(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 112154461 B
(45)授权公告日 2025.02.25

(21)申请号 201980032895.8

(22)申请日 2019.05.20

(65)同一申请的已公布的文献号
申请公布号 CN 112154461 A

(43)申请公布日 2020.12.29

(30)优先权数据
62/673,812 2018.05.18 US

(85)PCT国际申请进入国家阶段日
2020.11.16

(86)PCT国际申请的申请数据
PCT/EP2019/062943 2019.05.20

(87)PCT国际申请的公布数据
WO2019/219969 EN 2019.11.21

(73)专利权人 渊慧科技有限公司
地址 英国伦敦

(72)发明人 H.宋 A.塔切蒂 P.W.巴塔格利亚 V.扎姆巴尔迪

(74)专利代理机构 北京市柳沈律师事务所 11105
专利代理师 金玉洁

(51)Int.Cl.
G06N 3/042(2023.01)
G06N 3/0455(2023.01)
G06N 3/0442(2023.01)
G06N 3/0464(2023.01)
G06N 3/092(2023.01)
G06N 3/09(2023.01)

(56)对比文件

Thomas Kipf et al..Neural Relational Inference for Interacting Systems.arXiv:1802.04687.2018,1-8、11-17.

Tingwu Wang et al..NERVENET: LEARNING STRUCTURED POLICY WITH GRAPH NEURAL NETWORKS.6TH INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS.2018,摘要,图1-2,第1-3段.

Hoang M. Le et al..Coordinated Multi-Agent Imitation Learning.Proceedings of the 34 th International Conference on Machine Learning.2017,摘要,图1,第1段.

Ryan Lowe et al..Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments.ARXIV.ORG,CORNELL UNIVERSITY LIBRARY.2017,摘要,图1-2,第1-5段.

审查员 李琳

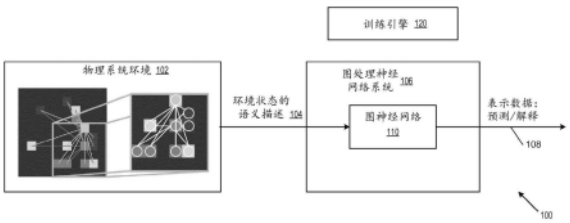
权利要求书4页 说明书14页 附图6页

(54)发明名称

用于多代理环境中的行为预测和强化学习的图神经网络系统

(57)摘要

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,用于预测在具有多个代理的环境中的代理的动作或对所述代理的影响,特别是用于强化学习。在一个方面,关系前向模型(RFM)系统接收表示多个代理中的每个代理的代理动作的代理数据,并实现:编码器图神经网络子系统,将代理数据处理为图数据以提供编码后的图数据;循环图神经网络子系统,处理编码后的图数据以提供处理后的图数据;解码器图神经网络子系统,解码处理后的图数据以提供解码后的图数据;以及输出,提供与一个或更多个代理的预测动作有关的、解码后的图数据的节点属性和/或边属性的表示数据。强化学习系统包括RFM系统。



CN 112154461 B

1.一种计算机实现的强化学习系统,包括基于共享环境中的多个代理的行为的观察、通过有监督训练所训练的用于预测或解释在共享环境中的多个代理的动作的神经网络系统,其中所述强化学习系统被配置为选择将由与所述共享环境交互的代理之一执行的动作,其中所述共享环境是真实世界环境,并且其中所述代理是机械代理,或者包括执行用于控制制造或服务设施中的设备物件以及用于一起工作以控制所述设备物件的动作的控制设备,或者所述代理通过执行将任务分配给特定计算资源的动作来管理跨计算资源的任务分布,或者所述共享环境是包括运动模拟环境的模拟环境,并且所述代理是在所述模拟环境中导航的模拟运载工具,该神经网络系统包括:

输入,用于接收代表多个代理中的每个代理的代理动作的代理数据;以及

一个或多个处理器,被配置为实现:

编码器图神经网络子系统,将代理数据作为图数据进行处理以提供编码后的图数据,

其中,代理数据代表多个代理中每一个的代理动作;

其中,图数据包括(i)至少表示图的节点和边的数据以及(ii)图中至少一些节点的节点属性,

其中,节点表示环境中的代理和一个或更多个非代理实体,

其中,边连接图中的节点,

其中,节点属性表示代理的代理动作,以及

其中,编码后的图数据包括表示图数据的更新后的版本的节点属性和边属性;

循环图神经网络子系统,包括处理编码后的图数据并提供处理后的图数据的循环神经网络,该处理后的图数据包括编码后的图数据的节点属性和边属性的更新后的版本;

解码器图神经网络子系统,解码处理后的图数据并提供解码后的图数据,该解码后的图数据包括处理后的图数据的节点属性和边属性的更新后的版本;以及

系统输出,提供表示数据,该表示数据包括针对一个或更多个代理的、解码后的图数据的节点属性和边属性之一或两者的表示,其中该表示与所述一个或更多个代理的预测动作或解释动作有关,

其中,所述强化学习系统包括:输入,用于获得并处理表示共享环境的状态的观察的状态数据和表示从代理执行动作的结果接收到的奖励的奖励数据,所述奖励数据隐含地定义了要执行的任务;策略神经网络;以及组合器,用于组合所述表示数据和所述状态数据以提供增强的观察,其中所述策略神经网络被配置为处理所述增强的观察以选择动作。

2.根据权利要求1所述的强化学习系统,其中,表示代理动作的代理数据包括多个代理中的每个代理的代理位置和运动数据,以及其中,用于确定每个代理的动作的节点属性还包括每个代理的位置和运动的属性。

3.根据权利要求1所述的强化学习系统,其中,每个代理通过边连接到每个其它代理,以及其中,每个非代理实体通过边连接到每个代理。

4.根据权利要求1所述的强化学习系统,其中,系统输出包括组合解码后的图数据中的节点的节点属性以输出表示数据的一个或更多个输出神经网络层,以及其中,该表示包括该节点所表示的代理的预测动作。

5.根据权利要求4所述的强化学习系统,其中,表示数据定义从表示一个或更多个代理的一个或更多个节点的节点属性得出的数据的空间图,以及其中,在空间图中,在相应节点

的位置处或邻近相应节点的位置表示从节点属性得出的数据。

6.根据权利要求1所述的强化学习系统,其中,表示数据包括针对连接到一个或更多个节点的边的、解码后的图数据的边属性的表示,以及其中,从边的边属性的组合中确定用于该边的边属性的表示。

7.根据权利要求6所述的强化学习系统,其中,表示数据定义空间图,以及其中,在空间图中,在边的起始节点位置处表示用于该边的边属性的表示。

8.根据权利要求1所述的强化学习系统,其中,编码器图神经网络子系统、循环图神经网络子系统和解码器图神经网络子系统中的一个或更多个配置为:

对于每个边,使用边神经网络来处理边特征以确定输出边特征,

对于每个节点,聚合连接到该节点的边的输出边特征以确定该节点的聚合边特征,以及

对于每个节点,使用节点神经网络来处理聚合边特征和节点特征以确定输出节点特征。

9.根据权利要求8所述的强化学习系统,其中,处理边特征包括:对于每个边,将边特征和由该边连接的节点的节点特征提供到边神经网络以确定输出边特征。

10.根据权利要求8所述的强化学习系统,其中,编码器图神经网络子系统、循环图神经网络子系统和解码器图神经网络子系统中的一个或更多个还配置为使用全局特征神经网络来确定全局特征向量,全局特征向量表示输出边特征和输出节点特征,以及其中,随后的图神经网络子系统配置为当确定输出边特征和输出节点特征时处理全局特征向量。

11.一种使用基于共享环境中的多个代理的行为的观察、通过有监督训练所训练的神经网络系统预测或解释在共享环境中的多个代理的动作的方法,其中所述代理与所述共享环境交互,其中所述共享环境是真实世界环境,并且其中所述代理是机械代理,或者包括执行用于控制制造或服务设施中的设备物件以及用于一起工作以控制所述设备物件的动作的控制设备,或者所述代理通过执行将任务分配给特定计算资源的动作来管理跨计算资源的任务分布,或者所述共享环境是包括运动模拟环境的模拟环境,并且所述代理是在所述模拟环境中导航的模拟运载工具,该方法包括:

接收表示多个代理中的每个代理的动作的代理数据;

将代理数据处理为图数据以提供编码后的图数据,

其中,图数据包括至少表示图的节点和边的数据,其中,每个代理由节点表示,其中,环境中的非代理实体每个由节点表示,其中,节点具有用于确定每个代理的动作的节点属性,其中,边将代理彼此连接并将代理连接到非代理实体,以及其中,编码后的图数据包括表示图数据的更新后的版本的节点属性和边属性;

使用循环图神经网络来处理编码后的图数据,以提供处理后的图数据,该处理后的图数据包括编码后的图数据的节点属性和边属性的更新后的版本;

解码处理后的图数据以提供解码后的图数据,该解码后的图数据包括处理后的图数据的节点属性和边属性的更新后的版本;以及

输出一个或更多个代理的解码后的图数据的节点属性和边属性之一或两者的表示,其中,该表示与该一个或更多个代理的预测行为或解释行为有关,

其中,所述方法还包括:

获得并处理表示共享环境的状态的观察的状态数据和表示从代理执行动作的结果接收到的奖励的奖励数据,所述奖励数据隐含地定义了要执行的任务;

组合所述表示数据和所述状态数据以提供增强的观察;

处理所述增强的观察以选择动作。

12.根据权利要求11所述的方法,其中,所述行为包括代理的动作,以及其中,输出所述表示包括处理解码后的图数据的节点的节点属性以确定由该节点所表示的代理的预测动作。

13.根据权利要求11所述的方法,其用于解释代理的动作,其中,输出所述表示包括处理解码后的图数据的将影响节点连接到代理节点的边的边属性以确定表示影响节点对于代理节点的重要性的数据。

14.一种或更多种非易失性计算机可读存储介质,其存储指令,所述指令当由一台或更多台计算机运行时使所述一台或更多台计算机实现包括通过有监督训练所训练的神经网络系统的强化学习系统,所述神经网络系统包括:

输入,用于接收代表多个代理中的每个代理的代理动作的代理数据;

编码器图神经网络子系统,将代理数据处理为图数据以提供编码后的图数据,

其中,代理数据代表多个代理中每一个的代理动作;

其中,图数据包括(i)至少表示图的节点和边的数据以及(ii)图中至少一些节点的节点属性,

其中,节点表示环境中的代理和一个或更多个非代理实体,

其中,边连接图中的节点,

其中,节点属性表示代理的代理动作,以及

其中,编码后的图数据包括表示图数据的更新后的版本的节点属性和边属性;

循环图神经网络子系统,包括处理编码后的图数据并提供处理后的图数据的循环神经网络,该处理后的图数据包括编码后的图数据的节点属性和边属性的更新后的版本;

解码器图神经网络子系统,解码处理后的图数据并提供解码后的图数据,该解码后的图数据包括处理后的图数据的节点属性和边属性的更新后的版本;以及

系统输出,提供表示数据,该表示数据包括针对一个或更多个代理的、解码后的图数据的节点属性和边属性之一或两者的表示,其中该表示与一个或更多个代理的预测动作或解释动作有关,

其中,所述强化学习系统包括:输入,用于获得并处理表示共享环境的状态的观察的状态数据和表示从代理执行动作的结果接收到的奖励的奖励数据,所述奖励数据隐含地定义了要执行的任务;策略神经网络;以及组合器,用于组合所述表示数据和所述状态数据以提供增强的观察,其中所述策略神经网络被配置为处理所述增强的观察以选择动作,

其中所述强化学习系统被配置为选择将由与所述共享环境交互的代理之一执行的动作,其中所述共享环境是真实世界环境,并且其中所述代理是机械代理,或者包括执行用于控制制造或服务设施中的设备物件以及用于一起工作以控制所述设备物件的动作的控制设备,或者所述代理通过执行将任务分配给特定计算资源的动作来管理跨计算资源的任务分布,或者所述共享环境是包括运动模拟环境的模拟环境,并且所述代理是在所述模拟环境中导航的模拟运载工具。

15.根据权利要求14所述的一种或更多种非易失性计算机可读存储介质,其中,表示代理动作的代理数据包括多个代理中的每个代理的代理位置和运动数据,以及其中,用于确定每个代理的动作的节点属性还包括每个代理的位置和运动的属性。

16.根据权利要求14所述的一种或更多种非易失性计算机可读存储介质,其中,每个代理通过边连接到每个其它代理,以及其中,每个非代理实体通过边连接到每个代理。

17.根据权利要求14所述的一种或更多种非易失性计算机可读存储介质,其中,系统输出包括组合解码后的图数据中的节点的节点属性以输出表示数据的一个或更多个输出神经网络层,以及其中,该表示包括该节点所表示的代理的预测动作。

18.根据权利要求17所述的一种或更多种非易失性计算机可读存储介质,其中,表示数据定义从表示一个或更多个代理的一个或更多个节点的节点属性得出的数据的空间图,以及其中,在空间图中,在相应节点的位置处或邻近相应节点的位置表示从节点属性得出的数据。

19.根据权利要求14所述的一种或更多种非易失性计算机可读存储介质,其中,表示数据包括针对连接到一个或更多个节点的边的、解码后的图数据的边属性的表示,以及其中,从边的边属性的组合中确定用于该边的边属性的表示。

20.根据权利要求19所述的一种或更多种非易失性计算机可读存储介质,其中,表示数据定义空间图,以及其中,在空间图中,在边的起始节点位置处表示用于该边的边属性的表示。

用于多代理环境中的行为预测和强化学习的图神经网络系统

背景技术

[0001] 本说明书涉及用于预测具有多个代理的环境中的代理的动作或对所述代理的影响的神经网络,特别是用于强化学习的神经网络。

[0002] 神经网络是采用一层或更多层非线性单元来预测接收到的输入的输出的机器学习模型。一些神经网络除了包括输出层之外还包括一个或更多个隐藏层。每个隐藏层的输出用作网络中下一层的输入,即下一个隐藏层或输出层。网络的每一层根据相应参数集的当前值从接收到的输入生成输出。

[0003] 一些神经网络表现出包括由边连接的节点的图结构;图可以是其中节点可由多个边连接的多图(multigraph)。节点和边可以具有相关联的节点特征和边特征;这些可以使用可由神经网络实现的节点函数和边函数来更新。

发明内容

[0004] 该说明书描述了神经网络系统和方法,该神经网络系统和方法被实现为在一个或更多个位置中的一台或更多台计算机上用于处理表示多个代理的行为的数据的计算机程序,以用于预测代理的动作或用于确定对代理的运作的影响。代理可以是工厂车间的机器人、或者自主或半自主运载工具。所描述的神经网络系统可以用于强化学习,例如,以通过预期其它代理的动作来改善性能,或者用于学习合作行为。

[0005] 因此,在一个方面,一种用于预测或解释在共享环境中的多个代理的动作的关系前向模型(RFM)神经网络系统包括:输入,接收表示多个代理中的每个代理的代理动作的代理数据;以及一个或更多个处理器。所述一个或更多个处理器配置为实现将代理数据处理为图数据(结合图数据处理)以提供编码后的图数据,图数据可以包括至少表示图的节点和边的数据;边可以是有向的或无向的。每个代理可以由节点表示。环境中的非代理实体每个也可以由节点表示。节点具有例如用于确定每个代理的动作的节点属性。节点可以每个具有相同属性的集合。提供给编码器的图数据可以缺少边属性。边可以将代理彼此连接以及将代理连接到非代理实体。编码后的图数据可以包括表示图数据的更新后的版本的节点属性和边属性。

[0006] 所述一个或更多个处理器还可以配置为实现处理图神经网络子系统,特别是循环图神经网络子系统。循环/处理图神经网络子系统可以包括处理编码后的图数据并提供处理后的图数据的循环神经网络,处理后的图数据包括编码后的图数据的节点属性和边属性的更新后的版本。所述一个或更多个处理器还可以配置为实现解码处理后的图数据并提供解码后的图数据的解码器图神经网络子系统,解码后的图数据包括处理后的图数据的节点属性和边属性的更新后的版本。该系统可以具有提供表示数据的系统输出,该表示数据包括针对一个或更多个代理(例如,所有)代理的、解码后的图数据的节点属性和边属性之一或两者的表示。该表示可以与一个或更多个代理的例如分别从解码后的图数据的节点属性或边属性得出的预测动作或解释动作有关。

[0007] 在一些实施方式中,动作可以包括代理的运动。因此,由系统捕获的代理数据可以

包括每个代理的代理位置和运动数据。于是节点属性可以包括每个代理的位置和运动的属性。在一些实施方式中,每个代理可以通过边连接到每个其它代理,并且每个非代理实体可以通过边连接到每个代理。然而,在实施方式中,非代理实体例如静止实体没有通过边连接。

[0008] 在一些实施方式中,系统输出包括一个或更多个输出神经网络层,例如多层感知器(MLP)。节点属性可以被表示为向量。所述一个或更多个输出神经网络层可以组合解码后的图数据中的节点的节点属性,以便输出表示数据。从节点得出的表示可以包括由该节点表示的代理的预测动作。表示数据可以定义从表示一个或更多个代理的一个或更多个节点的节点属性得出的数据的空间图,例如热图。在这样的图中,可以在相应节点的位置处表示从节点属性得出的数据。例如,在动作可以包括代理的移动的情况下,图可以表示所表示的每个代理处于某个位置的概率。

[0009] 在一些实施方式中,表示数据包括针对连接到一个或更多个节点的边的、解码后的图数据的边属性的表示。可以从边的边属性的组合中确定该边的边属性的表示。例如,边属性可以被表示为向量,并且边的边属性的组合可以是向量范数,诸如p范数,其中p是整数。

[0010] 在实施方式中,边是有方向的(尽管节点可以通过边在两个相反的方向上连接)。因此,边可以从起始节点连接到末端节点,例如从代理或非代理实体节点连接到代理节点。边的表示,例如向量范数,可以表示起始节点对它所连接的代理节点的重要性或影响。

[0011] 边表示数据可以定义空间图,诸如热图。在空间图中,边的边属性的表示,例如属性的向量范数,可以位于边的起始节点位置。这样的空间图可以基于每个代理来定义,即对于所考虑的每个代理可以存在一个图。

[0012] 假设边的表示,例如向量范数,表示边的重要性。因此,例如,通过指示哪些节点影响了动作或通过识别哪些其它(多个)节点对特定代理最有影响力(例如,通过对其它节点进行排名),解码后的图的边属性可以对可用于解释代理行为的信息进行编码。该解释信息的变化可以随着时间的推移被追踪。

[0013] 如上所述的神经网络系统可以通过有监督训练,例如,基于对共享环境中多个代理的行为的观察来训练。

[0014] 在如上所述的包括编码器图神经网络子系统、处理图神经网络子系统和解码器图神经网络子系统中的一个或更多个的神经网络系统中,编码器图神经网络子系统、处理图神经网络子系统和解码器图神经网络子系统中的一个或更多个可以配置为实现如下的图网络处理:对于每个边,使用边神经网络来处理边特征以确定输出边特征。对于每个节点,聚合连接到该节点的边的输出边特征以确定该节点的聚合边特征。对于每个节点,使用节点神经网络来处理聚合边特征和节点特征以确定输出节点特征。该过程可以执行一次或迭代地执行多次。处理边特征可以包括:对于每个边,将边特征和由该边连接的节点的节点特征提供给边神经网络以确定输出边特征。

[0015] 编码器图神经网络子系统、处理图神经网络子系统和解码器图神经网络子系统中的一个或更多个还可以配置为使用全局特征神经网络来确定全局特征向量。全局特征向量可以表示输出边特征和输出节点特征。在编码器图神经网络子系统确定全局特征向量的情况下,随后的处理图神经网络子系统和解码器图神经网络子系统也可以对全局特征向量进

行操作。图神经网络子系统,诸如处理(循环)图神经网络子系统,可以包括循环图网络。于是,边神经网络、节点神经网络和全局特征神经网络(如下所述)中的一个或更多个可以包括循环神经网络,例如GRU(门控循环单元)神经网络。

[0016] 在一些实施方式中,该系统可以被包括在强化学习系统中。强化学习系统可以配置为选择将要由与共享环境交互的代理之一执行的动作。强化学习系统可以包括获得状态数据和奖励数据的输入,状态数据表示共享环境的状态,奖励数据表示由于代理执行动作而收到的奖励。例如,可以通过从环境捕获一个或更多个观察(诸如图像)并使用观察处理神经网络(诸如卷积神经网络)处理这些观察而得出状态数据。强化学习系统还可以包括处理状态数据和奖励数据以选择动作的动作选择策略神经网络。动作选择策略神经网络可以配置为接收和处理表示数据以选择动作。

[0017] 观察还可以用于训练系统以预测/解释代理的动作。例如,观察处理神经网络或另一个神经网络可以配置为识别环境中的代理的动作。观察处理神经网络或另一个神经网络可以另外地或备选地配置为识别代理和/或非代理实体,例如用于根据存在的实体来适配图。

[0018] 一般,强化学习系统可以是任何类型。例如,它可以是直接将策略和价值函数参数化的基于策略的系统,诸如优势演员评论家(Advantage Actor Critic,A2C或A3C)系统(例如,Mnih et al.2016)。备选地,它可以是其中输出近似于动作-价值函数以及可选地近似于状态的值以用于确定动作的Q学习系统,诸如深度Q学习网络(DQN)系统或Double-DQN系统。在另一备选方案中,它可以是分布式强化学习系统,诸如IMPALA(重要性加权演员学习者,Importance-Weighted Actor-Learner),Espholt et al.,arXiv:1802.01561。在连续控制设置中,它可以例如使用DDPG(深度确定性策略梯度)技术(arXiv 1509.02971)或其变体直接输出动作,例如转矩或加速度值。

[0019] 在一些实施方式中,强化学习系统可以配置为训练神经网络系统,以在强化学习系统的训练期间从对共享环境的观察中预测或解释代理的动作。因此,可以共同训练两个系统。用于预测或解释代理的行为的神经网络系统可能是强化学习系统专有的,即,它不需要依赖于对强化学习系统不可用的数据进行训练。在一些实施方式中,这两个系统可以共同位于代理之一中,并且可以彼此相伴地学习。

[0020] 可选地,图网络处理可以包括由神经网络实现的自引用函数。例如,节点更新可以取决于节点的初始特征。这样的处理在促进代理预测其自己的动作方面可以是有利的,促进代理预测其自己的动作方面可以增强动作选择所基于的信息。可选地,可以针对多个时间步推出用于预测或解释代理的动作的神经网络系统,以为强化学习系统提供附加信息。因此,该系统可以用于基于想象力规划(Weber et al.,arXiv:1707.06203)。

[0021] 还描述了一种预测或解释在共享环境中的多个代理的动作的方法。该方法可以包括接收表示多个代理中的每个代理的动作的代理数据。该方法还可以包括结合图数据处理代理数据以提供编码后的图数据。图数据可以包括至少表示图的节点和边的数据,其中,每个代理由节点表示,其中,环境中的非代理实体每个由节点表示,其中,节点具有用于确定每个代理的动作的节点属性,其中,边将代理彼此连接并将代理连接到非代理实体,以及其中,编码后的图数据包括表示图数据的更新后的版本的节点属性和边属性。该方法还可以包括使用循环图神经网络来处理编码后的图数据以提供处理后的图数据,该处理后的图数

据包括编码后的图数据的节点属性和边属性的更新后的版本。该方法还可以包括解码处理后的图数据以提供解码后的图数据,该解码后的图数据包括处理后的图数据的节点属性和边属性的更新后的版本。该方法还可以包括输出针对一个或更多个代理的、解码后的图数据的节点属性和边属性之一或两者的表示,其中,该表示与代理的预测行为或解释行为有关。

[0022] 所述行为可以包括代理的动作。输出所述表示可以包括处理解码后的图数据的节点的节点属性,以确定该节点所表示的代理的预测动作。

[0023] 该方法可以用于解释代理的动作。然后,输出所述表示可以包括处理解码后的图数据的边的边属性(该边将可以是代理或非代理节点的影响(起始)节点连接到代理节点),以确定表示该影响节点对代理节点的重要性的数据。

[0024] 该方法可以用于强化学习。例如,强化学习可以取决于由多个代理共享的环境的状态来学习用于选择将要由代理执行的动作的策略。强化学习方法可以包括使用所描述的方法来预测环境中的代理的行为、然后使用所预测行为来学习策略。

[0025] 一般,所描述的系统和方法的实施方式可以与真实或模拟的代理和环境一起使用。环境可以是真实环境,代理可以是机械代理,诸如与真实环境交互的机器人。在一些其它实施方式中,代理可以包括在制造或服务设施中操作并且一起工作以控制设施的方面(例如,服务器场的操作温度)的控制设备。稍后描述其它示例。

[0026] 本说明书中描述的主题可以在特定实施例中实现,以实现以下优点中的一个或更多个。

[0027] 所描述的神经网络系统的一些实施方式能够表征多个自主代理的行为,诸如在工厂中的机器人或者自主或半自主运载工具。于是可以使用该信息来训练代理以更有效或更高效地完成任务。例如,可以训练诸如机械代理的代理以凭借改善的性能和/或使用较少的数据和处理资源来完成任务。更具体地,与一些其它技术相比,训练更快并且使用更少的数据和计算资源,并且所描述的系统可以预测动作和影响两者,这使得他们能够在其它技术可能发现困难的情况下学习,从而促进更好的性能。

[0028] 所描述的神经网络系统也可以由代理使用,使得该代理可以与其它代理合作来执行任务。其它代理可以是或可以不是同一类型;例如,它们可以是人类控制的或计算机控制的,或者它们可以来自不同的制造商。这种合作行为的一个示例是对工厂或仓库地面上的多个机器人的控制;为这样的机器人配备这里所述类型的神经网络系统可以促进改善的机器人控制,因为可以将其它机器人的行为纳入考虑。在另一示例中,自主或半自主运载工具可以使用所描述类型的神经网络系统来预测其它道路使用者或行人的行为,从而提高安全性。

[0029] 重要的是,因为所描述的用于预测/解释多个代理的动作的系统需要较少的经验来学习,所以即使其它代理仍在学习并且其行为在改变,所描述的系统也可以在与强化学习系统一起训练时取得良好的结果。

[0030] 在一些其它应用中,所描述的神经网络系统可以提供与代理的动机或意图有关的信息。即,从处理后的图的边得出的信息可以用于确定代理为何以这种方式采取行动或它可能如何行动。这可以深入了解在机器学习系统控制下的自主代理行动的方式。该信息对于监管和其它目的可以是有用的,因为如果可以识别出自主系统的行为动机,则更容易信

任此类系统。

[0031] 将一台或更多台计算机的系统配置为执行特定操作或动作意味着该系统已在其上安装了在操作中使该系统执行所述操作或动作的软件、固件、硬件或其组合。将一个或更多个计算机程序配置为执行特定操作或动作意味着所述一个或更多个程序包括指令,所述指令当由数据处理设备运行时使该设备执行所述操作或动作。

附图说明

[0032] 图1示出了用于预测或解释在共享物理系统环境中的多个代理的动作的示例神经网络系统。

[0033] 图2a和图2b示出了图神经网络子系统和该图神经网络子系统的操作。

[0034] 图3示出了用于预测或解释在共享物理系统环境中的多个代理的动作的图处理神经网络系统的细节。

[0035] 图4示出了使用图3的图处理神经网络系统来提供预测代理动作的表示数据的过程。

[0036] 图5a和图5b示出了代理在网格(上部)和(下部)上的定位,分别为动作预测的示例热图和代理-实体重要性关系的示例热图。

[0037] 图6示出了包括结合图3的图处理神经网络系统的强化学习系统的代理。

[0038] 在各个附图中,相同的附图标记和名称表示相同的元素。

具体实施方式

[0039] 图1示出了用于预测或解释在共享物理系统环境102中的多个代理的动作的示例神经网络系统100。神经网络系统100是被实现为在一个或更多个位置中的一台或更多台计算机上的计算机程序的系统的示例,可以在该计算机程序中实现下面描述的系统、组件和技术。

[0040] 代理可以包括机械代理,诸如在真实或模拟环境中的机器人或运载工具,并且系统可以例如出于安全或控制目的而预测或解释它们的行为。例如,神经网络系统100可以在机器人或运载工具控制系统中使用,以根据对一个或更多个其它代理(其可以是其它机器人/运载工具和/或人员)的行为的预测或解释来控制机器人/运载工具的行为。控制系统可以包括强化学习系统。对特定代理的行为的解释可以包括定义环境中的一个或更多个其它代理或非代理实体对该特定代理的重要性或影响的数据。该数据可以类似地被控制系统用来响应于对一个或更多个其它代理(例如,其它机器人/运载工具或人员)的行为的解释而控制代理。稍后给出进一步的示例。

[0041] 环境中的非代理实体可以包括:要收集、操纵、移动和/或落下的对象;静态对象或实体,诸如障碍物或标记;定义允许的或优选的移动或位置的实体,例如路线、车道、通道、大道、停车位(其可能被占用或未被占用)等;提供指令或控制命令的实体,例如标牌等。

[0042] 神经网络系统100接收环境104的状态的被格式化为如下面解释的图的语义描述,并处理数据以提供输出,该输出包括用于预测和/或解释环境中代理的行为的表示数据。神经网络系统100包括图处理神经网络系统106,其包括处理输入图从而以输出图的形式提供输出的数据108的一个或更多个图神经网络块110,可以从该输出图中得出表示数据以预测

和/或解释代理的行为。在一些实施方式中,例如增强学习应用,神经网络系统100可以包括训练引擎120以训练图处理神经网络系统106,如稍后所述。例如,在强化学习应用中,代理可以包括图处理神经网络系统106,以利用对其它代理的行为的预测来增强对环境的观察。

[0043] 环境状态的语义描述可以是这样一种描述,其中关于环境的数据是显式的而不是例如从观察推断出的。环境状态的语义描述可以包括代理数据以及与非代理(例如,静态实体)有关的数据。对于每个代理和非代理实体,并且在适用的情况下,该数据可以包括以下一项或更多项:位置,移动(例如,速度向量),类型(例如,独热编码(one-hot encoded)),状态或配置,以及(对于代理而言)代理所采取的最后动作。环境状态的语义描述也可以包括全局信息,诸如与环境相关联的得分或奖励,诸如游戏中的得分或与完成一项或更多项任务相关联的得分或奖励。

[0044] 环境状态的语义描述被编译成表示环境状态的图数据,作为输入图G_{in}^{t},特别是有向图。在输入图中,代理和非代理实体中的每个可以由节点表示,并且边可以将每个代理连接到每个其它代理以及连接到每个非代理实体(非代理实体不需要被连接)。环境状态的语义描述可以用于提供输入图的节点的属性。输入图可以没有边的属性;可以在输入图的边属性中指示代理合作/非合作(队友)属性。输出图具有边属性;这些边属性中的每个提供发送方节点对接收方节点的影响的表示。

[0045] 图2a示出了图神经网络子系统110。该子系统接受输入有向图G,该输入有向图G包括:一组节点特征$\{\boldsymbol{n}_i\}_{i=1\ldots N_n}$,其中$N_n$是节点的数量,每个$n_i$是节点特征的向量;一组有向边特征$\{e_j, s_j, r_j\}_{j=1\ldots N_e}$,其中$N_e$是边的数量,每个$e_j$是边特征的向量,$s_j$和$r_j$分别是发送方和接收方节点的索引;以及全局特征的向量g。对于输入图,可以从节点属性(诸如代理位置和速度)得出节点特征的向量;在属性不适用的情况下,例如非代理实体的状态,它们可以被赋予零值。对于输入图,边可以没有属性,并且全局特征的向量可以在适用时表示全局得分或奖励。

[0046] 图神经网络子系统110处理输入图$G=(g,\{n_i\},\{e_j,s_j,r_j\})$以确定输出图$G^{*}=\left(\boldsymbol{g}^{*},\{\boldsymbol{n}_i^{*}\},\left\{\boldsymbol{e}_j^{*},s_j,r_j\right\}\right)$。一般,尽管不是必须的,但是输入图和输出图可以具有不同的特征。

[0047] 图神经网络子系统110具有三个子函数,边式函数(edge-wise function)f_e、节点式函数(node-wise function)f_n和全局函数f_g。这些中的每个利用不同的相应神经网络(即具有不同参数(权重)的神经网络,即分别是边神经网络、节点神经网络和全局特征网络)来实现。在变型中,可以省略一些神经网络和/或稍后描述的更新。

[0048] 在一些实施方式中,这些函数中的每个利用相应的多层感知器(MLP)来实现。在一些实施方式中,可以使用循环神经网络来实现这些函数中的一个或更多个。在这种情况(未示出)下,函数即循环神经网络将另外的隐藏图G_h当作输入,并另外提供更新后的隐藏图G_h^{*}作为输出。输入图和隐藏图可以例如使用GRU(门控循环单元)样式或LSTM(长短期记忆)样式门控方案被组合。例如通过复制,可以分割输出以获得更新后的隐藏图G_h^{*}和输出图。

[0049] 在实施方式中,图神经网络子系统110配置为通过首先应用边式函数f_e以更新(在

每个指定方向上的)所有边、然后应用节点式函数f_n以更新所有节点、以及最后应用全局函数f_g以更新全局特征向量来处理输入图。

[0050] 图2b示出了示例图神经网络子系统110的操作。在步骤200,对于每个边{e_j,s_j,r_j},该过程收集发送方和接收方节点特征$\boldsymbol{n}_{s_i}$、$\boldsymbol{n}_{r_i}$以及边特征向量和全局特征向量,并使用边神经网络计算输出边向量$\boldsymbol{e}_i^* = f_e\left(\boldsymbol{g},\boldsymbol{n}_{s_i},\boldsymbol{n}_{r_i},\boldsymbol{e}_i\right)$。然后在步骤202,对于每个节点{$n_i$},该过程使用聚合函数将该节点的边向量聚合为接收方,以确定一组聚合的边特征$\hat{\boldsymbol{e}}_i$。聚合函数关于边向量的排列应该是不变的。例如,它可以包括平均值、最大值或最小值的确定。在一些实施方式中,聚合函数可以包括针对边的元素式求和(elementwise summation),例如$\hat{\boldsymbol{e}}_i = \Sigma_j \boldsymbol{e}_j^*$,其中$r_j$=i。然后,使用节点神经网络$\boldsymbol{n}_i^* = f_n(\boldsymbol{g},\boldsymbol{n}_i,\hat{\boldsymbol{e}}_i)$从所述一组聚合的边特征、当前节点特征向量和全局特征向量中计算输出节点向量$\boldsymbol{n}_i^*$。最后,对于每个图,该过程在步骤204例如通过元素式求和$\hat{\boldsymbol{e}} = \Sigma_j \boldsymbol{e}_j^*$、$\hat{\boldsymbol{n}} = \Sigma_i \boldsymbol{n}_i^*$聚合所有边特征向量和所有节点特征向量,以及当前全局特征向量g,使用全局特征神经网络$\boldsymbol{g}^* = f_g(\boldsymbol{g},\hat{\boldsymbol{n}},\hat{\boldsymbol{e}})$计算输出全局特征向量$g^*$。

[0051] 图3显示了图处理神经网络系统106的实施方式。该实施方式包括耦合到循环图神经网络子系统(例如,图GRU)304的编码器图神经网络子系统(GN编码器)302,该图GRU 304耦合到解码器图神经网络子系统(GN解码器)306。每个图神经网络子系统可以如参照图2所述的那样。时间步t处的输入图G_{in}^t由GN编码器302处理,以提供编码后的图数据。编码后的图数据由循环图神经网络子系统304连同先前时间步的隐藏图G_{hid}^{t-1}处理,以提供时间步的更新后的隐藏图G_{hid}^t和处理后的图数据,其可称为潜图(latent graph)。潜图由GN解码器306解码,以提供时间步的输出图G_{out}^t。

[0052] 可以从输出图的一个或更多个属性直接得出表示数据,或者可以例如通过MLP 308进一步处理输出图以提供表示数据。例如,输出图的每个代理节点的节点属性可以由MLP 308(其可以是单层"MLP")处理,以将这一组节点属性减少为一组输出值用于预测代理的动作。这一组输出值可以包括例如代理的可用动作的一组对数,或者可以定义例如连续动作的分布。一般,图处理神经网络系统106可以应用于离散的或连续的动作。

[0053] 图处理神经网络系统106可以通过共同训练图神经网络子系统来训练;图神经网络子系统可以基于任何输出属性来训练。基于输入-输出对的数据集、反向传播任何合适的损失函数的梯度、以及使用通过时间的反向传播来训练循环图神经网络子系统而使用有监督训练。例如,在一个实施方式中,以例如128轮为批次使用梯度下降来使代理的预测动作和地面真值(ground-truth)动作之间的交叉熵损失最小化。

[0054] 图3的包括三个分离的图神经网络子系统、编码器、循环块和解码器的架构允许系统在包括时间循环之前对原始输入数据执行关系推理、然后对循环子系统的输出再次执行关系推理。这允许循环图神经网络子系统构造实体之间关系的记忆,而不仅仅是其当前状态的记忆。

[0055] 在一个示例实施方式中,图3的架构如下:图编码器神经网络子系统302包括单独的64单元MLP,其具有一个隐藏层和针对边、节点和全局神经网络中的每个的ReLU激活;循环图神经网络子系统304的边、节点和全局神经网络中的每个包括隐藏状态大小为32的GRU;解码器神经网络子系统306与编码器相同;聚合函数包括求和。

[0056] 图4示出了使用图处理神经网络系统106来提供预测代理动作的表示数据的过程。对于每个时间步t,该过程输入定义环境状态的(语义)描述的数据,并使用该数据构建输入图(400)。然后,该过程使用GN编码器302对输入图进行编码以提供编码后的图数据(402),利用图GRU 304处理编码后的图数据以提供潜图(404),并使用GN解码器对潜图进行解码以提供输出图(406)。然后,该过程可以从输出图例如通过使用MLP 308和/或如下所述处理输出图来确定表示数据,以预测或解释一个或更多个代理的动作(408)。

[0057] 图5a(下部)示出了来自图处理神经网络系统106的动作预测的作为热图的示例表示。图5a(上部)示出了代理500a、500b在x、y位置网格上的定位。在该示例中,可能的代理动作包括代理不移动和代理移动到网格上四个相邻位置之一(5个动作之一)。下部图示出了对于每个代理位置,代理移动到网格上四个相邻位置中的每个以及不移动的概率。每个位置以针对相应动作的预测对数(例如,使用灰度表示)被编码。

[0058] 图5b(下图)示出了来自图处理神经网络系统106的代理-实体重要性关系的作为热图的示例表示。这些关系可用于解释代理行为,并且可以从输出图的边属性中得出。例如,为了确定特定代理与其它实体之间关系的强度,可以确定将该代理的节点与每个其它实体的节点连接的每个边向量的欧几里得范数。在图5b(下部)中,每个其它实体的位置以特定代理500a与实体之间关系的强度(例如,使用灰度)被编码(图5b上部,与图5a上部相同)。因此,该表示可以指示实体对该特定代理的影响程度。

[0059] 图6示出了代理600,其是在共享环境102中的代理之一。例如,代理600可以是在具有其它机器人或运载工具以及可选地具有非代理实体的环境中的机器人或运载工具。该代理将图3的图处理神经网络系统(关系前向模型,RFM)106合并到强化学习(RL)系统中,在这里将其表示为动作选择策略网络(Π)602。其它代理中的一些、全部或没有任何一个可以包括类似的RL系统。

[0060] 在所示示例中,例如,RFM 106接收以环境图的形式的对环境状态的语义描述,并且还为一个或更多个其它代理(例如队友)接收该代理所采取的最后动作。该信息被编译成输入图(未显示预处理)。在其它实施方式中,该代理可以从对环境的观察中推断该信息中的一些或全部。

[0061] 该代理还接收观察,例如对强化学习系统的环境的自我为中心的观察。观察可以包括静止图像或运动图像,和/或任何形式的传感器数据,和/或表征环境当前状态的任何其它数据。RL系统还从与代理因执行动作而接收的奖励有关的环境接收例如定义数值的奖励数据。来自RFM 106的表示数据输出108被提供给RL系统,并且可以通过组合器604与对环境的观察组合,以提供增强的观察606。

[0062] 表示数据可以包括以如图5a(下部)所示热图的形式的对一个或更多个其它代理的动作的预测。当观察和热图(或其它表示)两者被自我为中心地表示时,动作预测可以包括附于观察的附加图像平面。每个其它代理可以具有单独的附加图像平面,或者可以将图像平面中的数据组合为单个附加图像平面。

[0063] 表示数据还可以或替代地包括一些或所有其它实体对该代理的动作(或对其它代理的动作)的显著性的(例如以图5b(下部)所示热图的形式的)表示。其它实体可以是代理和/或非代理实体。当观察和热图(或其它表示)两者被自我为中心地表示时,其它实体对该代理的影响或其它实体对其它代理的影响可以类似地包括附于观察的附加图像平面。

[0064] 增强学习系统处理增强的观察和奖励数据,并选择将要由代理为了执行任务而执行的动作。任务可以由奖励数据隐含地定义。动作可以从离散的一组动作中选择,也可以是连续的控制动作;稍后给出任务和动作的示例。一般,使用动作选择策略网络602来选择动作;这可以输出用于选择动作的Q值,或者其可以定义用于选择动作的分布,或者可以直接输出动作。可以使用任何适当的强化学习技术(例如,Q学习技术或演员-评论家技术)来训练动作选择策略网络602。

[0065] 更详细地,在一个示例中,动作选择策略网络602的动作选择输出可以包括可由代理执行的一组可能动作中的每个动作的相应数值概率值。RL系统可以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来选择将要由代理执行的动作。

[0066] 在另一示例中,动作选择输出可以例如通过定义应该应用于机器人代理的关节的转矩值或将要应用于机器人或运载工具驾驶的加速度值而直接定义将要由代理执行的动作。

[0067] 在另一示例中,动作选择输出可以包括可由代理执行的一组可能动作中的每个动作的相应Q值。RL系统可以处理Q值(例如,使用soft-max函数)以为每个可能动作生成相应的概率值,该概率值可以用于选择将要由代理执行的动作。RL系统还可以选择具有最高Q值的动作作为将要由代理执行的动作。动作的Q值是对代理响应于当前观察执行该动作并随后根据动作选择策略网络参数的当前值选择由代理执行的将来动作将会产生的回报的估计。这里,回报是指代理收到的奖励的累积度量,例如时间折扣的奖励总和。

[0068] 在一些情况下,RL系统可以根据探索策略选择将要由代理执行的动作。例如,探索策略可以是∈贪心探索策略,其中RL系统以概率1-∈根据动作选择输出选择将要由代理执行的动作,并以概率∈随机选择将要由代理执行的动作。在该示例中,∈是0和1之间的标量值。

[0069] 在实施方式中,训练引擎120可以训练RFM 106和RL系统两者。它们可以被共同训练,即,代理600可以具有与动作选择策略网络602一起被训练的私有RFM 106。可以使用如前所述的有监督学习以预测动作观察动作对来训练RFM 106。如图6所示,在实施方式中,梯度不通过RL系统反向传播到RFM 106中。

[0070] 在一个实施方式中,训练引擎120使用演员-评论家技术来训练动作选择策略网络602。在该实施方式中,动作选择策略网络602配置为除了动作选择输出之外还生成值估计。值估计表示对回报(例如,在给定环境的当前状态的情况下根据动作选择网络参数的当前值选择由代理执行的将来动作将会产生的时间折扣回报)的估计。训练引擎可以使用以下公式给出的强化学习目标函数$\mathcal{L}_{RL}$的梯度来训练动作选择策略网络:

[0071] $$\mathcal{L}_{RL} = \mathcal{L}_{\pi} + \alpha\mathcal{L}_{V} + \beta\mathcal{L}_{H}$$

[0072] $$\mathcal{L}_{\pi} = -\mathbb{E}_{s_t\sim\pi}\left[\hat{R}_t\right]$$

[0073] $$\mathcal{L}_V = \mathbb{E}_{s_t \sim \pi}\left[\left(\hat{R}_t - V(s_t,\theta)\right)^2\right]$$

[0074] $$\mathcal{L}_H = -\mathbb{E}_{s_t \sim \pi}\left[H\left(\pi(\cdot\,|s_t,\theta)\right)\right]$$

[0075] 其中α和β是正常数，$\mathbb{E}_{s_t \sim \pi}[\cdot]$是指相对于当前动作选择策略的期望值(即，由动作选择策略网络参数θ定义)，$V(s_t,\theta)$是指动作选择策略网络为观察s_t而生成的值估计，$H(\pi(\cdot|s_t,\theta))$是正则化项，其是指由动作选择网络为观察s_t生成的在可能动作上的概率分布的熵，$\hat{R}_t$是指例如由以下给出的n步超前收益：

[0076] $$\hat{R}_t = \sum_{i=1}^{n-1} \gamma^i r_{t+i} + \gamma^n V(s_{t+n},\theta)$$

[0077] 其中γ是0与1之间的折扣因子，r_{t+i}是在时间步t+i处收到的奖励，$V(s_{t+n},\theta)$是指在时间步t+n处的值估计。

[0078] 一般，环境中的实体可以是自然的或人造的，并且该环境可以是真实环境或模拟的真实环境或虚拟环境。代理可以包括计算机控制的或人控制的机器，诸如机器人或自主的陆地、海上或空中运载工具。代理还可以包括人和/或动物。代理还可以包括静止或移动的软件代理，即配置为自主操作和/或与其它软件代理或人员一起操作以执行诸如计算机或通信网络的配置或维护、或制造工厂或数据中心/服务器场的配置或维护的任务的计算机程序。

[0079] 例如，系统的一些实施方式可以用于对运载工具交通或行人交通的预测或控制(例如，用于交通信号控制以减少拥堵)，或用于对执行任务或玩游戏的人员团队的预测或控制(例如，通过基于来自系统的输出例如表示数据向人员提供信号)。在一些其它实施方式中，系统可以用于对执行诸如仓库或物流自动化的任务的机器人的合作控制、包裹递送控制(例如，使用机器人)、无人机队控制等。

[0080] 前述类型的任何自主或半自主代理可以包括与RFM系统结合操作以控制代理的强化学习系统。如前所述的类型的多个自主或半自主代理可以包括RFM系统以及与RFM系统结合操作以促进复杂环境中的合作行为的控制系统，例如强化学习系统。在一些实施方式中，代理可以是不同的类型。例如，在仓库背景的自主运载工具或仓库中，来自多个不同制造商的控制机器人可以操作。在这种情况下，为这些机器人/运载工具中的每个配备强化学习系统和RFM系统的组合允许不同实体学习一起工作。利用其它类型的代理可以获得类似的益处。

[0081] 非代理实体可以包括上述代理的环境中的任何非代理对象。

[0082] 在图6所示类型的代理的其中关系前向模型被组合到强化学习系统中的另一些应用中，环境是真实环境并且代理是与真实环境相互的机电代理。例如，代理可以是与环境交互以完成特定任务(例如，将感兴趣的对象定位在环境中，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)的机器人或其它静止或移动的机器；或者代理可以是在环境中航行的自主或半自主的陆地或空中或海上运载工具。

[0083] 在这些实施方式中，观察可以包括例如图像、对象位置数据和在代理与环境交互时捕获观察的传感器数据(例如，来自图像、距离或位置传感器或者来自致动器的传感器数

据)中的一个或更多个。在机器人或其它机械代理或运载工具的情况下,观察可以类似地包括代理的一个或更多个部分的位置、线速度或角速度、力、转矩或加速度以及整体或相对姿势中的一个或更多个。观察可以在1维、2维或3维中定义,并且可以是绝对和/或相对的观察。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更多个:关节位置,关节速度,关节力,转矩或加速度,以及机器人的一部分(诸如臂和/或由机器人持拿的物件)的整体或相对姿势。观察还可以包括例如:感测到的电子信号,诸如电动机电流或温度信号;和/或例如来自相机或LIDAR传感器的图像或视频数据,例如来自代理的传感器的数据或来自在环境中与代理分开定位的传感器的数据。

[0084] 在这些实施方式中,动作可以是:控制机器人的控制输入,例如,针对机器人关节的转矩或更高级别的控制命令;或控制自主或半自主的陆地或空中或海上运载工具的控制输入,例如对于运载工具的控制表面或其它控制元件的转矩或更高级别的控制命令;或例如电动机控制数据。换句话说,动作可以包括例如机器人的一个或更多个关节或另一机械代理的部分的位置、速度或力/转矩/加速度数据。动作数据可以包括这些动作的数据和/或电子控制数据(诸如电动机控制数据)或更一般地,用于控制环境中的一个或更多个电子设备的数据,该电子设备的控制对观察到的环境状态有影响。例如,在自主或半自主的陆地或空中或海上运载工具的情况下,动作可以包括控制运载工具的航行(例如,转向和移动(例如制动和/或加速))的动作。

[0085] 例如,模拟环境可以是机器人或运载工具代理的模拟,并且可以在该模拟上训练强化学习系统。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,并且代理是在运动模拟中导航的模拟运载工具。在这些实施方式中,动作可以是控制模拟用户或模拟运载工具的控制输入。在使用真实环境中的系统之前,模拟环境对于训练强化学习系统可以是有用的。在另一示例中,模拟环境可以是视频游戏,并且代理可以是玩该视频游戏的模拟用户。一般,在模拟环境的情况下,观察可以包括一个或更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更多个先前描述的动作或动作类型的模拟版本。

[0086] 在电子代理的情况下,观察可以包括来自监视工厂或服务设施的部分的一个或更多个传感器(例如,电流、电压、功率、温度和其它传感器)的数据、和/或代表设备的电子和/或机械物件的功能的电子信号。在一些应用中,代理可以控制包括设备物件的真实环境中(例如在诸如数据中心、服务器场或电网干线电源或水分配系统的设施中)的动作,或者在制造工厂或服务设施中的运作。然后,观察可以与工厂或设施的操作有关。例如,作为前述内容的补充或替代,它们可以包括对设备的电力使用或水使用的观察、或者对发电或分配控制的观察、或者对资源使用或废物产生的观察。代理可以例如通过减少资源使用来控制环境中的动作以提高效率,和/或例如通过减少浪费来减少环境中的操作的环境影响。例如,代理可以控制设施中的电消耗或其它功率消耗或水的使用、和/或设施和/或设施中的物件的温度。动作可以包括控制工厂/设施的设备物件或将操作条件强加于工厂/设施的设备物件上的动作、和/或导致工厂/设施的操作中的设置(例如调整或打开/关闭工厂/设施的组件)改变的动作。

[0087] 在另一些应用中,环境是真实环境,并且代理管理跨计算资源(例如,在移动设备上和/或在数据中心中)的任务分配。在这些实施方式中,动作可以包括将任务分配给特定

的计算资源。

[0088] 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中、在有形地体现的计算机软件或固件中、在计算机硬件(包括本说明书中公开的结构及其结构等同物)中、或在它们中的一个或更多个的组合中实现。本说明书中描述的主题的实施例可以被实现为一个或更多个计算机程序,即在有形的非暂时性程序载体上编码以由数据处理装置运行或控制数据处理装置的操作的计算机程序指令的一个或更多个模块。备选地或另外地,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、光学信号或电磁信号)上,该人工生成的传播信号被生成以对信息进行编码用于传输到合适的接收器装置以由数据处理装置执行。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、或它们中的一个或更多个的组合。然而,计算机存储介质不是传播信号。

[0089] 术语“数据处理装置”涵盖用于处理数据的各种装置、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。该装置可以包括专用逻辑电路,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,该装置还可以包括为所讨论的计算机程序创建运行环境的代码,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更多个的组合的代码。

[0090] 计算机程序(也可称为或描述为程序、软件、软件应用、模块、软件模块、脚本或代码)可以以任何形式的编程语言(包括编译或解释的语言)或声明性或过程语言来编写,并且其可以以任何形式进行部署,包括部署为独立程序或部署为模块、组件、子例程或适用于计算环境的其它单元。计算机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的一部分文件(例如存储在标记语言文档中的一个或更多个脚本)中,在专用于所讨论的程序的单个文件中,或在多个协调文件(例如,存储一个或更多个模块、子程序或部分代码的文件)中。计算机程序可以被部署为在一台计算机上或在位于一个站点或分布于多个站点并通过通信网络互连的多台计算机上运行。

[0091] 如本说明书中所使用的,“引擎”或“软件引擎”指的是软件实现的输入/输出系统,其提供与输入不同的输出。引擎可以是功能的编码块,诸如库、平台、软件开发套件(“SDK”)或对象。可以在包括一个或更多个处理器和计算机可读介质的任何适当类型的计算设备(例如,服务器、移动电话、平板计算机、笔记本计算机、音乐播放器、电子书阅读器、膝上型或台式计算机、PDA、智能电话或其它固定或便携式设备)上实现每个引擎。另外,可以在同一计算设备上或在不同的计算设备上实现两个或更多个引擎。

[0092] 本说明书中描述的过程和逻辑流可以由一个或更多个可编程计算机执行,所述一个或更多个可编程计算机运行一个或更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。过程和逻辑流也可以由专用逻辑电路(例如,FPGA(现场可编程门阵列)或ASIC(专用集成电路))执行,并且装置也可以被实现为该专用逻辑电路。例如,过程和逻辑流可以由图形处理单元(GPU)执行,并且装置也可以被实现为图形处理单元(GPU)。

[0093] 例如,适合于运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,或者基于任何其它种类的中央处理单元。一般,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运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更多个存储器设备。一般,计算机还将包括用于存储数据

的一个或更多个大容量存储设备(例如,磁盘、磁光盘或光盘),或被操作地耦合以从所述一个或更多个大容量存储设备接收数据或向其传输数据,或两者。然而,计算机不必具有这样的设备。此外,计算机可以被嵌入另一个设备中,例如,移动电话、个人数字助理(PDA)、移动音频或视频播放器、游戏主机、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)等等。

[0094] 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盘;磁光盘;以及CD ROM和DVD-ROM盘。处理器和存储器可以由专用逻辑电路补充或并入专用逻辑电路中。

[0095] 为了支持与用户的交互,可以在具有用于向用户显示信息的显示器设备(例如,CRT(阴极射线管)或LCD(液晶显示器)监视器)以及用户可通过其向计算机提供输入的键盘和定点设备(例如,鼠标或轨迹球)的计算机上实现本说明书中描述的主题的实施例。其它种类的设备也可以用于支持与用户的交互;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(包括声音输入、语音输入或触觉输入)接收来自用户的输入。此外,计算机可以通过向用户使用的设备发送文档以及从用户使用的设备接收文档来与用户交互;例如,通过响应于从网络浏览器接收到的请求向用户的客户端设备上的网络浏览器发送网页。

[0096] 本说明书中描述的主题的实施例可以在包括后端组件(例如,作为数据服务器)或包括中间件组件(例如,应用服务器)或包括前端组件(例如,具有图形用户界面或网络浏览器的客户端计算机,用户可通过该图形用户界面或网络浏览器与本说明书中描述的主题的实施方式交互)或一个或更多个这样的后端组件、中间件组件或前端组件的任何组合的计算系统中实现。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。

[0097] 计算系统可以包括客户端和服务器。客户端和服务器一般彼此远离,并且通常通过通信网络交互。客户端和服务器之间的关系是通过在相应计算机上运行且彼此具有客户端-服务器关系的计算机程序产生的。

[0098] 虽然本说明书包含许多特定的实施细节,但是这些细节不应被解释为对任何发明或所要求保护的内容的范围的限制,而是应被解释为对特定发明的特定实施例所特有的特征的描述。在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可以在单个实施例中组合地实现。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在多个实施例中或以任何合适的子组合来实现。而且,尽管特征可以在上文中被描述为以某些组合起作用并且甚至最初如此要求保护,但是在一些情况下可以从组合中切除所要求保护的组合中的一个或更多个特征,并且所要求保护的组合可以针对子组合或了组合的变型。

[0099] 类似地,虽然在附图中按特定顺序描绘了操作,但是这不应被理解为要求按所示的特定顺序或按连续的顺序执行这些操作或要求执行所有示出的操作来获得期望的结果。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,并且应理解,所描述的程序组件和系统一般可以一起集成在单个软件产品中或打包成多个软件产品。

[0100] 已经描述了主题的特定实施例。其它实施例在所附权利要求的范围内。例如,权利

要求中记载的动作可以按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连续的顺序来实现期望的结果。在某些实施方式中,多任务和并行处理可以是有利的。

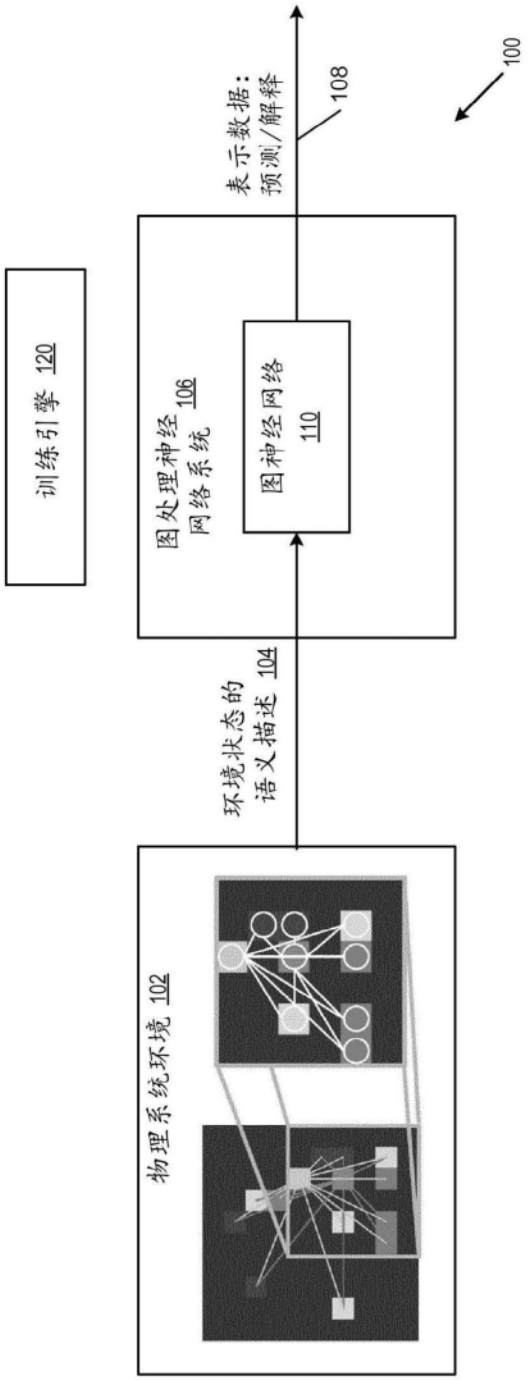

100
108
表示数据：
预测/解释
训练引擎 120
图处理神经
网络系统 106
图神经网络
110
环境状态的
语义描述 104
物理系统环境 102

图1

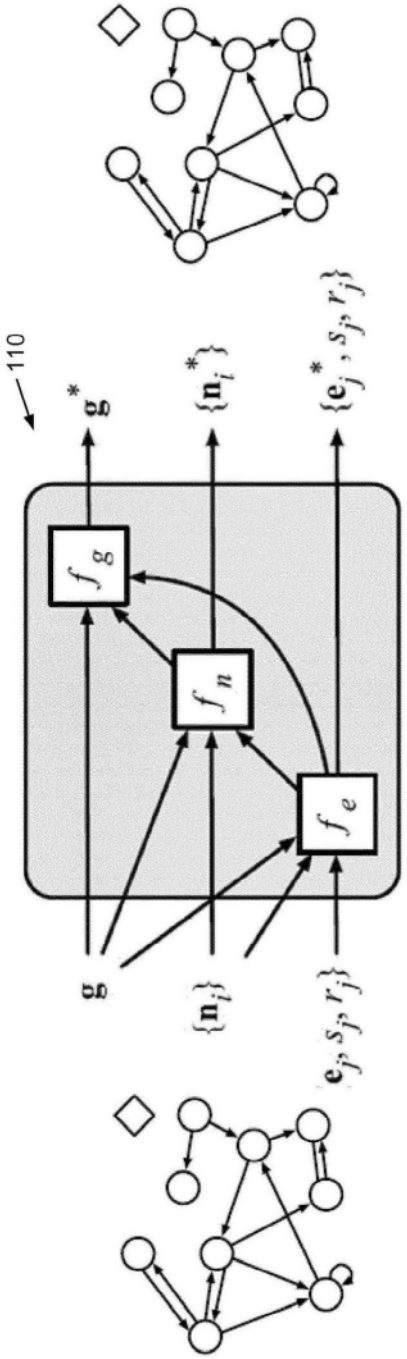

110
$\mathbf{g}^*$
$\{\mathbf{n}_i^*\}$
$\{\mathbf{e}_j^*, s_j, r_j\}$
f_g
f_n
f_e
$\mathbf{g}$
$\{\mathbf{n}_i\}$
$\mathbf{e}_j, s_j, r_j\}$

图2a

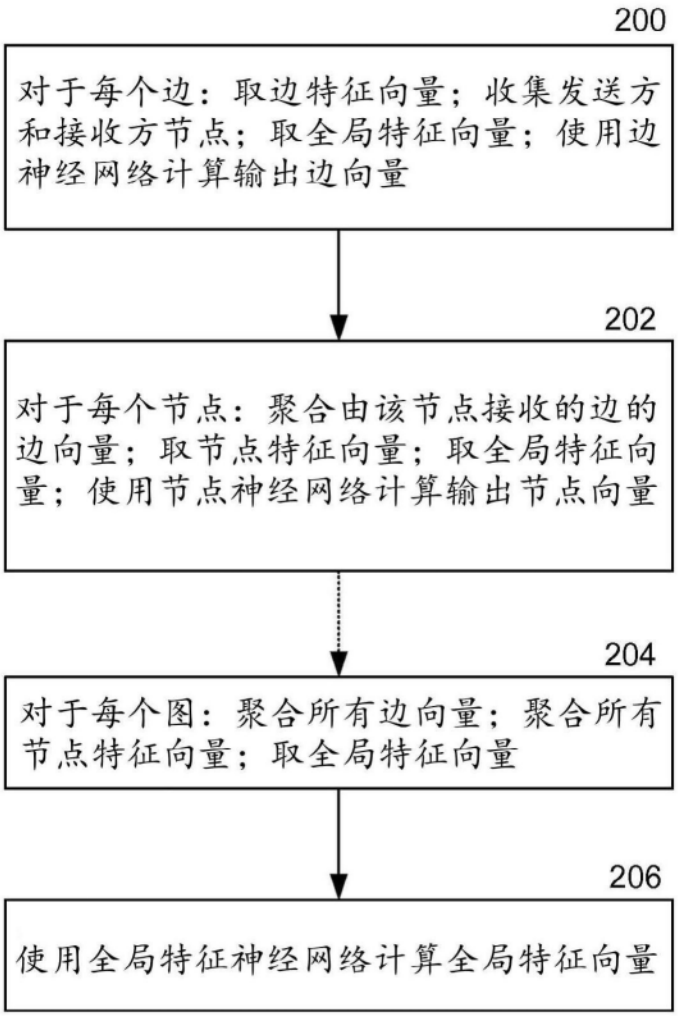

200
对于每个边：取边特征向量；收集发送方和接收方节点；取全局特征向量；使用边神经网络计算输出边向量
202
对于每个节点：聚合由该节点接收的边的边向量；取节点特征向量；取全局特征向量；使用节点神经网络计算输出节点向量
204
对于每个图：聚合所有边向量；聚合所有节点特征向量；取全局特征向量
206
使用全局特征神经网络计算全局特征向量

图2b

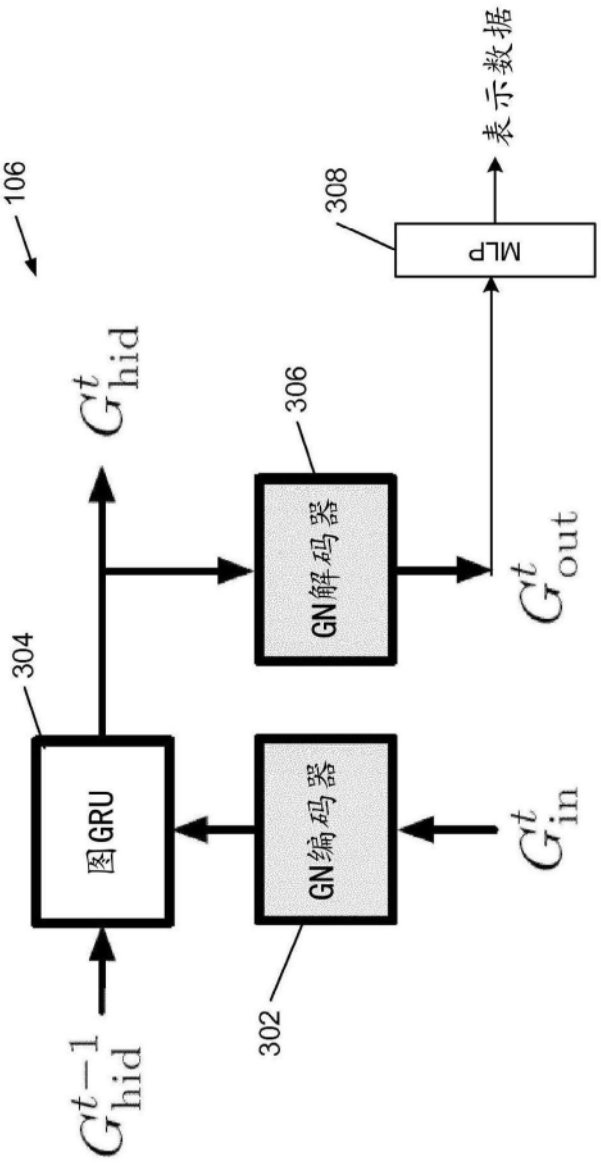

106
308
MLP
表示数据
306
GN解码器
G^t_{hid}
G^t_{out}
304
图GRU
GN编码器
G^t_{in}
302
G^{t-1}_{hid}

图3

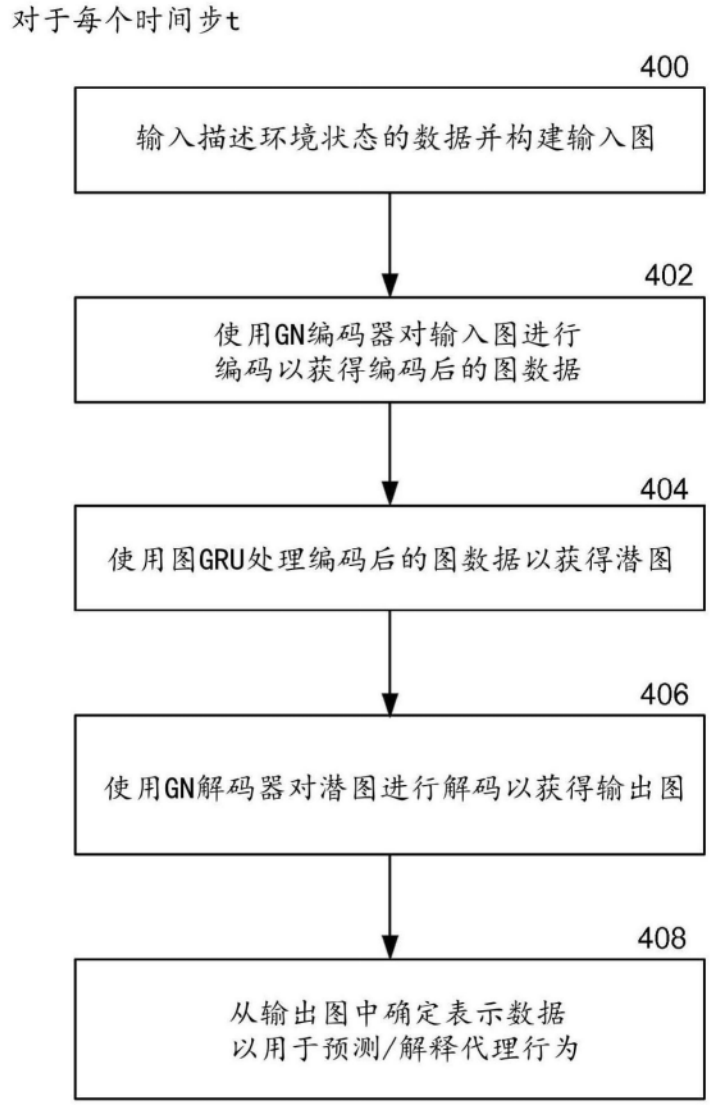

对于每个时间步t
400
输入描述环境状态的数据并构建输入图
402
使用GN编码器对输入图进行
编码以获得编码后的图数据
404
使用图GRU处理编码后的图数据以获得潜图
406
使用GN解码器对潜图进行解码以获得输出图
408
从输出图中确定表示数据
以用于预测/解释代理行为

图4

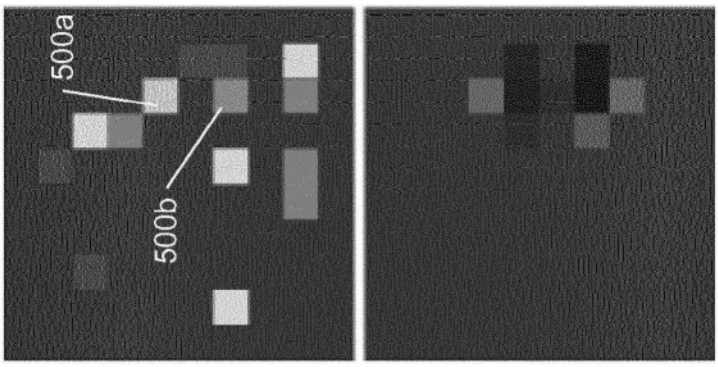

500a
500b

图5a

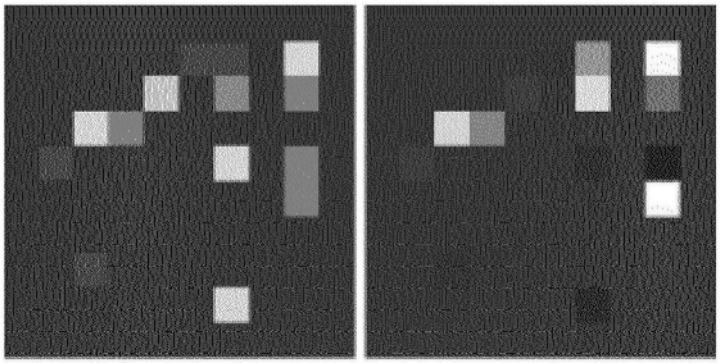


图5b

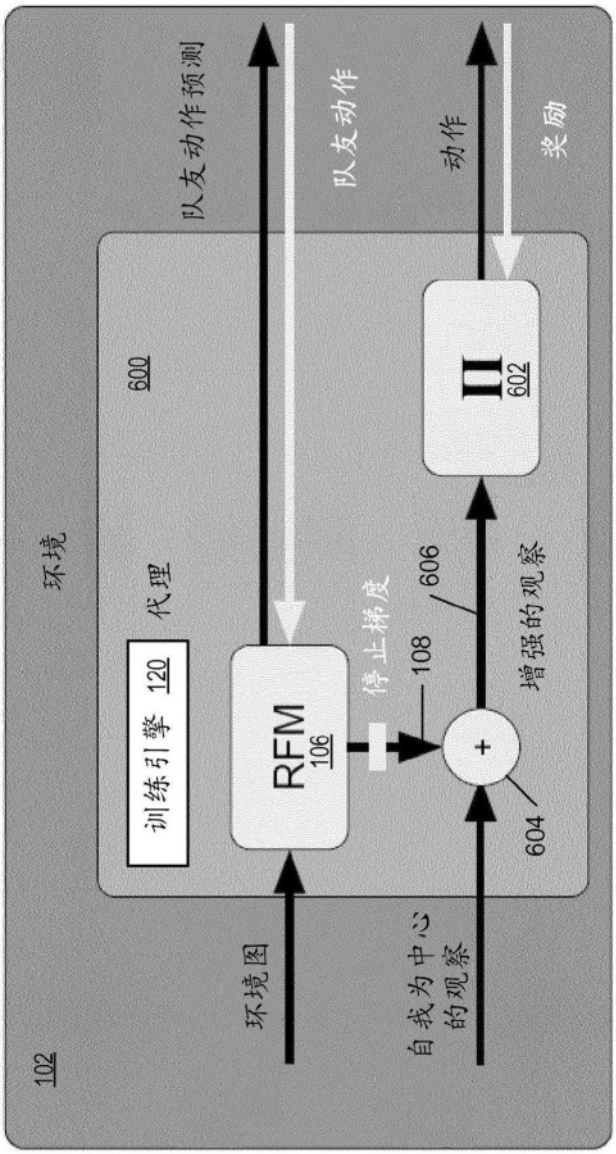

环境
102
环境图
自我为中心的观察
代理
600
训练引擎 120
RFM
106
停止梯度
108
606
604
增强的观察
II
602
队友动作预测
队友动作
动作
奖励

图6