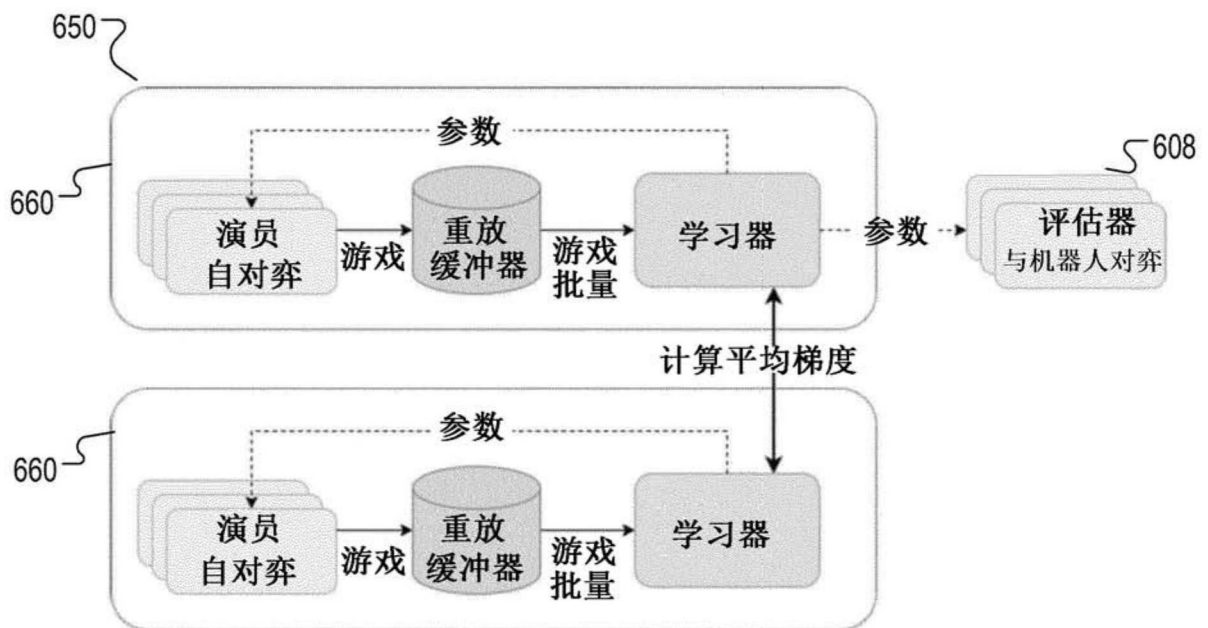


图6B