



(12) 发明专利申请

(10) 申请公布号 CN 117730329 A

(43) 申请公布日 2024. 03. 19

(21) 申请号 202280052793.4

(22) 申请日 2022.07.27

(30) 优先权数据

63/226,124 2021.07.27 US

(85) PCT国际申请进入国家阶段日

2024.01.26

(86) PCT国际申请的申请数据

PCT/EP2022/071137 2022.07.27

(87) PCT国际申请的公布数据

W02023/006848 EN 2023.02.02

(71) 申请人 渊慧科技有限公司

地址 英国伦敦

(72) 发明人 麦克斯维尔·埃利奥特·耶德贝里

沃伊切赫·恰尔内茨基

(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 3/08 (2023.01)

G06N 3/006 (2023.01)

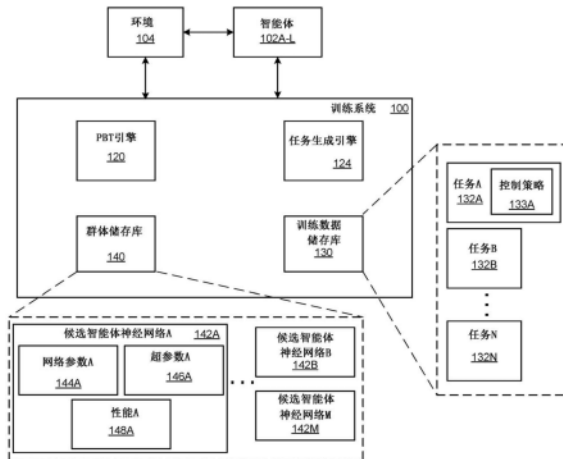
权利要求书5页 说明书23页 附图8页

(54) 发明名称

通过开放式学习来训练智能体神经网络

(57) 摘要

用于训练智能体神经网络以用于控制智能体执行多个任务的方法、系统和装置,包括编码在计算机存储介质上的计算机程序。方法之一包括维护指定一个或多个候选智能体神经网络的群体的群体数据;以及在一个或多个任务的相应集合上训练每个候选智能体神经网络以更新群体数据中的候选智能体神经网络的参数的参数值,训练包括针对每个候选智能体神经网络:获得识别候选任务的数据;获得指定用于候选任务的控制策略的数据;确定是否在候选任务上训练候选智能体神经网络;以及响应于确定在候选任务上训练候选智能体神经网络,在候选任务上训练候选智能体神经网络。



1. 一种用于训练智能体神经网络以用于控制智能体执行多个任务的计算机实现的方法,所述方法包括:

维护指定一个或多个候选智能体神经网络的群体的群体数据,所述群体数据针对所述群体中的每个候选智能体神经网络指定 (i) 所述候选智能体神经网络的参数的相应参数值和 (ii) 用于训练所述候选智能体神经网络的相应超参数值;以及

在一个或多个任务的相应集合上训练每个候选智能体神经网络以更新所述群体数据中的所述候选智能体神经网络的所述参数的所述参数值,所述训练包括,针对每个候选智能体神经网络:

获得识别用于训练所述候选智能体神经网络的候选任务的数据;

获得指定用于所述候选任务的控制策略的数据;

基于 (i) 所述候选智能体神经网络在所述候选任务上的性能、(ii) 用于所述候选任务的所述控制策略在所述候选任务上的性能以及 (iii) 所述候选智能体神经网络的所述超参数值中的一个或多个超参数值,来确定是否在所述候选任务上训练所述候选智能体神经网络;以及

响应于确定在所述候选任务上训练所述候选智能体神经网络,在所述候选任务上训练所述候选智能体神经网络以更新所述智能体神经网络的所述参数的所述参数值。

2. 根据权利要求1所述的方法,其中,在所述候选任务上训练所述候选智能体神经网络包括通过强化学习在所述候选任务上训练所述候选智能体神经网络。

3. 根据任一前述权利要求所述的方法,其中,用于所述候选任务的所述控制策略是均匀随机动作策略,所述均匀随机动作策略在控制所述智能体执行所述候选任务时采取从动作集合中随机均匀选择的动作。

4. 根据权利要求1或2所述的方法,其中,用于所述候选任务的所述控制策略是单个任务策略,所述单个任务策略在控制所述智能体执行所述候选任务时使用仅在所述候选任务上已被训练的单个候选智能体神经网络。

5. 根据权利要求1或2中任一项所述的方法,其中,用于所述候选任务的所述控制策略是在控制所述智能体执行所述候选任务时使用所述候选智能体神经网络的实例但具有来自所述候选智能体神经网络的训练期间的较早点的所述候选智能体神经网络的所述参数的历史参数值的策略。

6. 根据任一前述权利要求所述的方法,其中,基于 (i) 所述候选智能体神经网络在所述候选任务上的性能、(ii) 用于所述候选任务的所述控制策略在所述候选任务上的性能以及 (iii) 所述候选智能体神经网络的所述超参数值中的一个或多个超参数值,来确定是否在所述候选任务上训练所述候选智能体神经网络包括:

针对所述候选任务的多个任务回合中的每个任务回合,确定通过控制所述智能体使用所述候选智能体神经网络执行所述候选任务的所述任务回合而接收到的相应候选智能体回报,以及确定通过控制所述智能体使用所述控制策略执行所述候选任务的所述任务回合而接收到的相应控制策略回报。

7. 根据权利要求6所述的方法,其中,基于 (i) 所述候选智能体神经网络在所述候选任务上的性能、(ii) 用于所述候选任务的所述控制策略在所述候选任务上的性能以及 (iii) 所述候选智能体神经网络的所述超参数值中的一个或多个超参数值,来确定是否在所述候

选任务上训练所述候选智能体神经网络包括：

仅当由所述相应候选智能体回报所测量的所述候选智能体神经网络的性能不超过由所述候选智能体神经网络的所述超参数值定义的阈值性能水平时，确定在所述候选任务上训练所述候选智能体神经网络。

8. 根据权利要求7所述的方法，其中，仅当由所述相应候选智能体回报所测量的所述候选智能体神经网络的性能不超过由所述候选智能体神经网络的所述超参数值定义的阈值性能水平时，确定在所述候选任务上训练所述候选智能体神经网络包括：

仅当所述相应候选智能体回报超过第一指定值的一部分任务回合小于第二指定值时，确定在所述候选任务上训练所述候选智能体神经网络，其中：

(i) 所述第一指定值、(ii) 所述第二指定值或 (iii) 两者都是所述候选智能体神经网络的超参数值。

9. 根据权利要求6-8中任一项所述的方法，其中，基于 (i) 所述候选智能体神经网络在所述候选任务上的性能、(ii) 用于所述候选任务的所述控制策略在所述候选任务上的性能以及 (iii) 所述候选智能体神经网络的所述超参数值中的一个或多个超参数值，来确定是否在所述候选任务上训练所述候选智能体神经网络包括：

仅当由所述相应候选智能体回报所测量的所述候选智能体神经网络的性能超过由所述相应控制策略回报所测量的所述控制策略的性能多于由所述候选智能体神经网络的所述超参数值定义的阈值性能水平时，确定在所述候选任务上训练所述候选智能体神经网络。

10. 根据权利要求9所述的方法，其中，仅当由所述相应候选智能体回报所测量的所述候选智能体神经网络的性能超过由所述相应控制策略回报所测量的所述控制策略的性能多于由所述候选智能体神经网络的所述超参数值定义的阈值性能水平时，确定在所述候选任务上训练所述候选智能体神经网络包括：

仅当所述相应候选智能体回报超过所述相应控制策略回报至少第三指定值的一部分任务回合大于第四指定值时，确定在所述候选任务上训练所述候选智能体神经网络，其中：

(i) 所述第三指定值、(ii) 所述第四指定值或 (iii) 两者都是所述候选智能体神经网络的超参数值。

11. 根据权利要求6至10中任一项所述的方法，其中，基于 (i) 所述候选智能体神经网络在所述候选任务上的性能、(ii) 用于所述候选任务的所述控制策略在所述候选任务上的性能以及 (iii) 所述候选智能体神经网络的所述超参数值中的一个或多个超参数值，来确定是否在所述候选任务上训练所述候选智能体神经网络包括：

仅当由所述相应控制策略回报所测量的所述控制策略的性能低于由所述候选智能体神经网络的所述超参数值定义的阈值性能水平时，确定在所述候选任务上训练所述候选智能体神经网络。

12. 根据权利要求11所述的方法，其中仅当由所述相应控制策略回报所测量的所述控制策略的性能低于由所述候选智能体神经网络的所述超参数值定义的阈值性能水平时，确定在所述候选任务上训练所述候选智能体神经网络包括：

仅当所述相应控制策略回报的平均值小于第五指定值时，确定在所述候选任务上训练所述候选智能体神经网络，其中：

所述第五指定值是所述候选智能体神经网络的超参数值。

13. 根据任一前述权利要求所述的方法, 进一步包括:

在一个或多个任务的相应集合上训练每个候选智能体神经网络之后:

调整所述群体中的所述候选智能体神经网络中的一个或多个候选智能体神经网络的超参数值。

14. 根据权利要求13所述的方法, 其中, 一个或多个候选智能体神经网络的所述群体包括多个神经网络, 并且其中, 调整所述群体中的所述候选智能体神经网络中的一个或多个候选智能体神经网络的所述超参数值包括:

计算所述多个候选智能体神经网络中的每个候选智能体神经网络的相应适合度测量, 所述相应适合度测量跨多个验证任务测量所述多个候选智能体神经网络中的每个候选智能体神经网络的相应性能; 以及

将基于群体的训练技术应用于所述相应适合度测量, 以调整所述候选智能体神经网络中的一个或多个候选智能体神经网络的所述超参数值。

15. 根据权利要求14所述的方法, 其中, 对于所述候选智能体神经网络中的一个或多个候选智能体神经网络中的每个候选智能体神经网络, 所述基于群体的训练技术还将所述候选智能体神经网络的相应参数值设置为等于所述群体数据中的另一候选智能体神经网络的相应参数值。

16. 根据权利要求14或15中任一项所述的方法, 其中, 计算所述相应适合度测量包括: 确定所述候选智能体神经网络在所述多个验证任务中的每个验证任务上的相应归一化百分位度量。

17. 根据权利要求14至16中任一项所述的方法, 其中, 应用所述基于群体的训练技术包括:

确定第一候选智能体神经网络的相应适合度测量帕累托占优第二候选智能体神经网络的相应适合度测量; 以及

作为响应, 将所述第二候选智能体神经网络的所述超参数值调整为等于所述第一候选智能体神经网络的所述超参数值的变异版本。

18. 根据任一前述权利要求所述的方法, 其中, 在所述智能体神经网络的训练期间的在训练代序列的当前训练代期间发生在一个或多个任务的相应集合上的所述候选智能体神经网络中的每个候选智能体神经网络的训练, 并且其中, 在所述候选任务上训练所述候选智能体神经网络以更新所述智能体神经网络的所述参数的所述参数值包括:

在所述候选任务上训练所述候选智能体神经网络以优化预期回报, 同时在所述序列中紧接在所述当前训练代之前的先前训练代结束时从所述群体中的最佳执行候选智能体神经网络中蒸馏。

19. 根据任一前述权利要求所述的方法, 其中, 在所述先前训练代期间, 所述群体中的所述候选智能体神经网络是在自奖励游戏目标上训练的。

20. 一种由一个或多个计算机执行并且用于控制与环境交互的智能体以使所述智能体在所述环境中执行任务的方法, 所述方法包括:

接收表征所述环境的当前状态的观测;

接收表示要被满足的目标的目标数据, 以便在所述环境中执行所述任务, 其中, 所述目

标被表示为相应谓词集合上的选项集合；

使用智能体神经网络处理所述观测和所述目标数据以生成动作选择输出，包括：

使用状态编码器神经网络处理所述观测以生成表示所述环境的当前状态的当前隐藏状态；

从所述目标数据生成所述目标的目标嵌入和所述选项中的每个选项的相应选项嵌入；

使用注意力神经网络处理包括所述当前隐藏状态和所述目标嵌入的输入以生成目标注意力隐藏状态；

对于所述选项中的每个选项，使用所述注意力神经网络处理包括所述当前隐藏状态和所述选项的所述相应选项嵌入的输入，以生成所述选项的相应选项注意力隐藏状态；

使用值神经网络头处理所述目标注意力隐藏状态以生成所述目标的目标值估计，所述目标值估计表示如果所述智能体尝试从所述当前状态开始满足所述目标则将实现的估计回报；

对于所述选项中的每个选项，使用所述值神经网络头处理所述选项的所述相应选项注意力隐藏状态，以生成所述选项的相应选项值估计，所述相应选项值估计表示如果所述智能体尝试从所述当前状态开始满足所述选项则将实现的估计回报；

根据所述目标值估计和所述选项的所述相应选项值估计，生成所述目标和所述选项中的每个选项的相应权重；

根据所述相应权重组合所述目标注意力隐藏状态和所述相应选项注意力隐藏状态以生成组合隐藏状态；并且

使用策略神经网络头处理所述组合隐藏状态以生成所述动作选择输出；

使用所述动作选择输出来选择要由所述智能体执行的动作；并且

使所述智能体执行所选择的动作。

21. 根据权利要求20所述的方法，其中，所述状态编码器神经网络是递归神经网络。

22. 根据权利要求20或权利要求21所述的方法，其中，使用智能体神经网络处理所述观测和所述目标数据以生成动作选择输出进一步包括：

使用谓词预测器神经网络处理所述目标数据和所述当前隐藏状态以生成谓词预测，其中，包括所述当前隐藏状态和所述目标嵌入的所述输入进一步包括所述谓词预测，并且对于每个选项，包括所述当前隐藏状态和所述相应选项嵌入的所述输入进一步包括所述谓词预测。

23. 根据权利要求20-22中任一项所述的方法，其中，根据所述目标值估计和所述选项的所述相应选项值估计，生成所述目标和所述选项中的每个选项的相应权重包括：对所述目标值估计和所述选项的所述相应选项值估计应用softmax。

24. 根据权利要求20-22中任一项所述的方法，其中，根据所述目标值估计和所述选项的所述相应选项值估计，生成所述目标和所述选项中的每个选项的相应权重包括：将权重1指派给所述目标值估计和所述选项的所述相应选项值估计中的最高值估计，并且将权重0指派给目标值估计和所述选项的所述相应选项值估计的所有其他值估计。

25. 根据权利要求20-24中任一项所述的方法，其中，所述动作选择输出定义能够由所述智能体执行的动作集合上的概率分布。

26. 根据权利要求20-24中任一项所述的方法，其中，所述动作选择输出包括能够由所

述智能体执行的动作集合中的每个动作的相应Q值。

27. 根据权利要求20-24中任一项所述的方法, 其中, 所述动作选择输出是来自连续动作空间的动作。

28. 一种系统, 包括:

一个或多个计算机; 以及

通信地耦合到所述一个或多个计算机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得所述一个或多个计算机执行根据权利要求1-27中任一项所述的相应方法的操作。

29. 一种或多种存储指令的非暂时性计算机存储介质, 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1-27中任一项所述的相应方法的操作。

30. 根据任一前述权利要求所述的方法、系统或计算机存储介质, 其中, 所述智能体神经网络用于在现实世界环境中控制所述智能体, 并且被配置为处理与所述现实世界环境的状态有关的观测以生成与要由所述智能体在所述现实世界环境中执行的动作有关的动作选择输出。

31. 根据权利要求30所述的方法、系统或计算机存储介质, 其中, 所述智能体是机械智能体, 并且其中, 所述智能体神经网络被用于控制所述现实世界环境中的所述机械智能体以执行所述任务或所述任务之一。

通过开放式学习来训练智能体神经网络

[0001] 相关申请的交叉引用

[0002] 本申请要求于2021年7月27日提交的美国临时申请63/226,124的优先权。现有申请的公开内容被认为是本申请的公开内容的一部分并且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来处理数据。

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,并且基于所接收的输入和模型的值来生成输出。

[0005] 一些机器学习模型是采用多层模型来生成针对接收到的输入的输出的深度模型。例如,深度神经网络是包括输出层和一个或多个隐藏层的深度机器学习模型,每个隐藏层将非线性变换应用于接收到的输入以生成输出。

发明内容

[0006] 本说明书总体上描述了一种被实现为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,该系统训练用于选择要由与环境交互的智能体执行的动作的智能体神经网络。具体地,系统训练智能体神经网络,使得智能体神经网络可以用于控制智能体执行多个任务中的任何任务。每个任务可以包括例如导航到环境中的指定位置、识别环境中的特定对象、以指定方式操纵特定对象等中的一个或多个。

[0007] 在一个方面,描述了一种用于训练智能体神经网络以用于控制智能体执行多个任务的计算机实现的方法。该方法包括维护指定一个或多个候选智能体神经网络的群体的群体数据,该群体数据针对群体中的每个候选智能体神经网络指定(i)候选智能体神经网络的参数的相应参数值和(ii)用于训练候选智能体神经网络的相应超参数值。该方法涉及在一个或多个任务——通常是多个任务——的相应集合上训练每个候选智能体神经网络,以更新群体数据中的候选智能体神经网络的参数的参数值。在实施方式中,训练包括,对于每个候选智能体神经网络:获得识别用于训练候选智能体神经网络的候选任务(例如,多个任务中的一个)的数据;获得指定用于候选任务的控制策略的数据;以及基于(i)候选智能体神经网络在候选任务上的性能、(ii)用于候选任务的控制策略在候选任务上的性能以及(iii)候选智能体神经网络的超参数值中的一个或多个超参数值,确定是否在候选任务上训练候选智能体神经网络。响应于确定在候选任务上训练候选智能体神经网络,该方法包括例如使用强化学习来在候选任务上训练候选智能体神经网络以更新智能体神经网络的参数的参数值。可以在每次获得识别候选任务的数据时,即使用多个任务中的一个作为候选任务,重复执行训练。

[0008] 在这样的方法的实施方式中,响应于智能体的性能而动态地生成由智能体消耗的

训练任务。效果是任务分布贯穿训练被改变,并且其本身可以被优化以改善智能体的性能。更具体地,如上所列出的,基于三个因素的组合来确定是否训练候选智能体神经网络。例如,一般而言,只有当性能(显著地)优于控制策略时,才应当在任务上训练智能体,使得智能体执行有意义的动作,即影响回报的动作。智能体的性能通常可以例如根据来自任务的回报来确定。控制策略可以是(用于控制智能体执行候选任务的)任何动作选择策略,例如均匀随机动作策略或智能体的过去策略。

[0009] 经训练的智能体神经网络可以用于在一个或多个环境中控制智能体以执行多个任务。环境可以是现实世界环境,但是训练中的一些或全部可以是在现实世界环境的模拟中。在一些实施方式中,特别是在一些或全部训练处于模拟中的情况下,用于训练候选智能体神经网络的“提议”任务是程序地——即,基于任务参数自动地——生成的,并且这些任务然后可以用于确定是否训练候选智能体神经网络。

[0010] 上述方法可以用于获得包括一个或多个候选智能体神经网络的群体。当群体包括多于一个候选智能体神经网络时,训练多个候选智能体神经网络的群体。然后,可以选择一个或多个作为经训练的智能体神经网络。在多个候选智能体神经网络的群体中,这些神经网络可以但不必具有彼此相同的架构;通常,它们具有不同的相应参数值,例如权重,并且可以具有不同的相应超参数值。

[0011] 在一些实施方式中,候选智能体神经网络的群体用于基于群体的训练。也就是说,训练候选智能体神经网络的群体,并且为智能体中的每个智能体确定适合度测量,例如以比较两个(或更多个)智能体。这可以例如通过使用适合度函数评估智能体中的每个智能体来完成;存在许多合适的适合度测量;通常可以根据来自任务的回报来确定适合度。

[0012] 如上所述,每个候选智能体神经网络由相应的超参数值控制。可以使用适合度测量来修改候选智能体神经网络的群体,以为了改善群体的性能为目的,特别是通过调整候选智能体神经网络中的一个或多个候选智能体神经网络的超参数值。该调整可以以许多方式执行。作为一个示例,超参数值可以被随机扰动。

[0013] 作为另一示例,如果确定第一候选智能体神经网络(通过多个任务)的适合度测量帕累托(Pareto)占优(dominate)第二候选智能体神经网络(通过多个任务)的适合度测量,则可以将第二候选智能体神经网络的超参数值调整为等于第一候选智能体神经网络的超参数值的变异(修改)版本。如果适合度测量至少在多个任务上一样好并且对于任务中的至少一个任务而言更好,则可以认为达成了帕累托占优。任务可以是候选任务或其他任务,例如验证任务。

[0014] 在一些实施方式中,较不优选的候选智能体神经网络——即(例如,当比较两个智能体时)具有较小适合度测量的候选智能体神经网络——可以将其相应的参数值(例如,权重)以及可选地还有其超参数值设置为用于另一候选智能体神经网络——例如(例如,当比较两个智能体时)具有较大适合度测量的候选智能体神经网络——的相应参数值以及可选地超参数值。

[0015] 如上所提及,智能体的适合度或性能通常可以根据来自任务的回报来确定。然而,当存在许多候选任务时,回报可以在宽范围内变化。因此,在一些实施方式中,计算任务的适合度(或性能)测量包括确定相应的归一化百分位度量,例如通过确定由智能体在任务上达成的第k个——例如第50个——百分位分数,其中分数使用(除以)归一化常数来被归一

化,该归一化常数是由达成任务上的最佳分数的一个或多个智能体(候选智能体神经网络)的混合或集合达成的分数。这里,分数基于智能体针对任务达成的奖励或回报。该方法避免了将性能与例如用于任务的最优策略进行比较的需要,在确定最佳策略时有用但不直接。

[0016] 可以在候选任务上训练候选智能体神经网络以优化来自任务的预期回报,同时从群体中的最佳执行候选智能体神经网络中蒸馏(distill),例如,如在序列中紧接在当前训练代(generation)之前的先前训练代结束时确定的。这样的蒸馏时的训练可以包括使用蒸馏损失——例如基于这些输出之间的差异的损失——来训练,该蒸馏损失鼓励朝向最佳执行的候选智能体神经网络的动作选择输出训练候选智能体神经网络的动作选择输出。

[0017] 训练可以涉及自奖励游戏目标,即其中顺序地训练候选智能体神经网络以达成目标 g 并且达成 $\text{not}(g)$ (其中 $\text{not}(g)$ 是目标的否定)。这可以鼓励探索;可以看作智能体与自身竞争。目标可以被表示为相应谓词集合上的选项集合(析取),例如表示每个选项的一个或多个必要条件。目标可以但不必被表示为自然语言。

[0018] 通常,智能体神经网络(以及候选智能体神经网络中的每一个)被配置用于在一个或多个环境中控制智能体以执行多个任务。因此,通常,智能体神经网络(以及候选智能体神经网络中的每一个)被配置为接收和处理表征环境的当前状态的观测,并且在一些实施方式中,还可以被配置为接收和处理识别要执行的(特定)任务的数据,诸如表示要满足的目标以便在环境中执行任务的目标数据(可替代地,例如,可以从环境推断任务)。

[0019] 通常,智能体神经网络(以及候选智能体神经网络中的每一个)被配置为生成动作选择输出,该动作选择输出表征要由智能体响应于观测而执行的动作,以便执行一个或多个任务。在实施方式中,经训练的智能体神经网络用于选择要由智能体在现实世界环境中执行的动作,并且所选择的动作涉及要由智能体在现实世界环境中执行的动作(并且观测涉及现实世界环境的观测)。作为一个示例,智能体可以包括机械智能体,诸如机器人或自主或半自主车辆;稍后给出智能体的其他示例。

[0020] 本说明书中描述的主题可以在特定实施方式中实现,以便实现以下优点中的一个或多个。

[0021] 所描述的技术可以实现开放式强化学习训练过程,在该开放式强化学习训练过程期间,训练任务分布和训练目标以使得智能体神经网络的多个实例很少停止学习的方式被动态地适配,以连续并且有效地训练网络实例,以使得由经训练的网络实例中的一个或多个经训练的网络实例控制的智能体跨不同任务和环境的规模空间达成稳健并且大体上有能力的性能。所描述的用于训练智能体神经网络的技术普遍适用于其中可以部署智能体以执行任何类型的技术上具有挑战性的任务的任何类型的复杂环境。

[0022] 因此,使用所描述的技术导致大体上有能力的智能体神经网络,该智能体神经网络不仅可以胜过现有技术,而且可以胜过一系列智能体控制任务上的人类级性能,同时附加地可以例如通过微调而推广和容易地适应于新任务,包括自然地区别于网络可能已经在其上训练的现有任务的新颖任务。除了实现改进的智能体性能之外,通过跨大量多个任务——例如,无限多任务连续体——训练智能体神经网络实例的群体,与需要针对每个单个新任务或新环境从头开始训练一个神经网络模型的常规方法相比,该训练过程还消耗更少的计算资源,例如,存储器和处理功率。

[0023] 在附图和下面的描述中阐述了本说明书的主题的一个或多个实施方式的细节。根

据说明书、附图和权利要求,主题的其他特征、方面和优点将变得显而易见。

附图说明

[0024] 图1示出了示例训练系统。

[0025] 图2是用于训练候选智能体神经网络的群体的示例过程的流程图。

[0026] 图3示出了训练候选智能体神经网络的群体的示例图示。

[0027] 图4A示出了示例强化学习智能体控制系统,并且图4B示出了智能体神经网络的示例实施方式。

[0028] 图5是用于控制与环境交互的智能体的示例过程的流程图。

[0029] 图6是用于使用智能体神经网络来生成动作选择输出的示例过程的流程图。

[0030] 图7示出了确定归一化百分位度量的示例图示。

[0031] 各个附图中相同的附图标记和标记指示相同的元件。

具体实施方式

[0032] 图1示出了示例训练系统100。训练系统100是被实现为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,在其中可以实现下面描述的系统、组件和技术。

[0033] 训练系统100是实现开放式强化学习训练过程以训练智能体神经网络的多个实例的系统,每个实例可以用于选择要由环境104中的智能体——例如智能体102A——执行的动作,以便控制智能体执行一系列机器学习任务。在一些情况下,机器学习任务是可以由智能体本身执行的单智能体任务,而在其他情况下,机器学习任务是要求智能体与环境104中的例如智能体102B-L的一个或多个其他强化学习智能体交互——例如竞争或协调——的多智能体任务。也就是说,训练系统100获得(即,生成或接收)观测,其中每个观测表征环境104的相应状态,并且响应于每个观测,从预定动作集合中选择要由强化学习智能体102A响应于观测而执行的动作。响应于由智能体102A执行的动作中的一些或全部,训练系统100获得奖励。每个奖励是由于智能体102A执行动作而从环境104接收的数值,特别地,奖励将取决于环境104由于智能体102A执行动作而转换到的状态而不同。在一些情况下,预定动作集合可以定义离散动作空间,而在其他情况下,预定动作集合可以可替代地定义连续动作空间,即,单独动作中的所有动作值是从可能值的连续范围中选择的,或者可以进一步可替代地定义混合动作空间,即,单独动作中的一个或多个动作值是从可能值的连续范围中选择的。

[0034] 具体地,通过实现所描述的开放式强化学习训练过程,可以训练单个但大体上有能力的智能体神经网络,该智能体神经网络可以用于控制智能体以不仅实现或甚至超过现有技术,还可以在不同环境中跨数千或数百万任务实现人类级性能。

[0035] 在一些实施方式中,环境是现实世界环境,智能体是与现实世界环境交互的机械智能体,例如,机器人或在环境中操作或通过环境导航的自主或半自主陆地、空中或海洋交通工具,并且动作是由现实世界环境中的机械智能体采取以执行任务的动作。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,以定位环境中的感兴趣对象或将感兴趣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。

[0036] 在这些实施方式中,观测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测,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,观测可以包括表征机器人的当前状态的数据,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,重力补偿的扭矩反馈——以及由机器人保持的物项(item)的全局或相对姿态。在机器人或其他机械智能体或车辆的情况下,观测可以类似地包括智能体的一个或多个部分的位置、线性或角速度、力、扭矩或加速度以及全局或相对姿态中的一个或多个。观测可以在1个、2个或3个维度中定义,并且可以是绝对和/或相对观测。观测还可以包括例如感测到的电子信号,诸如马达电流或温度信号;和/或例如来自相机或LIDAR传感器的图像或视频数据,例如来自智能体的传感器的数据或来自与环境中的智能体分开定位的传感器的数据。

[0037] 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,例如,机器人的关节的扭矩,或更高级别的控制命令;或可以是用于控制自主或半自主陆地、空中、海运交通工具的控制信号,例如,到控制表面或其他控制元件——例如交通工具的转向控制元件——的扭矩,或更高级别的控制命令。控制信号可以包括例如用于机器人的一个或多个关节或另一机械智能体的部分的位置、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,诸如马达控制数据,或更一般地用于控制环境内的一个或多个电子设备的数据,该控制对环境的观测状态具有影响。例如,在自主或半自主陆地或空中或海洋交通工具的情况下,控制信号可以定义控制导航——例如转向——和移动——例如交通工具的制动和/或加速——的动作。

[0038] 在一些实施方式中,环境是上述现实世界环境的模拟,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车辆的模拟,并且强化学习系统可以在模拟上被训练,然后一旦被训练,就在现实世界中使用。

[0039] 在一些实施方式中,环境是用于制造产品——诸如化学、生物或机械产品或食物产品——的现实世界制造环境。如本文所用,“制造”产品还包括精炼起始材料(material)以产生产品,或处理起始材料例如以去除污染物,以生成清洁或再循环的产品。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(substance)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的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物产品。

[0040] 智能体可以包括被配置为控制制造单元或诸如机器人的机器的电子智能体,其操作以制造产品。也就是说,智能体可以包括被配置为控制化学、生物或机械产品的制造的控制系统。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组件在制造单元或机器之间的移动。

[0041] 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组件的任务。作为另一示例,由智能体执行的任务可以包括控制——例如最小化——资源的使用的任务,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。

[0042] 动作可以包括控制动作,以控制用于处理固体或液体材料的机器或制造单元的使用,以制造产品或其中间件或组件,或者控制产品的中间版本或组件在制造环境内——例如,在制造单元或机器之间——的移动。通常,动作可以是对环境的观测状态有影响的任何

动作,例如被配置为调整下面描述的任何感测参数的动作。这些可以包括调整制造单元的物理或化学条件的动作,或者控制机器的机械部件或机器人的关节的移動的动作。动作可以包括在制造单元或机器上施加操作条件的动作,或者导致设置改变以调整、控制或打开或关闭制造单元或机器的操作的动作。

[0043] 奖励或回报可以涉及任务的性能的度量。例如,在要制造产品的任务的情况下,度量可以包括所制造的产品数量、产品的质量、产品的生产速度的度量或对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、材料或其他资源数量的度量。在控制使用资源的任务的情况下,度量可以包括资源使用的任何度量。

[0044] 通常,环境状态的观测可以包括表示装备的电子和/或机械物项的功能的任何电子信号。例如,环境状态的表示可以从由感测制造环境状态的传感器——例如,感测制造单元或机器的状态或配置的传感器,或感测材料在制造单元或机器之间的移动的传感器——进行的观测中导出。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测;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,来自传感器的观测可以包括对机器的一个或多个部分的位置、线性或角速度、力、扭矩或加速度或姿态的观测,例如表征机器或机器人或由机器或机器人保持或处理的物项的当前状态的数据。观测还可以包括例如感测到的电子信号,诸如马达电流或温度信号,或者例如来自相机或LIDAR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的智能体的一部分或与环境中的智能体分开定位。

[0045] 在一些实施方式中,环境是包括多个电子装备物项的服务设施的现实世界环境,诸如服务器群或数据中心,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物项的操作环境的辅助控制装备,例如环境控制装备,诸如温度控制,例如冷却装备,或气流控制或空气调节装备。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,该电子智能体被配置为控制装备物项的操作,或者控制辅助的——例如环境的——控制装备的操作。

[0046] 通常,动作可以是对环境的观测状态有影响的任何动作,例如被配置为调整下面描述的任何感测参数的动作。这些可以包括控制装备物项或辅助控制装备、或者对装备物项或辅助控制装备施加操作条件的动作,例如导致对设置进行改变以调整、控制或打开或关闭装备项或辅助控制装备物项的操作的动作。

[0047] 通常,环境状态的观测可以包括表示设施或设施中的设备的功能的任何电子信号。例如,环境的状态的表示可以从由对设施的物理环境的状态进行感测的任何传感器进行的观测、或者由感测装备的一个或多个物项的状态或感测辅助控制装备的一个或多个物项的状态的任何传感器进行的观测中导出。这些传感器包括被配置为感测电气条件的传感器,电气条件诸如电流、电压、功率或能量;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。

[0048] 奖励或回报可以涉及任务的性能的度量。例如,在控制——例如最小化——资源

的使用的任务——诸如控制电力或水的使用的任务——的情况下,度量可以包括资源的使用的任何度量。

[0049] 在一些实施方式中,环境是发电设施——例如可再生发电设施,诸如太阳能发电场或风电场——的现实世界环境。该任务可以包括控制由设施生成的电力的控制任务,例如,控制电力到配电网的输送,例如,以满足需求或降低电网的元件之间不匹配的风险,或最大化由设施生成的电力。该智能体可以包括电子智能体,该电子智能体被配置为控制由设施生成电力或将生成的电力耦合到电网中。这些动作可以包括控制发电机的电气或机械配置的动作,诸如一个或多个可再生发电元件的电气或机械配置,例如控制风力涡轮机或太阳能电池板或多个电池板或镜子的配置,或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,例如转换的效率或能量输入到电能输出的耦合程度。电控制动作可以例如包括控制所生成的电力的电压、电流、频率或相位中的一个或多个的动作。

[0050] 奖励或回报可以涉及任务的性能的度量。例如,在控制电力到配电网的输送的任务的情况下,度量可以涉及传输的电力的测量,或者涉及发电设施和电网之间的电失配的测量,诸如电压、电流、频率或相位失配,或者涉及发电设施中的电力或能量损失的测量。在使电力到配电网的输送最大化的任务的情况下,度量可以涉及传输到电网的电力或能量的测量,或者涉及发电设施中的电力或能量损失的测量。

[0051] 通常,环境状态的观测可以包括表示发电装备在发电设施中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由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状况的任何传感器进行的观测中导出。这样的传感器可以包括被配置为感测装备的诸如电流、电压、功率或能量的电气状况的传感器;感测物理环境的温度或冷却的传感器;感测流体流动的传感器;或感测装备的物理配置的传感器;以及感测例如来自本地或远程传感器的电网的电气状况的观测的传感器。环境状态的观测还可以包括关于发电装备的操作的未来状况的一个或多个预测,诸如未来风水平或太阳辐照度的预测或电网的未来电气状况的预测。

[0052] 作为另一个示例,环境可以是化学品合成或蛋白质折叠环境,使得每种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,这些动作是用于折叠蛋白质链的可能的折叠动作或用于组装前体化学品/中间体的动作,并且要达成的结果可以包括例如折叠蛋白质,使得蛋白质稳定并且使得其达成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,其执行或控制由系统自动选择的蛋白质折叠动作或化学品合成步骤,而无需人类交互。观测可以包括蛋白质或化学品/中间体/前体的状态的直接或间接观测和/或可以从模拟中导出。

[0053] 以类似的方式,环境可以是药物设计环境,使得每个状态是潜在的制药化学药物(pharmaceutical drug)的相应状态,并且智能体是用于确定制药化学药物的元素和/或制药化学药物的合成途径的计算机系统。药物/合成可以基于从药物的靶标导出的奖励来设计,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。

[0054] 在一些进一步的应用中,环境是现实世界环境,并且智能体管理跨例如移动设备上和/或数据中心中的计算资源的任务分发。在这些实施方式中,动作可以包括将任务指派

给特定计算资源。

[0055] 作为进一步的示例,动作可以包括呈现广告,观测可以包括广告印象或点进计数或点进率,并且奖励可以表征由一个或多个用户采取的物项或内容的先前选择。

[0056] 在一些情况下,观测可以包括由第三方(例如,智能体的操作者)提供给智能体的文本或口头指令。例如,智能体可以是自主车辆,并且自主车辆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)。

[0057] 作为另一示例,环境可以是电气、机械或机电设计环境,例如在其中模拟电气、机械或机电实体的设计的环境。模拟环境可以是实体旨在在其中工作的现实世界环境的模拟。任务可以是设计实体。观测可以包括表征实体的观测,即实体的机械形状或电气、机械或机电配置的观测,或实体的参数或属性的观测。动作可以包括修改实体的动作,例如修改一个或多个观测的动作。奖励或回报可以包括实体的设计的性能的一个或多个度量。例如,奖励或回报可以涉及实体的一个或多个物理特性,诸如重量或强度,或者涉及实体的一个或多个电特性,诸如执行实体被设计用于的特定功能的效率的度量。设计过程可以包括例如以用于制造实体的计算机可执行指令的形式输出用于制造的设计。该过程可以包括根据设计制作实体。因此,可以例如通过强化学习来优化实体的设计,然后例如作为计算机可执行指令来优化用于制造实体的优化设计输出;然后可以制造具有优化设计的实体。

[0058] 如前所述,环境可以是模拟环境。通常,在模拟环境的情况下,观测可以包括一个或多个先前描述的观测或观测类型的模拟版本,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,并且智能体可以通过运动模拟导航的模拟交通工具。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控制输入。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。

[0059] 模拟环境可以是特定现实世界环境和智能体的模拟。例如,系统可以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之后,可以被部署用于控制作为模拟的主体的特定现实世界环境中的现实世界智能体。这可以避免对现实世界环境或现实世界智能体的不必要的磨损和损坏,并且可以允许在现实世界环境中很少发生或难以重新创建或重新创建不安全的情况下训练和评估控制神经网络。例如,可以使用特定现实世界环境的模拟中的机械智能体的模拟来部分地训练系统,并且然后部署系统以控制特定现实世界环境中的真实机械智能体。因此,在这样的情况下,模拟环境的观测与现实世界环境有关,并且模拟环境中的所选择的动作与要由现实世界环境中的机械智能体执行的动作有关。

[0060] 可选地,在任何上述实施方式中,在任何给定时间步处的观测可以包括来自先前时间步的数据,该数据可以有益于表征环境,例如,在先前时间步处执行的动作、在先前时间步处接收的奖励或两者。

[0061] 通常,智能体神经网络可以具有使其能够执行其描述的功能的任何适当的神经网络架构,例如,处理识别要执行的任务的数据和观测以生成表征要由智能体响应于观测而执行的动作的动作选择输出。例如,智能体神经网络可以包括任何适当类型(例如,全连接层、卷积层、注意力层、变换器层等)并且以任何适当配置(例如,作为线性层序列,具有或不具有残差连接)连接的任何适当数量的层(例如,5层、10层或25层)。

[0062] 在一个示例中,动作选择输出可以包括可以由智能体执行的可能动作集合中的每个动作的相应数字概率值。系统可以选择要由智能体执行的动作,例如,通过根据动作的概率值对动作进行采样,或者通过选择具有最高概率值的动作。

[0063] 在另一示例中,动作选择输出可以直接定义要由智能体执行的动作,例如,通过定义应当施加到机器人智能体的关节的扭矩值。

[0064] 在另一示例中,动作选择输出可以包括可以由智能体执行的可能动作集合中的每个动作的相应Q值。系统可以处理Q值(例如,使用softmax函数)以生成每个可能动作的相应概率值,其可以用于选择要由智能体执行的动作(如前所述)。系统还可以选择具有最高Q值的动作作为要由智能体执行的动作。

[0065] 动作的Q值是“回报”的估计,其将由智能体执行对当前观测的动作并且此后根据智能体神经网络的参数的当前值选择由智能体102执行的未来动作而产生。

[0066] 回报是指智能体接收的“奖励”的累积测量,例如奖励的时间折扣总和。智能体可以在每个时间步处接收相应的奖励,其中,奖励由标量数值指定并且表征例如智能体完成所指派的任务的进度。

[0067] 在一些情况下,系统可以根据探索策略选择要由智能体执行的动作。例如,探索策略可以是 ϵ 贪婪探索策略,其中,系统根据具有概率 $1-\epsilon$ 的动作选择输出来选择要由智能体执行的动作,并且随机地选择具有概率 ϵ 的动作。在该示例中, ϵ 是0和1之间的标量值。

[0068] 将参考图4进一步描述智能体神经网络的架构的一个特定示例。

[0069] 训练系统100维护存储指定一个或多个候选智能体神经网络142A-M的群体的群体数据的群体储存库140。群体储存库140被实现为一个或多个物理位置中的一个或多个逻辑存储设备,或者被实现为在一个或多个物理位置中的一个或多个存储设备中分配的逻辑存储空间。在训练期间的任何给定时间,群体储存库140存储指定候选智能体神经网络142A-M的当前群体的数据。

[0070] 具体地,群体储存库110针对当前群体中的每个候选智能体神经网络142A-M存储定义相应候选智能体神经网络的维护值的集合。维护值的集合包括网络参数、超参数,并且在实施方式中,维护在不同任务上的每个候选智能体神经网络142A-M的性能测量(或简称“性能”)也是方便的。例如,对于候选智能体神经网络A 142A,维护值的集合包括网络参数A 144A、超参数A 146A和性能A 148A(其可以包括每个不同任务的相应性能)。当在群体中存在多个候选智能体神经网络时,每个候选智能体神经网络通常将具有相同的架构,但是具有不同的相应参数值,并且在一些情况下具有与群体中的其他候选智能体神经网络不同的相应超参数值。

[0071] 候选智能体神经网络的超参数是影响如何通过训练更新网络参数的值——例如权重——的值。超参数可以包括折扣因子(discount factor)、学习率、目标函数值或指派给目标函数的各种项的权重等。此外,超参数可以包括用于确定是否在给定候选任务上训练候选智能体神经网络的一个或多个指定值。因此,通常在所描述的系统中,超参数控制参数的学习(这可以例如通过一个或多个目标函数的梯度的反向传播来发生)。

[0072] 为了训练智能体神经网络,训练系统100还维护存储用于训练候选智能体神经网络142A-M的训练数据的训练数据储存库130。训练数据储存库120被实现为一个或多个物理位置中的一个或多个逻辑存储设备,或者被实现为在一个或多个物理位置中的一个或多个

存储设备中分配的逻辑存储空间。训练数据储存库130存储定义可以在其上训练候选智能体神经网络的候选任务132A-N集合的数据,以及针对每个候选任务的可以用于控制智能体执行候选任务的对应的控制策略133A-N,例如,针对任务A 132A的控制策略133A。训练数据储存库130可选地存储定义用于评估候选智能体神经网络142A-M在验证任务上的性能的验证任务集合的数据。

[0073] 训练系统100可以以各种方式中的任何一种接收数据。例如,系统100可以通过数据通信网络(例如,使用训练系统100可用的应用编程接口(API))从系统的远程用户接收定义候选任务集合和/或验证任务集合的数据作为上传。作为另一示例,训练系统100可以从用户接收指定训练系统100已经维护的哪些数据应当被用作定义候选任务集合和/或验证任务集合的数据的输入。

[0074] 在一些实施方式中,训练数据储存库130在训练过程的过程中保持固定,而在其他实施方式中,训练数据储存库130随着训练过程的迭代被执行而扩展(例如,直到无限)或收缩,例如,新的候选任务可以被动态地添加到储存库,并且可以丢弃预定数量的候选智能体神经网络已经在其上全部达成阈值性能的现有任务。在这些其他实施方式中,训练系统100可以使用任务生成引擎124来重复地——即在训练过程中的多个训练迭代中的每个训练迭代处——从候选任务的空间生成新任务,例如以生成多任务连续体(continuum)。该候选任务的空间可以通过任务参数集来参数化,每个任务参数与一个或多个值——例如,分数、离散值或连续值——相关联。例如,任务生成引擎124可以例如通过从任务参数集中随机选择不同的任务参数值来随机生成任务。接下来,为了根据这些选择的任务参数值实际生成任务,训练系统100可以使用在Shaker,Noor,et al.Procedural content generation in games.Switzerland:Springer International Publishing,2016(Shaker,Noor等,游戏中的程序化内容生成,瑞士:施普林格国际出版社,2016)中描述的任何示例技术,补充或代替其他已知的数字内容生成技术。

[0075] 通常,任务参数集合可以包括可以共同定义候选任务的空间的各种可调整参数中的任何一个。例如,任务参数集合可以包括第一多个环境参数,该第一多个环境参数定义由智能体正在交互的环境的不同属性或特性。更详细地,环境参数可以定义或以其他方式指定:环境的静态拓扑(例如,环境的拓扑构建块的布局 and 结构)、环境的照明、表征可以提供给智能体的环境的状态的观测的类型或格式、环境中存在的动态对象的属性(例如,目标对象或障碍物的位置和物理属性)、可以由智能体响应于接收到观测而执行的预定动作集合、环境中存在的其他智能体的数量(在竞争或协作任务的情况下)、或其组合以及可能更多。

[0076] 作为另一示例,任务参数集可以包括第二多个目标参数,该第二多个目标参数定义或以其他方式指定一个或多个智能体中的每个智能体在与环境交互时要达成的不同目标。目标参数可以例如包括奖励函数的各种项,该奖励函数用于计算响应于智能体在与环境交互时执行不同动作而发布给智能体的奖励(表示为数值)。奖励通常反映智能体完成智能体被配置为执行的候选任务的指定目标的进度。

[0077] 训练系统100可以使用基于群体的训练(PBT)引擎120使用基于群体的训练技术来训练一个或多个候选智能体神经网络142A-M的群体,该训练技术在PCT专利申请No.PCT/EP2018/082162中更详细地描述,其通过引用并入本文。作为训练的一部分,PBT引擎120在一个或多个任务的相应集合上训练每个候选智能体神经网络,以更新群体储存库140中的

候选智能体神经网络的参数的参数值。然而,与传统的基于群体的训练不同,在任何给定候选任务上训练给定候选智能体神经网络之前,PBT引擎120获得识别用于训练候选智能体神经网络的候选任务的数据,并获得指定用于候选任务的控制策略的数据。

[0078] 具体地,PBT引擎120基于(i)候选智能体神经网络在候选任务上的性能(即,当根据当前存储在候选智能体神经网络的群体数据中的参数值使用时)、(ii)用于候选任务的控制策略在候选任务上的性能和(iii)候选智能体神经网络的一个或多个超参数值,来确定是否在候选任务上训练候选智能体神经网络。通常,仅当PBT引擎在给定当前学习阶段——即给定存储在群体数据中的候选者的参数的当前值——的情况下确定任务将对候选者“有用”时,PBT引擎120将在给定任务上训练候选智能体神经网络。PBT引擎120可以部分地基于相对于候选的性能的控制策略的性能来确定这一点。

[0079] 响应于确定在候选任务上训练候选智能体神经网络,PBT引擎120在候选任务上训练候选智能体神经网络以更新智能体神经网络的参数的参数值。响应于确定不在候选任务上训练候选智能体神经网络,系统避免在候选任务上训练候选智能体神经网络。

[0080] 因此,控制策略的使用改善了训练过程的有效性和效率,例如,通过仅当候选智能体神经网络的性能显著优于控制策略的性能时允许PBT引擎120在候选任务上训练候选智能体神经网络。在一些实施方式中,对于每个候选任务——例如任务A 132A,对应的控制策略——例如控制策略133A,可以是固定策略,例如,随机选择动作的随机动作策略,而在其他实施方式中,对应的控制策略可以由来自先前训练代的候选控制的策略,即,根据训练期间来自较早点的候选的历史网络参数值。在这些其他实施方式中,控制策略的使用附加地允许PBT引擎120确定候选智能体神经网络的性能相对于先前的训练迭代最近是否已经改善或恶化,并且因此采取适当的措施,例如调整超参数或简单地将其从当前群体中移除。

[0081] 在一些实施方式中,PBT引擎120在训练期间保持候选的超参数固定。在一些其他实施方式中,PBT引擎120在训练期间调整候选的超参数,以便在候选的参数值在训练过程中变化时确保上述动态任务选择继续仅选择对学习有用的任务。

[0082] 在其上训练候选的一些或所有任务可以是多智能体任务,即,要求使用候选控制的智能体与使用另一策略控制的一个或多个其他智能体交互的任务。对于这些任务,系统可以使用各种策略中的任何策略来控制一个或多个其他智能体。例如,系统可以使用固定策略,例如,随机选择动作的随机动作策略、其他智能体不执行任何动作的无操作(no-op)动作策略等、表示专家智能体的行为的专家策略或由来自先前训练代的高性能候选控制的策略。

[0083] 在训练之后,训练系统100可以选择群体中的候选之一用作智能体神经网络,即,用于控制智能体执行新任务。例如,训练系统100可以在训练已经完成之后选择最高性能的候选。可替代地,训练系统100可以使用群体中的多个候选的集成(ensemble)作为最终智能体神经网络。

[0084] 图2是用于训练候选智能体神经网络的群体的示例过程200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基于群体的神经网络训练系统——例如图1的训练系统100——可以执行过程200。

[0085] 如上所述,系统维护指定一个或多个候选智能体神经网络的群体的群体数据。群

体数据针对群体中的每个候选智能体神经网络指定 (i) 候选智能体神经网络的参数的相应参数值, (ii) 用于训练候选智能体神经网络的相应超参数值, 以及 (iii) 候选智能体神经网络的性能 (其可以包括每个不同任务的相应性能)。

[0086] 系统针对群体中的每个候选神经网络重复执行过程200。在一些实施方式中, 系统针对每个候选神经网络并行执行过程并且针对群体中的每个其他候选神经网络异步地执行过程200。

[0087] 系统获得识别用于训练候选智能体神经网络的候选任务的数据 (步骤202)。系统通常将在不同的迭代中获得不同的候选任务。不同的任务可以涉及由使用候选智能体神经网络控制的智能体与不同的环境交互、智能体在与环境交互时达成的不同目标、存在于环境中的不同数量的其他智能体 (在竞争或协作任务的情况下) 等。

[0088] 在一些实施方式中, 候选任务可以是由系统根据从候选任务的空间采样的不同任务参数集合自动并且动态地 (即, 在训练过程中) 生成的新任务。在其他实施方式中, 可以通过在每次迭代时从训练数据储存库中包括的候选任务集合中采样固定数量的候选任务来获得候选任务。在这些实施方式中的一些中, 系统可以针对群体中的所有候选智能体神经网络使用相同的任务采样计策 (strategy), 而在这些实施方式中的其他实施方式中, 系统可以针对群体使用不同的候选特定任务采样策略。附加地, 在系统维护指定先前已经在其上训练候选智能体神经网络的历史候选任务的列表的数据的实施方式中, 系统可以更新列表以包括所获得的候选任务。

[0089] 系统获得指定用于候选任务的控制策略的数据 (步骤204)。通常, 控制策略可以是可用于控制智能体与环境 (可能包括环境中的其他智能体) 交互的任何动作选择策略。例如, 用于候选任务的控制策略是均匀随机动作策略, 其在控制智能体执行候选任务时采取从预定动作集合中随机均匀选择的动作。作为另一示例, 用于候选任务的控制策略是在控制智能体执行候选任务时使用仅在候选任务上已经训练的单个候选智能体神经网络的单个任务策略。作为另一示例, 用于候选任务的控制策略是在控制智能体执行候选任务时使用候选智能体神经网络的实例但在候选智能体神经网络的训练期间具有来自较早时间点的候选智能体神经网络的网络参数的历史参数值的策略。

[0090] 系统基于 (i) 候选智能体神经网络在候选任务上的性能、(ii) 控制策略在候选任务上的性能以及 (iii) 候选智能体神经网络的一个或多个超参数值, 来确定是否在候选任务上训练候选智能体神经网络 (步骤206)。

[0091] 具体地, 为了做出该确定, 系统可以使用规划算法, 例如, 蒙特卡罗树搜索 (MCTS) 算法或另一前瞻规划算法, 来做出关于在给定时间步处环境的初始状态之后的多个未来状态的预测。这被称为任务回合。假设智能体执行通过使用候选智能体神经网络选择的某些动作, 任务回合表示在给定时间步之后的时间的环境的预演 (rollout)。系统可以运行规划算法以针对候选任务重复地生成多个任务回合, 例如五个、十个、二十个或更多个任务回合。

[0092] 任务回合是指智能体与环境交互的时间步的序列。例如, 当智能体已经与环境交互了预定义数量的时间步时, 或者当智能体完成任务时, 任务回合可以终止。

[0093] 在一些实施方式中, 每个任务回合可以包括多个轨迹的序列, 其中每个轨迹进而可以包括表征环境的相应当前状态的相应当前观测、由智能体响应于当前观测而执行的相

应当前动作、表征环境的相应下一状态的相应下一状态、以及响应于智能体执行当前动作而接收的奖励。

[0094] 对于候选任务的多个任务回合中的每个任务回合,系统通过控制智能体使用候选智能体神经网络执行候选任务的任务回合来确定接收到的相应候选智能体回报。系统还确定通过控制智能体使用控制策略执行候选任务的任务回合而接收的相应控制策略回报。候选智能体回报和控制策略回报两者都可以是根据由所选择的任务参数定义的奖励函数计算的任务特定的回报,例如,作为响应于在与环境交互时执行一个或多个动作而由智能体接收的奖励的累积测量。

[0095] 在一些实施方式中,仅当由相应候选智能体回报所测量的候选智能体神经网络的性能不超过由候选智能体神经网络的超参数值定义的阈值性能水平时,系统确定继续在候选任务上训练候选智能体神经网络。

[0096] 在这些实施方式中,仅当相应候选智能体回报超过第一指定值的一部分任务回合小于第二指定值时,系统可以确定在候选任务上训练候选智能体神经网络。例如,第一指定值可以是智能体可以接收的总奖励的可能范围之间的整数或浮点值,并且第二指定值可以是0和1之间的小数值,其中第一指定值、第二指定值或两者是候选智能体神经网络的超参数值。

[0097] 数学表示中的该标准的示例可以是:

[0098] $\Pr[R_{\pi}(x) > m_s] < m_{solved}$,

[0099] 其中, $R_{\pi}(x)$ 是相应候选智能体回报,并且 m_s 和 m_{solved} 分别是第一指定值和第二指定值。

[0100] 在一些实施方式中,仅当由相应候选智能体回报所测量的候选智能体神经网络的性能超过由相应控制策略回报所测量的控制策略的性能多于由候选智能体神经网络的超参数值定义的阈值性能水平时,系统确定在候选任务上训练候选智能体神经网络。

[0101] 在这些实施方式中,仅当相应候选智能体回报超过相应控制策略回报至少第三指定值的一部分任务回合大于第四指定值时,系统可以确定在候选任务上训练候选智能体神经网络。例如,第三指定值可以是智能体可接收的总奖励的可能范围之间的整数或浮点值,并且第四指定值可以是0和1之间的小数值,其中第三指定值、第四指定值或两者是候选智能体神经网络的超参数值。第三和第四指定值可以分别与第一指定值和第二指定值相同或不同。

[0102] 数学表示中的该标准的示例可以是:

[0103] $\Pr[R_{\pi}(x) > R_{\pi_{cont}}(x) + m_{>}] > m_{>cont}$,

[0104] 其中, $R_{\pi}(x)$ 是相应候选智能体回报, $R_{\pi_{cont}}(x)$ 是相应控制策略回报,并且 $m_{>}$ 和 $m_{>cont}$ 分别是第三指定值和第四指定值。

[0105] 在一些实施方式中,仅当由相应控制策略回报所测量的控制策略的性能低于由候选智能体神经网络的超参数值定义的阈值性能水平时,系统确定在候选任务上训练候选智能体神经网络。也就是说,系统确定在需要比例如均匀随机动作策略更复杂或更高级的控制策略的候选任务上训练候选智能体神经网络,以便达成候选任务的指定目标。

[0106] 在这些实施方式中,仅当相应控制策略回报的平均值(或中值或中心趋势的其他度量)小于第五指定值时,系统可以确定在候选任务上训练候选智能体神经网络。例如,第

五指定值可以是可由智能体接收的总奖励的可能范围之间的整数或浮点值,其中第五指定值是候选智能体神经网络的超参数值。第五指定值可以与第一指定值或第三指定值相同或不同。

[0107] 数学表示中的该标准的示例可以是:

$$[0108] \quad V_{\pi_{\text{cont}}}(\mathbf{x}) < m_{\text{cont}},$$

[0109] 其中, $V_{\pi_{\text{cont}}}(\mathbf{x})$ 是相应控制策略的预期回报,并且 m_{cont} 是第五指定值。

[0110] 在任何上述实施方式中,响应于确定在候选任务上训练候选智能体神经网络,系统使用任何适当的强化学习技术在候选任务上训练候选智能体神经网络以更新智能体神经网络的网络参数的维护值(步骤208),即,通过优化目标函数,该目标函数取决于由使用候选智能体神经网络控制的智能体从环境接收的奖励。例如,强化学习技术可以是基于策略(On-policy)的RL训练技术,例如,在Song, et al. V-MPO: On-policy maximum a posteriori policy optimization for discrete and continuous control, International Conference on Learning Representations, 2019 (Song等, V-MPO: 关于离散和连续控制的On-policy最大后验策略优化,国际学习表征会议, 2019) 中更详细地描述的RL算法之一,并且系统可以在候选任务上训练候选智能体神经网络达固定数量的迭代或设定的时间段。

[0111] 与强化学习技术一起应用的是基于群体的训练技术。具体地,系统使用候选智能体神经网络的超参数和网络参数的维护值在一个或多个候选任务的相应集合上训练候选神经网络,以迭代地生成候选智能体神经网络的更新的网络参数,直到满足PBT终止标准。PBT终止标准是一个或多个条件集合,当候选智能体神经网络满足所述一个或多个条件集合时,使得系统利用新的网络参数、新的超参数和新的性能测量来更新候选智能体神经网络的储存库。满足PBT终止标准的示例是当候选智能体神经网络已经针对迭代训练过程的固定次数的迭代(例如, $1e4$ 、 $1e6$ 、 $1e8$ 等)或设定的时间段(例如,一小时、两小时、十小时等)在一个或多个候选任务的相应集合上被训练时。满足PBT终止标准的另一个示例是当候选智能体神经网络下降到低于某个性能阈值时。

[0112] 在满足PBT终止标准时,系统执行群体储存库更新过程,包括根据候选智能体神经网络的网络参数的更新值确定候选智能体神经网络的更新性能。更新的性能反映作为更新的网络参数的结果的候选智能体神经网络的潜在性能增加。

[0113] 系统还确定候选神经网络的超参数和网络参数的新值。在一些实施方式中,系统至少基于群体储存库中的候选智能体神经网络的群体的维护性能和候选智能体神经网络的更新性能来确定超参数的新值(即,调整候选智能体神经网络的维护的超参数值)和候选智能体神经网络的网络参数的新值。

[0114] 在确定候选智能体神经网络的新网络参数、新超参数和新性能之后,迭代训练过程继续。也就是说,系统使用候选神经网络的新的超参数和新的网络参数通过对一个或多个新获得的候选任务的相应集合的强化学习来训练候选智能体神经网络,以迭代地生成用于候选智能体神经网络的更新的网络参数。系统将继续候选智能体神经网络的迭代训练过程,直到满足下一个PBT终止标准(并且系统重复候选智能体神经网络的群体储存库更新过程)。虽然在一些实施方式中,系统可以终止候选智能体神经网络的迭代训练过程,例如,当PBT终止标准已经满足预定次数或当性能标准被满足以指示系统停止训练时,在其他实施

方式中,系统可以继续迭代训练过程直到无限。

[0115] 图3示出了训练候选智能体神经网络的群体的示例图示。例如图1的训练系统100的训练系统并行地执行用于群体(“群体1”)中的候选智能体神经网络A-G 142A-G的迭代训练过程,并且直到满足PBT终止标准。当训练系统执行候选智能体神经网络A-G 142A-G的迭代训练过程时,相应地更新候选智能体神经网络的网络参数A-G。

[0116] 系统基于在多个验证任务中的每个验证任务上的候选智能体神经网络的性能的评估来更新每个候选智能体神经网络的性能。可以通过从验证任务集合中采样来获得多个验证任务。验证集合中的任务数量通常远小于候选任务集合中的任务总数。例如,训练数据储存库可以存储一百万个、五百万个或更多个候选任务的集合,而验证任务的数量大约为数千。在一些实施方式中,采样可以是随机采样,而在其他实施方式中,采样可以可替代地是偏斜(skewed)采样,以便确保验证任务集合的均匀覆盖。每个验证任务的更新的性能可以取决于根据由所选择的任务参数定义的奖励函数计算的任务特定的回报,例如,作为响应于在与环境交互时执行使用候选智能体神经网络选择的一个或多个动作而由智能体接收的奖励的累积测量。

[0117] 与其中性能直接用于更新每个候选智能体神经网络的超参数和网络参数的传统的基于群体的训练不同,因为评估任务可以在其复杂度、回报的规模或两者方面变化,所以系统附加地计算每个候选智能体神经网络的相应PBT适合度测量(或简称“适合度测量”)310,其然后用于更新候选智能体神经网络的超参数和网络参数。适合度测量310允许跨评估任务集合更好地表征网络性能和鲁棒性。

[0118] 具体地,通过确定候选智能体神经网络在多个验证任务中的每个验证任务上的相应归一化百分位度量来计算相应适合度测量,该相应适合度测量可以被视为每个候选智能体神经网络跨多个验证任务的性能的多维测量。

[0119] 图7示出了确定归一化百分位度量的示例图示。如图所示,对于每个候选智能体神经网络,为了确定验证任务711、712、713、714和715中的每个验证任务的候选智能体神经网络的群体上的归一化百分位度量,系统可以使用候选在验证任务上的最佳性能(例如,任务711的最佳性能701B)作为归一化常数,然后使用归一化常数来归一化候选智能体神经网络的性能(例如,任务711的候选智能体神经网络性能701A)。对于每个候选智能体神经网络,然后对跨多个验证任务的候选的归一化性能(例如,任务714的归一化候选智能体神经网络性能724C)进行重新排序,例如,以单调递增的次序布置(如图7所示的归一化百分位曲线所指示的),从中可以确定每个验证任务上的候选的相应归一化百分位度量。

[0120] 在图7的示例中,候选在任何给定验证任务上的最佳性能可以是由候选智能体神经网络获得的最佳性能,如由候选在验证任务上接收的相应候选智能体回报(在单智能体任务的情况下)测量的,或者可以可替代地是由多个候选智能体神经网络的组(纳什均衡)之一获得的最佳性能,如由候选在验证任务上接收的相应候选智能体回报(在多智能体任务的情况下)测量的。

[0121] 然后,系统将基于群体的训练技术应用于相应的适合度测量,以调整群体中的一个或多个候选智能体神经网络142A-G中的每一个的相应超参数值。例如,响应于将候选智能体神经网络的相应适合度测量彼此进行比较并且确定候选智能体神经网络142D的适合度测量优于群体中的另一候选智能体神经网络——例如候选智能体神经网络142E——的

相应适合度测量,系统将候选智能体神经网络142D识别为当前群体中的“更好”性能的候选。例如,比较结果可以基于帕累托占优。作为这里的示例,如果针对例如多个验证任务的多个任务的候选的例如归一化百分位度量的所有适合度测量至少与针对例如多个验证任务的多个任务的另一候选的适合度测量一样好,但是对于例如验证任务的至少一个任务在例如归一化百分位度量的适合度测量上严格地是更好的,则该候选的性能可以被认为是帕累托占优另一候选的性能。

[0122] 在该示例中,系统然后可以通过“开发(exploit)”候选智能体神经网络142D的网络超参数来确定其他候选智能体神经网络的网络超参数的新值,即,将其他候选智能体神经网络的超数值调整为等于候选智能体神经网络142D的超数值。在该示例中,系统可以可替代地通过“探索”候选智能体神经网络142D的网络超参数来确定其他候选智能体神经网络的网络超参数的新值,即,将其他候选智能体神经网络的超数值调整为等于候选智能体神经网络142D的超数值的变异(例如,随机扰动)版本。

[0123] 系统类似地将基于群体的训练技术应用于相应的适合度测量,以调整群体中的每个候选智能体神经网络142A-G的相应参数值。继续上面的示例,其中候选智能体神经网络142D已经被识别为当前群体中的“更好”性能候选,系统可以将另一候选智能体神经网络——例如候选智能体神经网络142E——的相应参数值设置为等于候选智能体神经网络142D的相应参数值(或相应参数值的变异版本)。

[0124] 如图3所示,在一些实施方式中,系统附加地将世代(generational)训练技术结合到基于群体的训练技术中,以进一步提高RL训练的有效性和速度。在这些实施方式中,候选智能体神经网络的训练跨越训练代序列,其中在每个训练代处训练的相应候选智能体神经网络可以被统称为一个特定群体。例如,每个训练代可以包括迭代训练过程的固定迭代次数(例如,1e8、5e8、10e8等)或设定时间段(例如,12小时、24小时、48小时等)。

[0125] 尽管图3示出了其中在四个训练代的序列中的每个训练代处训练七个候选智能体神经网络的相应群体的示例实施方式,但是可以存在针对更多或更少的训练代训练的更多并且有时更多数量级的候选智能体神经网络。

[0126] 基于世代群体的训练使得可以在训练期间使用策略蒸馏(distillation)技术,这通常使得当前训练代中的候选智能体神经网络能够从先前训练代中的另一候选引导其行为。执行策略蒸馏技术涉及:在训练候选智能体神经网络的群体期间的训练代序列的当前训练代期间,在相同或不同的候选任务上训练候选智能体神经网络以优化预期回报,在所述候选任务上已经训练了来自先前训练代的“更好”性能的候选智能体神经网络(例如,在任务特定的性能或适合度测量方面),同时从最佳性能的候选智能体神经网络中蒸馏,例如在序列中紧接在当前训练代之前的先前训练代结束时从最佳性能候选中蒸馏。

[0127] 具体地,系统可以将辅助策略蒸馏损失项合并到候选智能体神经网络的RL训练中使用的目标函数中。

[0128] $\ell_t^{\text{exp-distill}} := (1 - r_{t-1}) \cdot \ell_t^{\text{distill}}$, 其中

[0129] $\ell_t^{\text{distill}} := \text{KL} \left[\left[\pi_t^{\text{teacher}} \right] \parallel \pi_t \right]$, 并且其中

[0130] π^{teacher} 和 π 分别是“更好”性能候选智能体神经网络和“学习”候选智能体神经网络的动作选择输出。在上面的示例等式中,策略蒸馏损失项在获得奖励的环境状态上被掩蔽,

即,辅助损失(以Kullback-Leibler散度)仅影响在没有接收到奖励的各种训练任务回合中的某些时间步上的RL训练。

[0131] 另外,在这些实施方式中,来自先前训练代的“更好”性能候选智能体神经网络(例如,在任务特定的性能或适合度测量方面)可以用于控制多智能体候选任务中的其他智能体,以增加RL训练的多样性。

[0132] 在一些实施方式中,可以在自奖励游戏目标(play objective)上训练与序列中的一个或多个训练代中的每个训练代相对应的群体中的至少一些候选智能体神经网络。这可以用于通过确保候选智能体神经网络仅无法在较少数量的任务上达成对应的目标来帮助RL训练。具体地,自奖励游戏目标奖励候选智能体神经网络以满足目标 g ,并且在成功之后,候选被奖励在相同环境内履行 $\text{not}(g)$,其中在每次满足之后重复目标的该翻转。例如,假设目标 g 是朝向目标位置移动对象,则 $\text{not}(g)$ 是移动对象远离目标位置。这可以被视为两个智能体以竞争的方式相对于它们自己游玩,其中一个智能体必须满足 g 并且另一个智能体必须满足 $\text{not}(g)$,然而智能体顺序地动作,并且由相同的候选智能体神经网络控制。

[0133] 图4A示出了示例强化学习智能体控制系统400。强化学习智能体控制系统400是被实现为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,其中可以实现下面描述的系统、组件和技术。

[0134] 强化学习系统400包括作为本说明书中描述的开放式强化学习训练过程的结果而获得的智能体神经网络420。

[0135] 强化学习智能体控制系统400通过使用智能体神经网络420选择要由智能体408执行的动作410并且然后使智能体408执行所选择的动作410来控制与环境402交互的智能体408。

[0136] 智能体408执行所选择的动作410通常使环境402转换到新状态。通过重复地使智能体408在环境402中动作,系统400可以控制智能体408以完成多个指定任务中的每一个。

[0137] 在多个时间步中的每个时间步处,智能体神经网络420被配置为根据网络参数的经训练的值来处理包括(i)表征环境402的当前状态的当前观测404和(ii)表示为了在环境中执行任务而要满足的目标的目标数据406的输入以生成动作选择输出。目标可以被表示为相应谓词(predicate)集合上的选项集合,其中,每个谓词可以将环境的当前状态映射到对应的奖励值,并且其中,每个选项是一个或多个谓词的合取(conjunction),例如,使得仅在所有谓词回报非零奖励时才接收奖励。在一些实施方式中,目标数据406可以被表示为数值的一个或多个张量,例如,每个谓词被表示为多热(multi-hot)编码向量。

[0138] 具体地,在一些实施方式中,表征环境402的当前状态的当前观测404可以包括定义环境中存在的不同实体(例如,目标对象、障碍物和其他智能体)的位置、取向、速度等的信息。在这些实施方式中,可以定义相对于实体中的一些或全部的物理关系形式的原子谓词集合 ϕ_j ,并且目标可以被定义为原子谓词集合 ϕ_j 上的布尔表达式。这些物理关系可以例如包括:在实体是目标对象、障碍物、智能体或诸如环境的地面的静态拓扑构建块的情况下,接近、在其上、看到和保持实体以及其取反。因此,示例谓词可以是near(purple sphere,opponent)(接近(紫色球,对手)),这表示使对手智能体之一接近环境中的紫色球的目标。在可能谓词集合固定的情况下,智能体的目标可以由针对该选项(合取(conjunction))的相关(或必要)谓词集合的选项集合(析取(disjunction))来表示。换句

话说,目标可以是每个观测S指派给谓词真值的二进制向量的映射 $\varphi_j: S \rightarrow \{0,1\}^d$,其中0指示当环境处于由观测S表征的状态时目标的不相关谓词,并且1指示当环境处于由观测S表征的状态时目标的相关谓词。例如,由目标数据406表示的目标可以看起来像

$$[0139] \quad \mathbf{g} = \underbrace{(\phi_{j_1} \wedge \phi_{j_2})}_{\text{选项1}} \vee \underbrace{(\phi_{j_2} \wedge \phi_{j_3} \wedge \phi_{j_4})}_{\text{选项2}}$$

[0140] 对于一些示例谓词,这可以意味着“保持紫色球(φ_{j_1})同时接近黄色球(φ_{j_2})或者接近黄色球(φ_{j_2})同时看到不保持黄色球(φ_{j_3})的对手智能体(φ_{j_4})”。

[0141] 然后,系统400使用动作选择输出来控制智能体,即,根据动作选择输出来选择要由智能体在当前时间步执行的动作410,然后使智能体执行动作410,例如,通过直接向智能体传送控制信号或通过向智能体的控制系统传送识别动作410的数据。上面参考图1描述了示例动作选择输出以及它们如何可以用于选择动作。

[0142] 智能体神经网络420利用使其能够执行其描述的功能的神经网络架构来实现。如图4A所示,智能体神经网络420包括状态编码器神经网络430、嵌入神经网络440、注意力神经网络445、值神经网络(值神经网络“V”头)450和策略神经网络(策略神经网络“ π ”头)460。也就是说,值神经网络也可以被称为值神经网络头;类似地,策略神经网络也可以被称为策略神经网络头;术语“头”可以但不必指示值神经网络和策略神经网络两者都接收由注意力神经网络445的一个或多个层生成的共享数据作为输入。神经网络420、430、440、445和450中的每一个包括智能体神经网络420中的多个神经网络层的不同子集。

[0143] 状态编码器神经网络430可以包括多个卷积层的堆叠,随后是一个或多个池化层(例如,最大池化层),并且随后是一个或多个递归层(例如,长短期记忆(LSTM)层)。状态编码器神经网络430被配置为——即通过训练——接收观测404,以通过处理接收到的观测404来更新状态编码器神经网络430的隐藏状态,即,生成表示环境的当前状态的当前隐藏状态432。通常,嵌入是具有固定维度的数值或其他值的有序合集。固定维度可以取决于包括在状态编码器神经网络430的每个LSTM层中的LSTM神经元的实际数量,例如128、256等。

[0144] 嵌入神经网络440可以是全连接神经网络,即,其包括多个全连接层,可选地跟随有激活层(例如,ReLU激活层),其被配置为处理目标数据406和当前隐藏状态432以生成(i)谓词预测,(ii)目标的目标嵌入,以及(iii)由目标数据表示的选项中的每个选项的相应选项嵌入。嵌入神经网络440进而可以包括谓词预测器神经网络、目标嵌入神经网络和选项嵌入神经网络,其中的每个包括嵌入神经网络440的多个全连接层的相应子集。

[0145] 谓词预测器神经网络被配置为处理目标数据406和当前隐藏状态432以生成谓词预测。在一些实施方式中,谓词预测可以是多标签二元分类预测,其指定来自原子谓词集合 φ_j 的哪些谓词与目标数据406中表示的目标相关(例如包括在目标中),假定环境处于由当前观测表征的状态。

[0146] 选项嵌入神经网络被配置为处理目标数据406和当前隐藏状态432、从目标数据406和当前隐藏状态432导出的数据或两者,以生成选项中的每个选项的相应选项嵌入。同样,目标嵌入神经网络被配置为处理目标数据406和当前隐藏状态432、从目标数据406和当前隐藏状态432导出的数据或两者,以生成目标的目标嵌入。通常,嵌入是具有固定维度的数值或其他值的有序合集。固定维度可以取决于嵌入神经网络440的每个全连接层中包括的实际隐藏单元数量,例如128、256等。

[0147] 注意力神经网络445被配置为处理包括(i)当前隐藏状态、(ii)目标嵌入和(iii)谓词预测的输入以生成目标注意力隐藏状态。如本文所使用的,注意力神经网络是包括一个或多个注意力层的神经网络,每个注意力层是包括注意力机制——例如缩放点积注意力机制——的神经网络层。为了生成目标注意力隐藏状态,注意力机制将查询和键值对集合映射到输出(目标注意力隐藏状态),其中查询可以来自目标嵌入(或以其他方式从其导出),并且其中键值对集合可以从当前隐藏状态432导出(例如,当前隐藏状态的线性投影和/或重塑版本)。

[0148] 另外,注意力神经网络445被配置为针对选项中的每个选项处理包括(i)当前隐藏状态、(ii)选项的相应选项嵌入和(iii)谓词预测的输入,以生成选项的相应选项注意力隐藏状态。为了为每个选项生成相应的选项注意力隐藏状态,注意力机制类似地将查询和键值对集合映射到输出(选项注意力隐藏状态),其中查询可以是选项的选项嵌入,并且其中,键值对集合可以从当前隐藏状态432导出。

[0149] 值神经网络450可以是全连接神经网络,即,包括一个或多个全连接层,可选地跟随有激活层(例如,ReLU激活层),其被配置为处理目标注意力隐藏状态以生成目标的目标值估计,该目标值估计表示如果智能体尝试从当前状态开始满足目标,则将达成的估计回报,例如,将由智能体接收的奖励的累积测量。值神经网络450还被配置为对于选项中的每个选项,使用值神经网络头处理选项的相应选项注意力隐藏状态,以生成选项的相应选项值估计,该选项值估计表示如果智能体尝试从当前状态开始满足选项则将达成的估计回报。

[0150] 然后将目标注意力隐藏状态和相应的选项注意力隐藏状态组合以生成要提供给策略神经网络460的组合隐藏状态,策略神经网络460被配置为处理组合隐藏状态以生成动作选择输出。在一些实施方式中,组合隐藏状态可以是未加权组合,而在其他实施方式中,组合隐藏状态可以是加权组合,即,根据它们相应的权重组合目标注意力隐藏状态和相应的选项注意力隐藏状态。

[0151] 在这些实施方式中的一些实施方式中,可以通过在选项的目标值估计和相应选项值估计上应用softmax函数以生成目标和选项中的每个选项的相应softmax分数,来计算对于选项的来自目标值估计和相应选项值估计的目标和选项中的每个选项的相应权重,然后将其用作用于计算加权组合的权重。通常,softmax函数是将数值转换成概率的函数。在这些实施方式的其他实施方式中,可以通过将权重1指派给选项的相应选项值估计和目标值估计的最高值估计以及将权重0指派给选项的相应选项值估计和目标值估计的所有其他值估计来计算相应权重。

[0152] 策略神经网络460可以包括一个或多个全连接层,随后是一个或多个输出层。在一些实施方式中,一个或多个输出层可以包括单个softmax层,该单个softmax层被配置为处理先前全连接层的输出以生成可以由智能体执行的该动作集合上的概率分布。在一些其他实施方式中,一个或多个输出层可以包括分别与可以由智能体执行的动作集合的不同子集相对应的多个softmax层,每个softmax层被配置为处理先前全连接层的输出以在该子集上生成相应的概率分布。又在其他实施方式中,一个或多个输出层可以包括一个或多个附加全连接层,其被配置为针对可以由智能体执行的动作集合中的每个动作生成相应的Q值。

[0153] 图4B示出了智能体神经网络420的一个示例实施方式的细节。

[0154] 图5是用于控制与环境交互的智能体的示例过程500的流程图。为了方便起见,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执行。例如,适当编程的强化学习系统——例如,图4的强化学习智能体控制系统400——可以执行过程500。

[0155] 系统接收表征环境的当前状态的观测(步骤502)。例如,观测可以包括音频数据段、图像或自然语言的句子。可选地,观测还可以包括从先前时间步导出的信息,例如,执行的先前动作、在先前时间步接收的奖励或两者。

[0156] 系统接收表示要被满足的目标的目标数据以便在环境中执行任务(步骤504)。在接收到的目标数据中,目标被表示为相应谓词集合上的选项集合。在图4的示例中,目标被表示为两个选项的集合,每个选项包括单个谓词“Hold purple sphere(保持紫色球)”或“See yellow cube(看到黄色立方体)”,尽管在其他示例中,目标可以由更多选项表示,每个选项由更多谓词组成,有时具有跨多个选项共享的相同谓词。

[0157] 系统使用智能体神经网络处理包括观测和目标数据的输入以生成动作选择输出(步骤506),如下面将参考图6进一步描述的。在一些实施方式中,动作选择输出可以定义可以由智能体执行的动作集合上的概率分布。在一些实施方式中,动作选择输出可以包括可以由智能体执行的动作集合中的每个动作的相应Q值。在一些实施方式中,动作选择输出是来自连续动作空间的动作,即,单个动作中的所有动作值是从可能值的连续范围中选择的。

[0158] 系统使用动作选择输出来选择要由智能体执行的动作(步骤508)。在动作选择输出定义概率分布的实施方式中,系统可以通过根据动作的概率值对动作进行采样或者通过选择具有最高概率值的动作来选择动作。在动作选择输出包括Q值的实施方式中,系统可以处理Q值(例如,使用softmax函数)以生成每个可能动作的相应概率值,其可以用于选择要由智能体执行的动作。系统还可以选择具有最高Q值的动作作为要由智能体执行的动作。

[0159] 系统例如通过指示智能体执行动作或向智能体的控制系统传递控制信号来使智能体执行所选择的动作(步骤510)。

[0160] 图6是用于使用智能体神经网络来生成动作选择输出的示例过程600的流程图。为了方便起见,过程600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执行。例如,适当编程的强化学习系统——例如,图4的强化学习系统400——可以执行过程600。

[0161] 系统使用智能体神经网络的状态编码器神经网络来处理观测以生成表示环境的当前状态的当前隐藏状态(步骤602)。状态编码器神经网络包括一个或多个递归层,并且被配置为通过处理接收到的当前观测来生成当前隐藏状态,以修改已经通过处理先前观测生成的状态编码器神经网络的隐藏状态。

[0162] 系统使用智能体神经网络的嵌入神经网络来处理目标数据和当前隐藏状态,以生成(i)谓词预测,(ii)目标的目标嵌入,以及(iii)由目标数据表示的每个选项的相应选项嵌入(步骤604)。

[0163] 系统使用注意力神经网络处理包括(i)当前隐藏状态、(ii)目标嵌入和(iii)谓词预测的输入以生成目标注意力隐藏状态(步骤606)。至少部分地通过对(i)当前隐藏状态、(ii)目标嵌入和(iii)谓词预测应用注意力机制来生成目标注意力隐藏状态。

[0164] 对于选项中的每个选项,系统使用注意力神经网络处理包括(i)当前隐藏状态、(ii)选项的相应选项嵌入和(iii)谓词预测的输入,以生成选项的相应选项注意力隐藏状态(步骤608)。至少部分地通过对(i)当前隐藏状态、(ii)选项的相应选项嵌入和(iii)谓词

预测,应用注意力机制来生成选项注意力隐藏状态。

[0165] 系统使用智能体神经网络的值神经网络来处理目标注意力隐藏状态,以生成目标的目标值估计,该目标值估计表示如果智能体尝试从当前状态开始满足目标则将达成的估计回报(步骤610)。

[0166] 对于选项中的每个选项,系统使用值神经网络处理选项的相应选项注意力隐藏状态,以生成选项的相应选项值估计,该选项值估计表示如果智能体尝试从当前状态开始满足选项则将达成的估计回报(步骤612)。

[0167] 系统从选项的目标值估计和相应选项值估计生成目标和每个选项的相应权重(步骤614)。在一些实施方式中,生成相应权重可以包括对选项的相应选项值估计和目标值估计应用softmax。在一些其他实施方式中,生成相应权重可以包括将权重1指派给选项的相应选项值估计和目标值估计的最高值估计,并且将权重0指派给选项的相应选项值估计和目标值估计的所有其他值估计。

[0168] 系统根据相应权重将目标注意力隐藏状态和相应选项注意力隐藏状态组合以生成组合隐藏状态(步骤616)。

[0169] 系统使用智能体神经网络的策略神经网络来处理组合隐藏状态以生成动作选择输出(步骤618)。

[0170] 本说明书结合系统和计算机程序组件使用术语“被配置”。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得装置执行操作或动作。

[0171] 在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系统中、在有形地体现的计算机软件或固件中、在计算机硬件中——包括在本说明书中公开的结构及其结构等同物——或在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,编码在有形非暂时性存储介质上的计算机程序指令的一个或多个模块,用于由数据处理装置执行或控制数据处理装置的操作。计算机存储介质可以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。可替代地或附加地,程序指令可以被编码在人工生成的传播信号——例如机器生成的电信号、光信号或电磁信号——上,该人工生成的传播信号被生成以编码用于传输到合适的接收器装置以供数据处理装置执行的信息。

[0172] 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。该装置还可以是或进一步包括专用逻辑电路系统,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,该装置可以可选地包括创建用于计算机程序的执行环境的代码,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。

[0173] 也可以被称为或描述为程序、软件、软件应用、应用、模块、软件模块、脚本或代码的计算机程序可以以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、子例程或适合于在计算环境中使用的其他单元。程序可以但不必与文件系统中的文件相对应。程序可以存储在保

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,存储在专用于所讨论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、子程序或代码部分的文件。可以部署计算机程序以在位于一个站点或分布在多个站点上并通过数据通信网络互连的一个计算机或多个计算机上执行。

[0174] 在本说明书中,术语“引擎”广泛地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,引擎将被实现为安装在一个或多个位置中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,可以在同一计算机或多个计算机上安装和运行多个引擎。

[0175] 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例如FPGA或ASIC的专用逻辑电路系统或者由专用逻辑电路和一个或多个编程计算机的组合来执行。

[0176] 适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执行指令的中央处理单元以及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可以由专用逻辑电路系统补充或并入专用逻辑电路系统中。通常,计算机还将包括或可操作地耦合以从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——接收数据或向其传送数据或两者。然而,计算机不需要具有这样的设备。此外,计算机可以被嵌入于另一设备中,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,仅举几例。

[0177] 适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盘;磁光盘;以及CDROM和DVD-ROM盘。

[0178] 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,该计算机具有用于向用户显示信息的显示设备,例如CRT(阴极射线管)或LCD(液晶显示器)监视器,以及用户可以通过其向计算机提供输入的键盘和指点设备,例如鼠标或轨迹球。其他种类的设备也可以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设备发送文档和从用户使用的设备接收文档来与用户交互;例如,通过响应于从用户的设备上的web浏览器接收到的请求而向web浏览器发送网页。此外,计算机可以通过向个人设备——例如正在运行消息传送应用的智能电话——发送文本消息或其他形式的消息并且作为回报从用户接收响应消息来与用户交互。

[0179] 用于实现机器学习模型的数据处理装置还可以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——即推断——工作负载的公共和计算密集型部分。

[0180] 可以使用机器学习框架——例如TensorFlow框架——来实现和部署机器学习模型。

[0181] 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

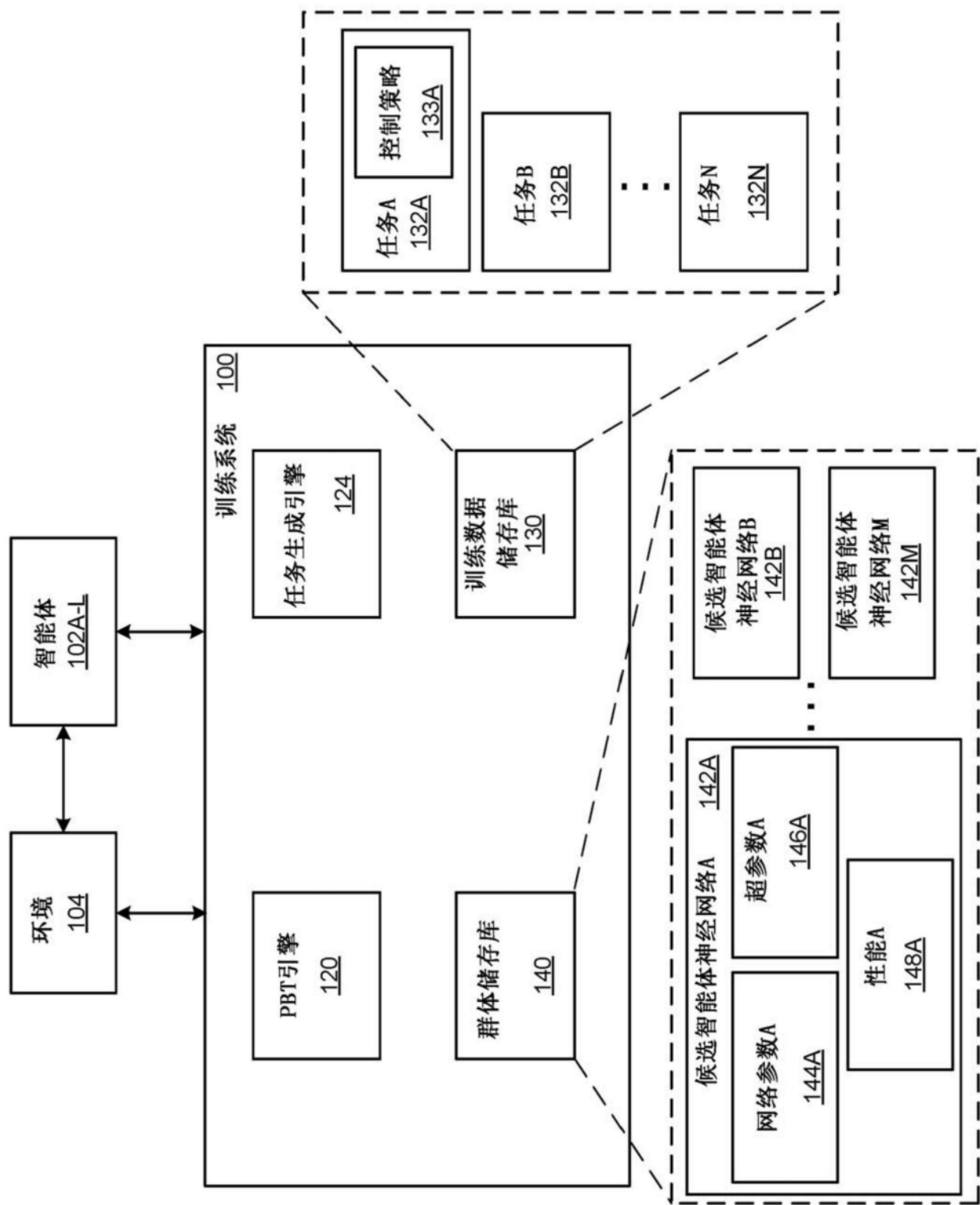
件,例如,作为数据服务器,或者包括中间件组件,例如,应用服务器,或者包括前端组件,例如,具有用户可以通过其与本说明书中描述的主题的实施方式进行交互的图形用户界面、web浏览器或应用的客户端计算机,或者一个或多个这样的后端、中间件或前端组件的任何组合。系统的组件可以通过任何形式或介质的数字数据通信——例如通信网络——互连。通信网络的示例包括局域网(LAN)和广域网(WAN)——例如因特网。

[0182] 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。客户端和服务端的关系通过在相应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,服务端将数据——例如HTML页面——传输到用户设备,例如,用于向充当客户端的设备交互的用户显示数据和从其接收用户输入。可以在服务端处从设备接收在用户设备处生成的数据,例如用户交互的结果。

[0183] 虽然本说明书包含许多具体实施方式细节,但是这些不应当被解释为对任何发明的范围或对可以要求保护的范围的限制,而是作为可以特定于特定发明的特定实施例的特征的描述。在本说明书中在分离实施例的上下文中描述的某些特征也可以在单个实施例中组合实现。相反,在单个实施例的上下文中描述的各种特征也可以分开地或以任何合适的子组合在多个实施例中实现。此外,尽管上面可以将特征描述为以某些组合起作用并且甚至最初如此要求保护,但是在一些情况下,来自要求保护的组合的一个或多个特征可以从组合中去除,并且要求保护的组合可以涉及子组合或子组合的变型。

[0184] 类似地,虽然在附图中以特定次序描绘并且在权利要求中叙述了操作,但是这不应被理解为要求以所示的特定次序或以顺序次序执行这样的操作,或者执行所有所示的操作以实现期望的结果。在某些情况下,多任务和并行处理可能是有利的。此外,上述实施例中的各种系统模块和组件的分离不应被理解为在所有实施例中都需要这样的分离,并且应当理解,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被封装到多个软件产品中。

[0185] 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书的范围内。例如,权利要求中记载的动作可以以不同的次序执行,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顺序次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。



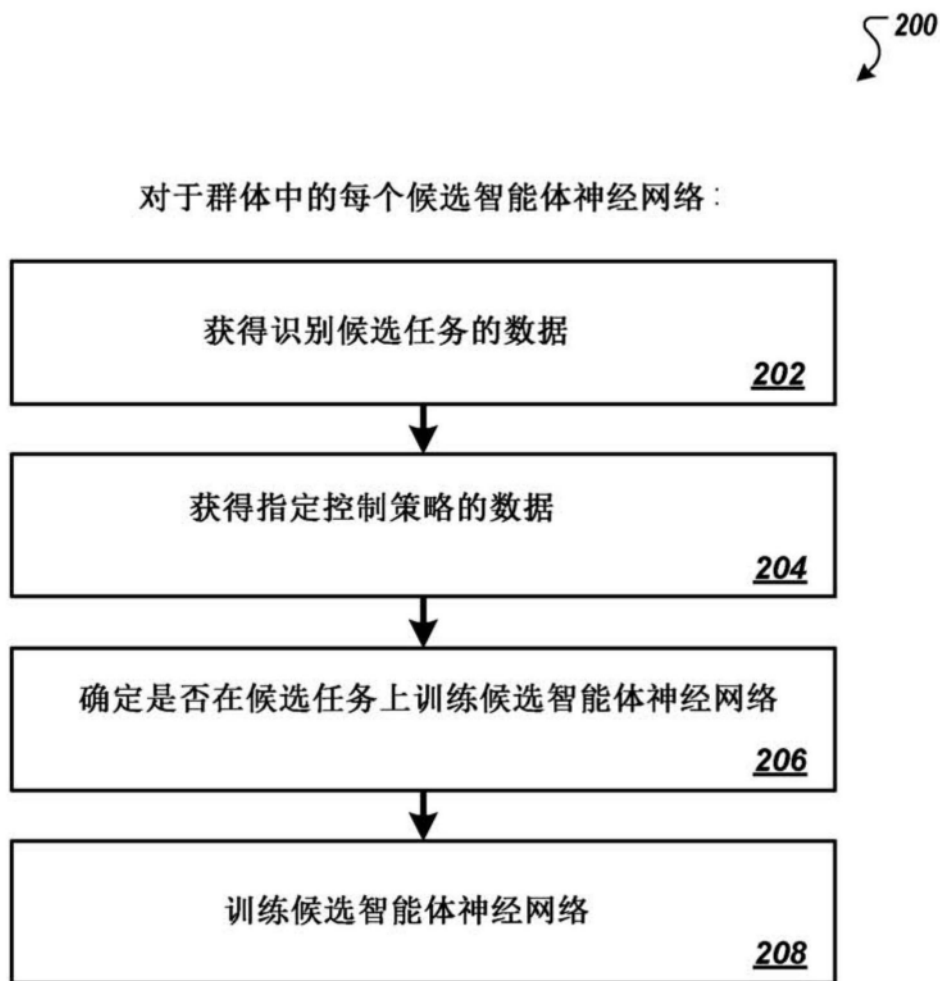


图2

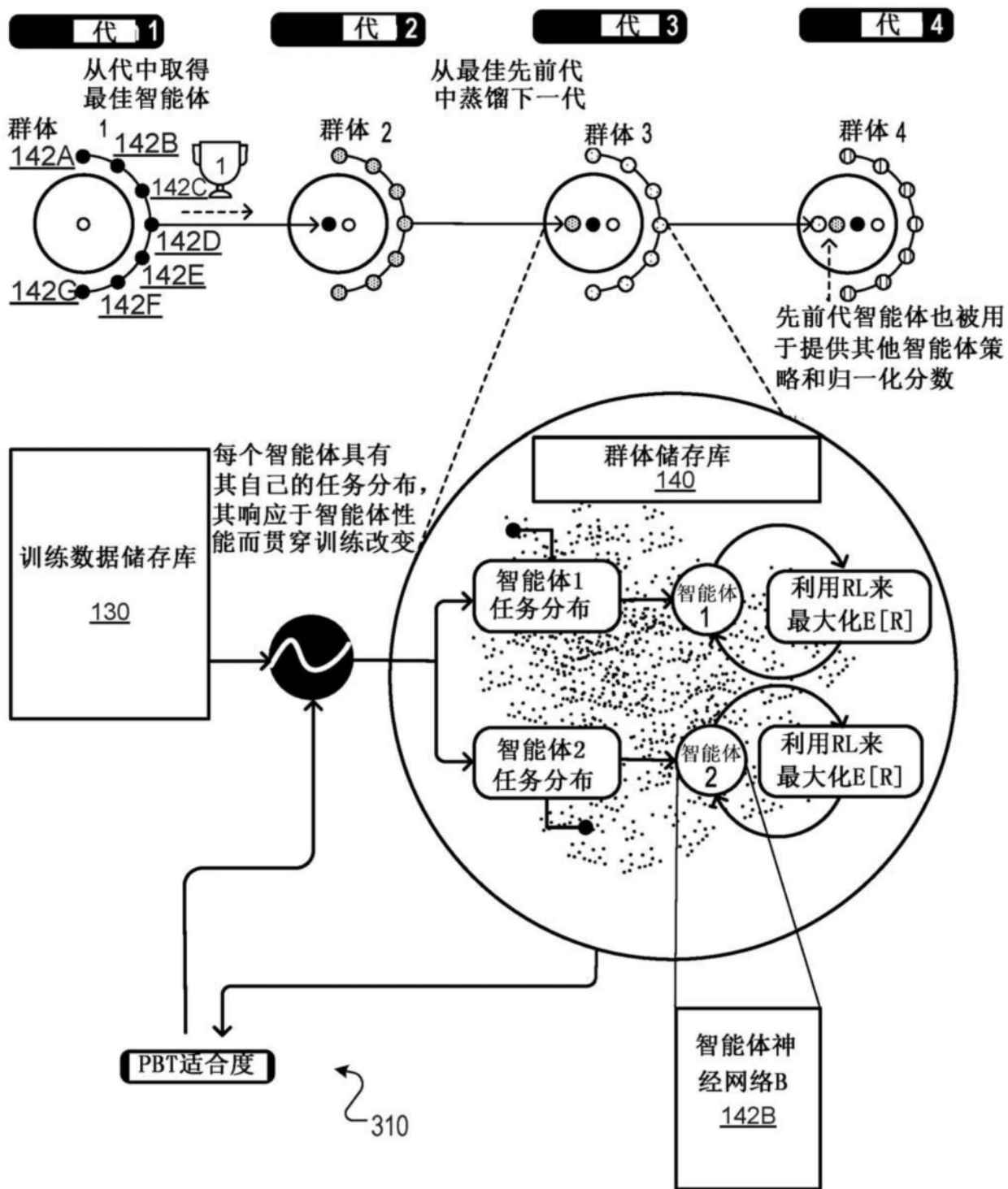


图3

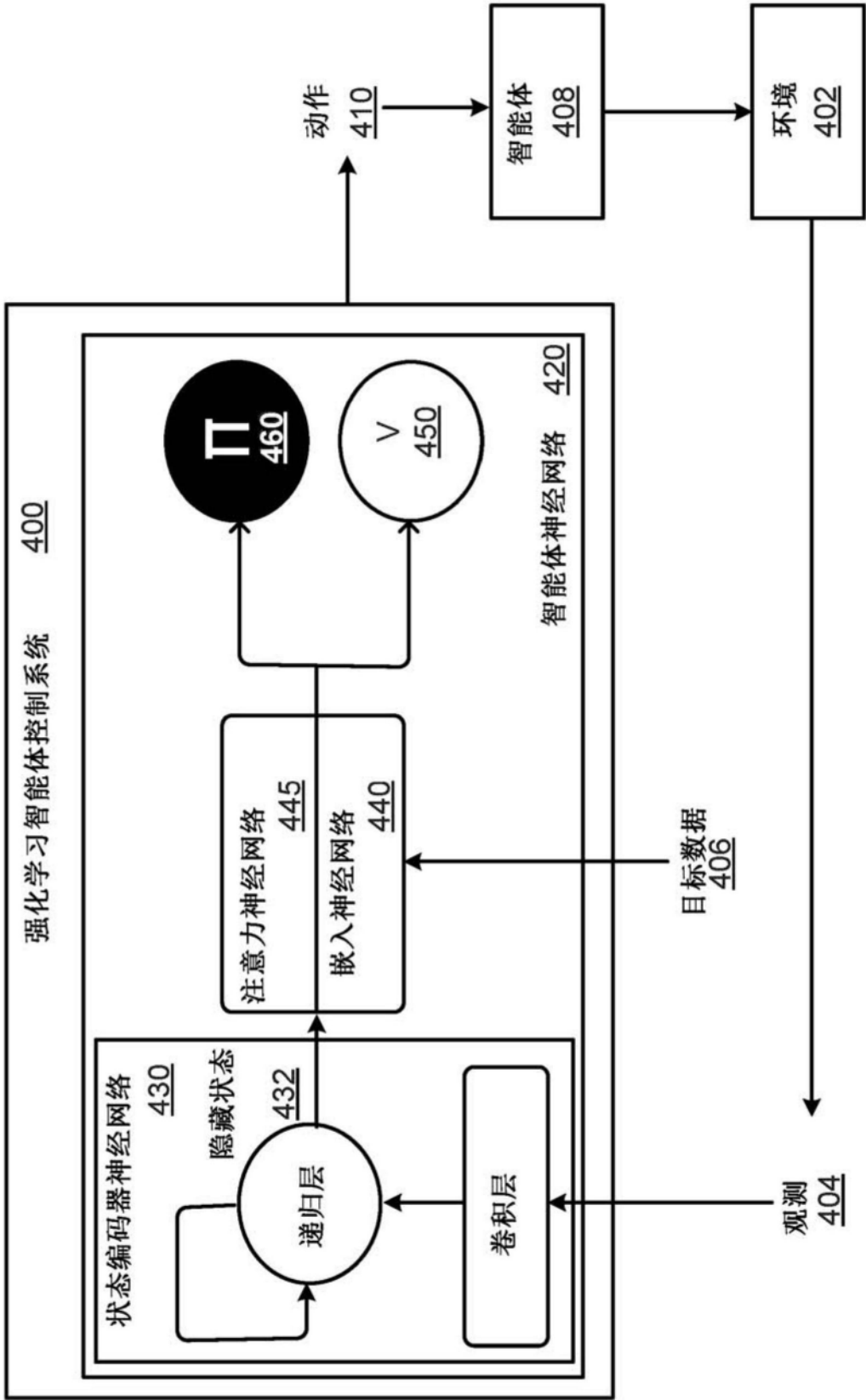


图4A

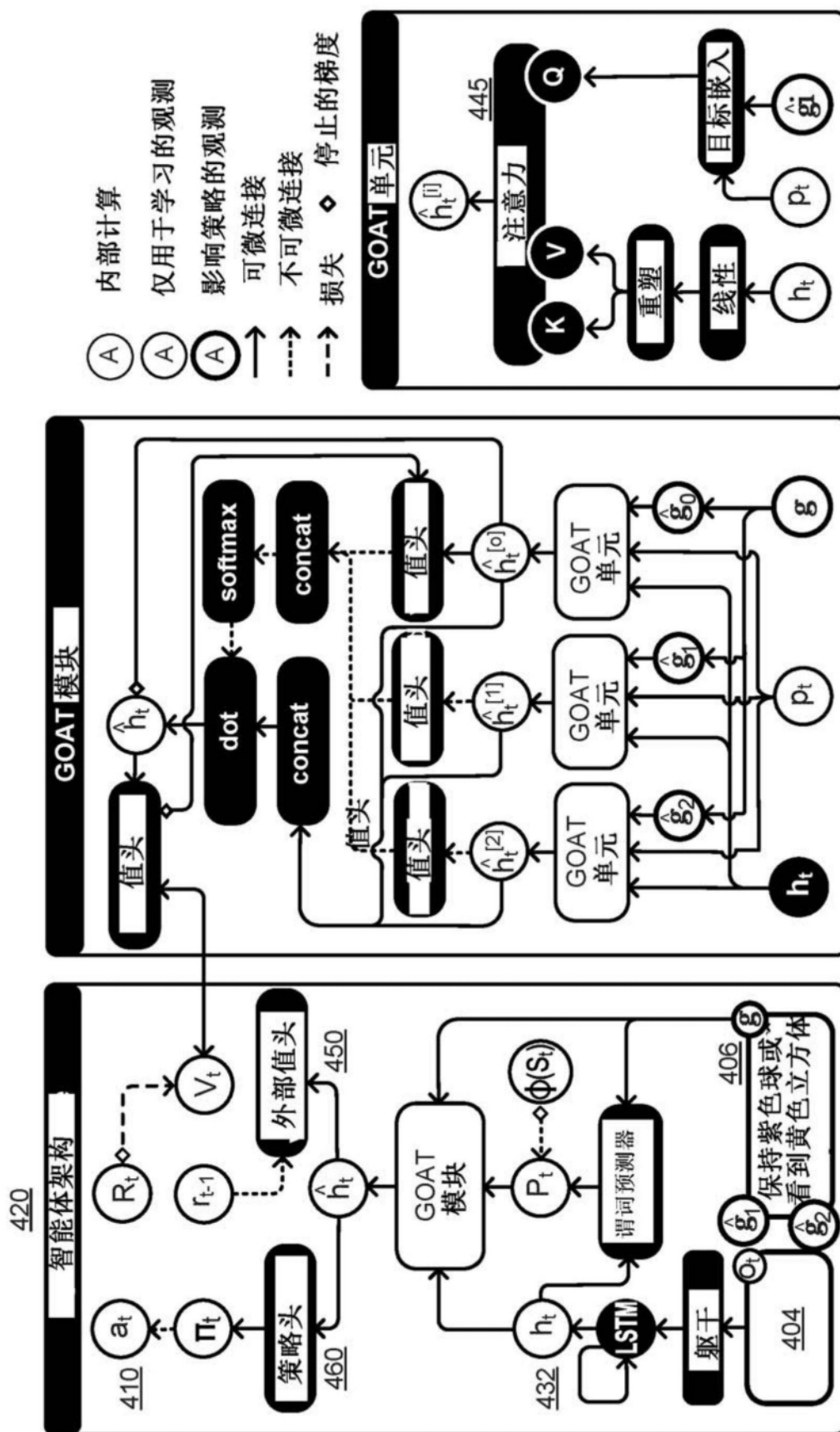


图4B

500

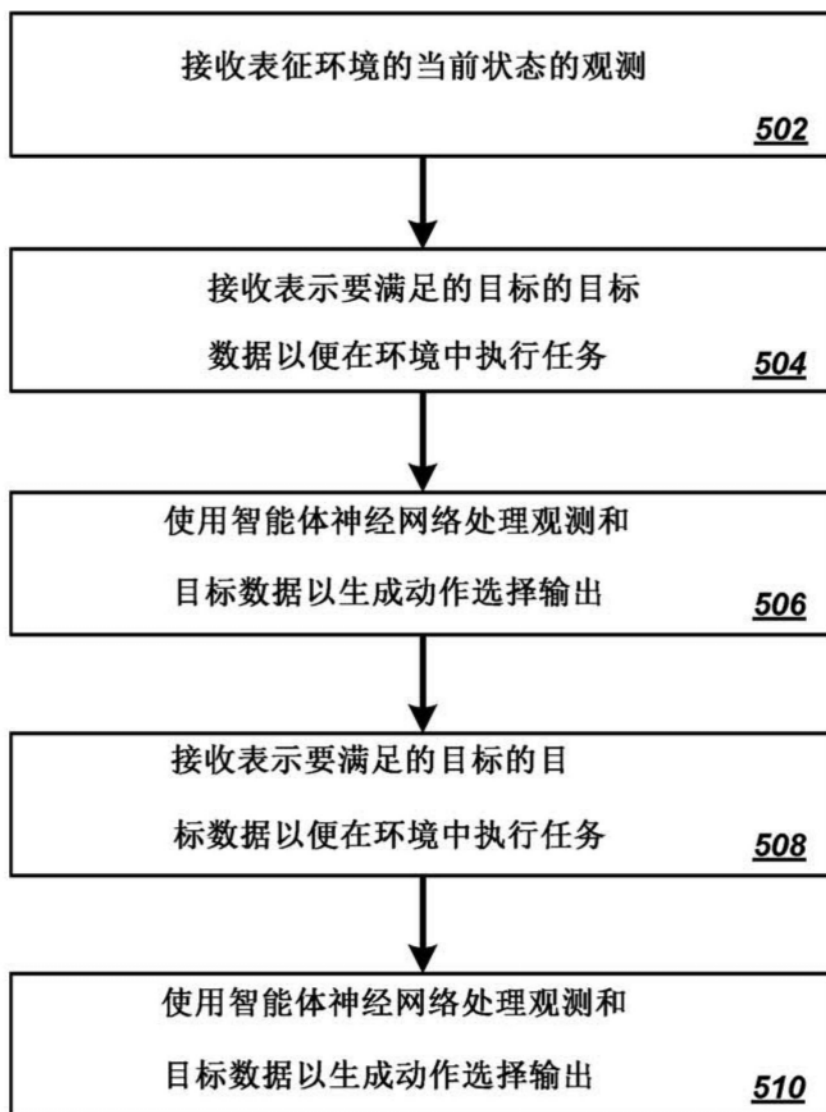


图5

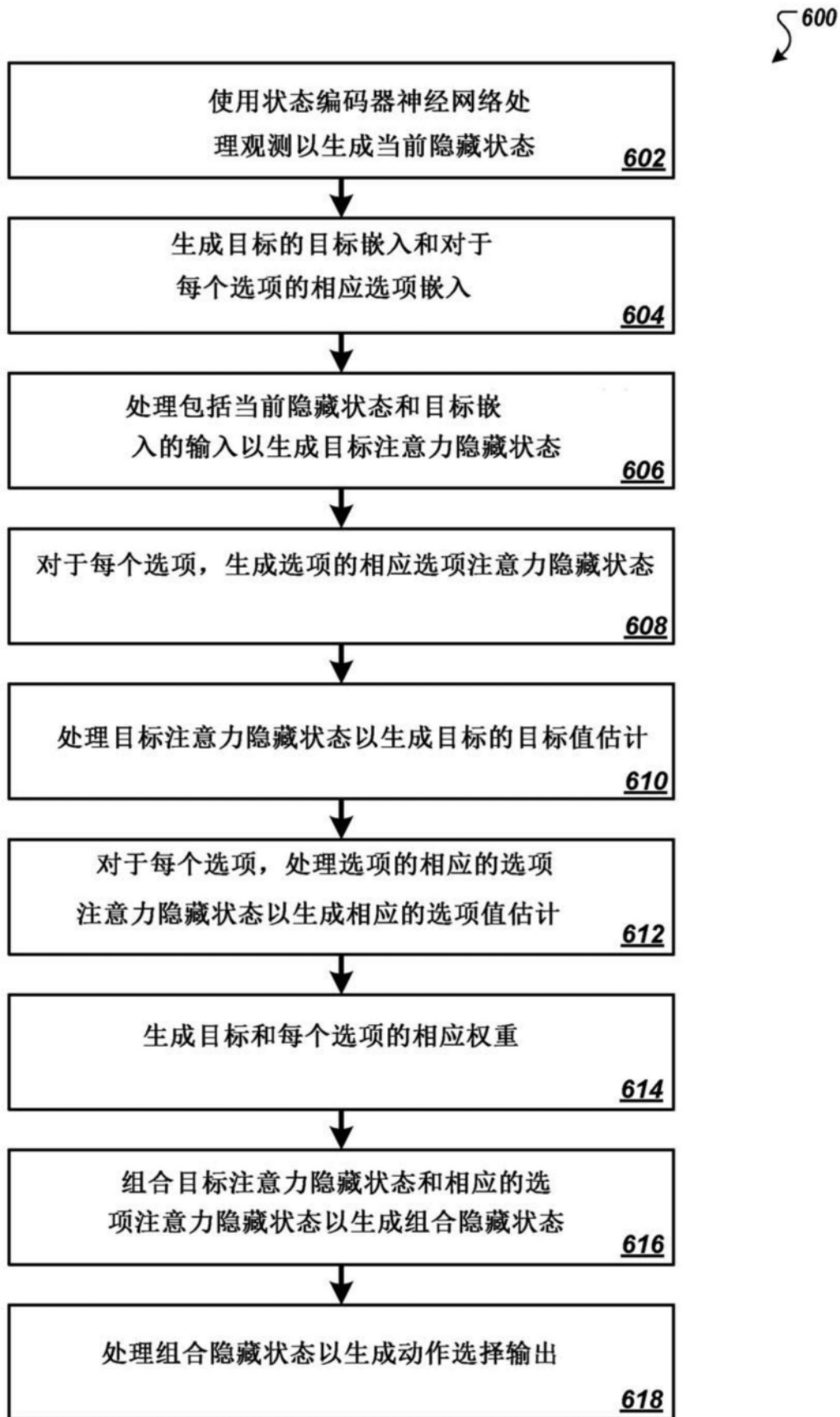


图6

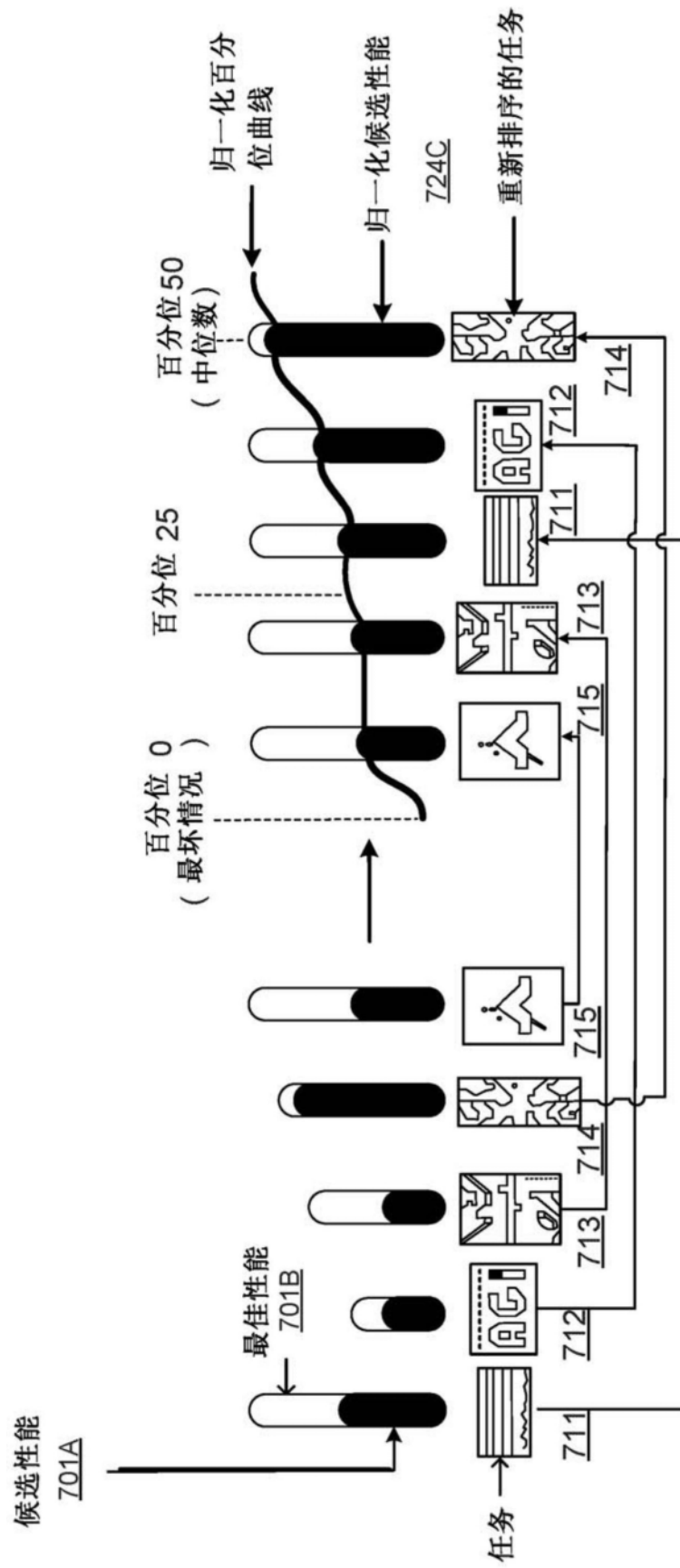


图7