



(12) 发明专利

(10) 授权公告号 CN 113211441 B

(45) 授权公告日 2022.09.09

(21) 申请号 202110522641.3

G06N 3/08 (2006.01)

(22) 申请日 2021.05.13

(56) 对比文件

(65) 同一申请的已公布的文献号
申请公布号 CN 113211441 A

CN 109190760 A, 2019.01.11

CN 110651279 A, 2020.01.03

CN 108229678 A, 2018.06.29

(43) 申请公布日 2021.08.06

CN 111260027 A, 2020.06.09

(66) 本国优先权数据

CN 111931910 A, 2020.11.13

202011378716.7 2020.11.30 CN

CN 109291052 A, 2019.02.01

(73) 专利权人 湖南太观科技有限公司

CN 107020636 A, 2017.08.08

地址 410000 湖南省长沙市开福区芙蓉北路街道金马路377号福天兴业综合楼407房

CN 111203878 A, 2020.05.29

CN 109240280 A, 2019.01.18

CN 111506405 A, 2020.08.07

(72) 发明人 宋子豪

CN 111768028 A, 2020.10.13

JP 2019516568 A, 2019.06.20

(74) 专利代理机构 北京林达刘知识产权代理事务所(普通合伙) 11277

CN 110119844 A, 2019.08.13

专利代理师 刘新宇

李帅龙等. 模仿学习方法综述及其在机器人领域的应用.《计算机工程与应用》.2019, 第17-30页.

(51) Int. Cl.

B25J 9/16 (2006.01)

审查员 何成斌

G06N 3/04 (2006.01)

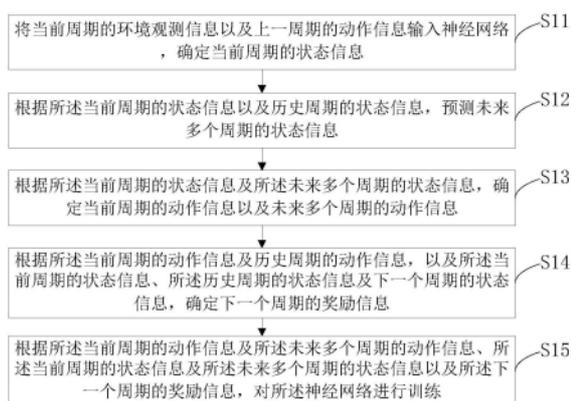
权利要求书3页 说明书16页 附图4页

(54) 发明名称

神经网络训练和机器人控制方法及装置

(57) 摘要

本公开涉及一种神经网络训练和机器人控制方法及装置,所述方法包括:将当前周期的环境观测信息以及上一周期的动作信息输入神经网络,确定当前周期的状态信息;根据当前及历史周期的状态信息,预测未来周期的状态信息;根据当前及未来周期的状态信息,确定当前及未来周期的动作信息;根据当前及历史周期的动作信息、状态信息,确定未来周期的奖励信息;根据当前及未来周期的动作信息、状态信息以及奖励信息,对所述神经网络进行训练。根据本公开的实施例的神经网络训练方法,在预测未来周期的状态信息时,可考虑历史周期的影响,在训练过程中使神经网络获得对训练过程的全局认知,提高训练效率,且易于适应新环境。



CN 113211441 B

1. 一种神经网络训练方法,其特征在于,所述方法包括:

将当前周期的环境观测信息以及上一个周期的动作信息输入神经网络,确定当前周期的状态信息,所述环境观测信息用于描述环境,所述环境包括机器人的运行环境,所述动作信息用于作用于当前周期的环境中,对环境造成改变,所述动作信息包括用于控制机器人动作的控制信息,所述状态信息用于描述机器人的控制状态;

根据所述当前周期的状态信息以及历史周期的状态信息,预测未来多个周期的状态信息;

根据所述当前周期的状态信息及所述未来多个周期的状态信息,确定当前周期的动作信息以及未来多个周期的动作信息;

根据所述当前周期的动作信息及历史周期的动作信息,以及所述当前周期的状态信息、所述历史周期的状态信息及下一个周期的状态信息,确定未来多个周期的奖励信息;

根据所述当前周期的动作信息及所述未来多个周期的动作信息、所述当前周期的状态信息及所述未来多个周期的状态信息,以及未来多个周期的奖励信息,对所述神经网络进行训练。

2. 根据权利要求1所述的方法,其特征在于,将当前周期的环境观测信息以及上一个周期的动作信息输入神经网络,确定当前周期的状态信息,包括:

根据所述当前周期的环境观测信息以及所述上一个周期的动作信息,确定隐状态;

根据所述隐状态确定历史周期中与所述隐状态特征距离最近的第一隐状态值;

根据所述第一隐状态值和所述隐状态,确定所述当前周期的状态信息。

3. 根据权利要求2所述的方法,其特征在于,所述方法还包括:

根据所述当前周期的状态信息,对历史周期中的第一隐状态值进行更新,获得当前周期的第一隐状态值。

4. 根据权利要求1所述的方法,其特征在于,根据所述当前周期的状态信息以及历史周期的状态信息,预测未来多个周期的状态信息,包括:

根据所述当前周期的状态信息和历史周期的状态信息,确定当前周期的状态信息转移至下一个周期的状态信息的马尔科夫转移概率;

根据所述当前周期的状态信息转移至下一个周期的状态信息的马尔科夫转移概率,以及所述当前周期的状态信息,确定下一个周期的状态信息;

根据所述下一个周期的状态信息,确定所述未来多个周期的状态信息。

5. 根据权利要求1所述的方法,其特征在于,根据所述当前周期的动作信息及历史周期的动作信息,以及所述当前周期的状态信息、所述历史周期的状态信息及下一个周期的状态信息,确定未来多个周期的奖励信息,包括:

根据所述当前周期的动作信息及历史周期的动作信息,以及所述当前周期的状态信息、所述历史周期的状态信息及下一个周期的状态信息,确定相对熵;

根据所述相对熵,确定所述未来多个周期的奖励信息;

根据所述相对熵,确定所述未来多个周期的奖励信息,包括:通过对所述神经网络进行训练,使得相对熵最小化,使得奖励信息最大化,以获得所述奖励信息。

6. 根据权利要求5所述的方法,其特征在于,根据所述当前周期的动作信息及历史周期的动作信息,以及所述当前周期的状态信息、所述历史周期的状态信息及下一个周期的状

态信息,确定相对熵,包括:

确定所述当前周期的动作信息及历史周期的动作信息组成的动作序列,以及所述当前周期的状态信息、所述历史周期的状态信息及下一个周期的状态信息组成的状态序列之间的第一互信息;

根据所述第一互信息确定所述相对熵。

7. 根据权利要求5所述的方法,其特征在于,根据所述当前周期的动作信息及历史周期的动作信息,以及所述当前周期的状态信息、所述历史周期的状态信息及下一个周期的状态信息,确定相对熵,包括:

根据所述当前周期的动作信息及历史周期的动作信息,以及所述当前周期的状态信息及所述历史周期的状态信息,确定技能序列;

确定所述技能序列,和所述当前周期的动作信息及历史周期的动作信息组成的动作序列之间的第二互信息;

根据所述第二互信息确定所述相对熵。

8. 根据权利要求5所述的方法,其特征在于,根据所述当前周期的动作信息及历史周期的动作信息,以及所述当前周期的状态信息、所述历史周期的状态信息及下一个周期的状态信息,确定相对熵,包括:

根据所述当前周期的状态信息和所述历史周期的状态信息,确定全局特征;

根据所述全局特征、所述当前周期的动作信息及历史周期的动作信息,以及所述当前周期的状态信息、所述历史周期的状态信息及下一个周期的状态信息,确定所述相对熵。

9. 根据权利要求1所述的方法,其特征在于,根据所述当前周期的动作信息及所述未来多个周期的动作信息、所述当前周期的状态信息及所述未来多个周期的状态信息,以及未来多个周期的奖励信息,对所述神经网络进行训练,包括:

根据所述未来多个周期的奖励信息、所述当前周期的动作信息及所述未来多个周期的动作信息,以及所述当前周期的状态信息及所述未来多个周期的状态信息,确定奖励价值信息;

根据所述奖励价值信息,训练所述神经网络。

10. 一种机器人控制方法,其特征在于,包括:

将当前周期的环境观测信息、上一个周期的动作信息、上一个周期的状态信息和当前周期的奖励信息输入根据权利要求1-9中任意一项所述的神经网络训练方法训练后的神经网络,获得当前周期的动作信息。

11. 一种神经网络训练装置,其特征在于,所述装置包括:

环境确定模块,用于将当前周期的环境观测信息以及上一个周期的动作信息输入神经网络,确定当前周期的状态信息,所述环境观测信息用于描述环境,所述环境包括机器人的运行环境,所述动作信息用于作用于当前周期的环境中,对环境造成改变,所述动作信息包括用于控制机器人动作的控制信息,所述状态信息用于描述机器人的控制状态;

环境预测模块,用于根据所述当前周期的状态信息以及历史周期的状态信息,预测未来多个周期的状态信息;

动作确定模块,用于根据所述当前周期的状态信息及所述未来多个周期的状态信息,确定当前周期的动作信息以及未来多个周期的动作信息;

奖励确定模块,用于根据所述当前周期的动作信息及历史周期的动作信息,以及所述当前周期的状态信息、所述历史周期的状态信息及下一个周期的状态信息,确定未来多个周期的奖励信息;

训练模块,用于根据所述当前周期的动作信息及所述未来多个周期的动作信息、所述当前周期的状态信息及所述未来多个周期的状态信息,以及未来多个周期的奖励信息,对所述神经网络进行训练。

12. 一种机器人控制装置,其特征在于,包括:

控制模块,用于将当前周期的环境观测信息、上一个周期的动作信息、上一个周期的状态信息和当前周期的奖励信息输入根据权利要求11所述的神经网络训练装置训练后的神经网络,获得当前周期的动作信息。

13. 一种电子设备,其特征在于,包括:

处理器;

用于存储处理器可执行指令的存储器;

其中,所述处理器被配置为调用所述存储器存储的指令,以执行权利要求1至10中任意一项所述的方法。

14. 一种计算机可读存储介质,其上存储有计算机程序指令,其特征在于,所述计算机程序指令被处理器执行时实现权利要求1至10中任意一项所述的方法。

神经网络训练和机器人控制方法及装置

技术领域

[0001] 本公开涉及计算机技术领域,尤其涉及一种神经网络训练和机器人控制方法及装置。

背景技术

[0002] 在机器人控制领域,机器人所处实际环境的状态是复杂的,机器人只能获得部分感知信号,例如,其零件位置、速度等。进而可输出控制信号,例如,关节的扭矩等。机器人通过感知信号观测到的环境的维度低于实际环境的维度。

[0003] 在相关技术中,机器人可观测外部真实环境的状态,获得观测信号,并通过神经网络获得控制状态,该控制状态是对外部高维度环境状态的低维度表示,可包括外部状态的特征信息和机器人的判断信息等。

[0004] 机器人可基于控制状态形成控制策略,例如,用于控制机器人行动的控制信息。其行动作用于外部环境,引起环境发生改变,并带来下一个状态的奖励信息,例如,将环境改变至目标状态的过程中,如果环境与目标状态之间的差距缩小,则产生正的奖励信息,反之,则产生负的奖励信息。综上所述,下一步的环境状态和奖励信息取决于当前的环境状态和动作信息,历史状态可反映在当前的环境状态中,但不直接影响未来状态。

[0005] 然而,真实环境中的复杂动态存在时间尺度上的长程依赖、空间尺度上的相互作用,而下一步的环境状态和奖励信息取决于当前的环境状态和动作信息,而忽略了现实中的动态关系,使得控制策略容易陷入某种单一的模式,不利于奖励信息的长期回报的最大化。此外,通过上述方式需要大量反复迭代,机器学习的效率较低,且奖励信息仅可表示某一个状态的环境反馈,较为敏感,训练过程中易使得机器人缺乏整体认知,难以适应新环境。

发明内容

[0006] 本公开提出了一种神经网络训练和机器人控制方法及装置。

[0007] 根据本公开的一方面,提供了一种神经网络训练方法,包括:将当前周期的环境观测信息以及上一个周期的动作信息输入神经网络,确定当前周期的状态信息,所述环境观测信息用于描述环境,所述环境包括机器人的运行环境,所述动作信息用于作用于当前周期的环境中,对环境造成改变,所述动作信息包括用于控制机器人动作的控制信息,所述状态信息用于描述机器人的控制状态;根据所述当前周期的状态信息以及历史周期的状态信息,预测未来多个周期的状态信息;根据所述当前周期的状态信息及所述未来多个周期的状态信息,确定当前周期的动作信息以及未来多个周期的动作信息;根据所述当前周期的动作信息及历史周期的动作信息,以及所述当前周期的状态信息、所述历史周期的状态信息及下一个周期的状态信息,确定未来多个周期的奖励信息;根据所述当前周期的动作信息及所述未来多个周期的动作信息、所述当前周期的状态信息及所述未来多个周期的状态信息,以及未来多个周期的奖励信息,对所述神经网络进行训练。

[0008] 在一种可能的实现方式中,根据当前周期的环境观测信息以及上一个周期的动作信息输入神经网络,确定当前周期的状态信息,包括:根据所述当前周期的环境观测信息以及所述上一个周期的动作信息,确定隐状态;根据所述隐状态确定历史周期中与所述隐状态特征距离最近的第一隐状态值;根据所述第一隐状态值和所述隐状态,确定所述当前周期的状态信息。

[0009] 在一种可能的实现方式中,所述方法还包括:根据所述当前周期的状态信息,对历史周期中的第一隐状态值进行更新,获得当前周期的第一隐状态值。

[0010] 在一种可能的实现方式中,根据所述当前周期的状态信息以及历史周期的状态信息,预测未来多个周期的状态信息,包括:根据所述当前周期的状态信息和历史周期的状态信息,确定当前周期的状态信息转移至下一个周期的状态信息的马尔科夫转移概率;根据所述当前周期的状态信息转移至下一个周期的状态信息的马尔科夫转移概率,以及所述当前周期的状态信息,确定下一个周期的状态信息;根据所述下一个周期的状态信息,确定所述未来多个周期的状态信息。

[0011] 在一种可能的实现方式中,根据所述当前周期的动作信息及历史周期的动作信息,以及所述当前周期的状态信息、所述历史周期的状态信息及下一个周期的状态信息,确定未来多个周期的奖励信息,包括:根据所述当前周期的动作信息及历史周期的动作信息,以及所述当前周期的状态信息、所述历史周期的状态信息及下一个周期的状态信息,确定相对熵;根据所述相对熵,确定所述确定未来多个周期的奖励信息。

[0012] 在一种可能的实现方式中,根据所述当前周期的动作信息及历史周期的动作信息,以及所述当前周期的状态信息、所述历史周期的状态信息及下一个周期的状态信息,确定相对熵,包括:确定所述当前周期的动作信息及历史周期的动作信息组成的动作序列,以及所述当前周期的状态信息、所述历史周期的状态信息及下一个周期的状态信息组成的状态序列之间的第一互信息;根据所述第一互信息确定所述相对熵。

[0013] 在一种可能的实现方式中,根据所述当前周期的动作信息及历史周期的动作信息,以及所述当前周期的状态信息、所述历史周期的状态信息及下一个周期的状态信息,确定相对熵,包括:根据所述当前周期的动作信息及历史周期的动作信息,以及所述当前周期的状态信息及所述历史周期的状态信息,确定技能序列;确定所述技能序列,和所述当前周期的动作信息及历史周期的动作信息组成的动作序列之间的第二互信息;根据所述第二互信息确定所述相对熵。

[0014] 在一种可能的实现方式中,根据所述当前周期的动作信息及历史周期的动作信息,以及所述当前周期的状态信息、所述历史周期的状态信息及下一个周期的状态信息,确定相对熵,包括:根据所述当前周期的状态信息和所述历史周期的状态信息,确定全局特征;根据所述全局特征、所述当前周期的动作信息及历史周期的动作信息,以及所述当前周期的状态信息、所述历史周期的状态信息及下一个周期的状态信息,确定所述相对熵。

[0015] 在一种可能的实现方式中,根据所述当前周期的动作信息及所述未来多个周期的动作信息、所述当前周期的状态信息及所述未来多个周期的状态信息,以及未来多个周期的奖励信息,对所述神经网络进行训练,包括:根据所述未来多个周期的奖励信息、所述当前周期的动作信息及所述未来多个周期的动作信息,以及所述当前周期的状态信息及所述未来多个周期的状态信息,确定奖励价值信息;根据所述奖励价值信息,训练所述神经网络。

络。

[0016] 根据本公开的一方面,提供了一种机器人控制方法,其特征在于,包括:将当前周期的环境观测信息、上一个周期的动作信息、上一个周期的状态信息和当前周期的奖励信息输入所述神经网络训练方法训练后的神经网络,获得当前周期的动作信息。

[0017] 根据本公开的一方面,提供了一种神经网络训练装置,包括:环境确定模块,用于将当前周期的环境观测信息以及上一个周期的动作信息输入神经网络,确定当前周期的状态信息,所述环境观测信息用于描述环境,所述环境包括机器人的运行环境,所述动作信息用于作用于当前周期的环境中,对环境造成改变,所述动作信息包括用于控制机器人动作的控制信息,所述状态信息用于描述机器人的控制状态;环境预测模块,用于根据所述当前周期的状态信息以及历史周期的状态信息,预测未来多个周期的状态信息;动作确定模块,用于根据所述当前周期的状态信息及所述未来多个周期的状态信息,确定当前周期的动作信息以及未来多个周期的动作信息;奖励确定模块,用于根据所述当前周期的动作信息及历史周期的动作信息,以及所述当前周期的状态信息、所述历史周期的状态信息及下一个周期的状态信息,确定未来多个周期的奖励信息;训练模块,用于根据所述当前周期的动作信息及所述未来多个周期的动作信息、所述当前周期的状态信息及所述未来多个周期的状态信息,以及未来多个周期的奖励信息,对所述神经网络进行训练。

[0018] 在一种可能的实现方式中,所述环境确定模块进一步用于根据所述当前周期的环境观测信息以及所述上一个周期的动作信息,确定隐状态;根据所述隐状态确定历史周期中与所述隐状态特征距离最近的第一隐状态值;根据所述第一隐状态值和所述隐状态,确定所述当前周期的状态信息。

[0019] 在一种可能的实现方式中,所述装置还包括:更新模块,用于根据所述当前周期的状态信息,对历史周期中的第一隐状态值进行更新,获得当前周期的第一隐状态值。

[0020] 在一种可能的实现方式中,所述环境预测模块进一步用于:根据所述当前周期的状态信息和历史周期的状态信息,确定当前周期的状态信息转移至下一个周期的状态信息的马尔科夫转移概率;根据所述当前周期的状态信息转移至下一个周期的状态信息的马尔科夫转移概率,以及所述当前周期的状态信息,确定下一个周期的状态信息;根据所述下一个周期的状态信息,确定所述未来多个周期的状态信息。

[0021] 在一种可能的实现方式中,所述奖励确定模块进一步用于:根据所述当前周期的动作信息及历史周期的动作信息,以及所述当前周期的状态信息、所述历史周期的状态信息及下一个周期的状态信息,确定相对熵;根据所述相对熵,确定所述未来多个周期的奖励信息。

[0022] 在一种可能的实现方式中,所述奖励确定模块进一步用于:确定所述当前周期的动作信息及历史周期的动作信息组成的动作序列,以及所述当前周期的状态信息、所述历史周期的状态信息及下一个周期的状态信息组成的状态序列之间的第一互信息;根据所述第一互信息确定所述相对熵。

[0023] 在一种可能的实现方式中,所述奖励确定模块进一步用于:根据所述当前周期的动作信息及历史周期的动作信息,以及所述当前周期的状态信息及所述历史周期的状态信息,确定技能序列;确定所述技能序列,和所述当前周期的动作信息及历史周期的动作信息组成的动作序列之间的第二互信息;根据所述第二互信息确定所述相对熵。

[0024] 在一种可能的实现方式中,所述奖励确定模块进一步用于:根据所述当前周期的状态信息和所述历史周期的状态信息,确定全局特征;根据所述全局特征、所述当前周期的动作信息及历史周期的动作信息,以及所述当前周期的状态信息、所述历史周期的状态信息及下一个周期的状态信息,确定所述相对熵。

[0025] 在一种可能的实现方式中,所述训练模块进一步用于:根据所述未来多个周期的奖励信息、所述当前周期的动作信息及所述未来多个周期的动作信息,以及所述当前周期的状态信息及所述未来多个周期的状态信息,确定奖励价值信息;根据所述奖励价值信息,训练所述神经网络。

[0026] 根据本公开的一方面,提供了一种机器人控制装置,包括:控制模块,用于将当前周期的环境观测信息、上一个周期的动作信息、上一个周期的状态信息和当前周期的奖励信息输入根据所述神经网络训练装置训练后的神经网络,获得当前周期的动作信息。

[0027] 根据本公开的一方面,提供了一种电子设备,包括:处理器;用于存储处理器可执行指令的存储器;其中,所述处理器被配置为调用所述存储器存储的指令,以执行上述方法。

[0028] 根据本公开的一方面,提供了一种计算机可读存储介质,其上存储有计算机程序指令,所述计算机程序指令被处理器执行时实现上述方法。

[0029] 应当理解的是,以上的一般描述和后文的细节描述仅是示例性和解释性的,而非限制本公开。根据下面参考附图对示例性实施例的详细说明,本公开的其它特征及方面将变得清楚。

附图说明

[0030] 此处的附图被并入说明书中并构成本说明书的一部分,这些附图示出了符合本公开的实施例,并与说明书一起用于说明本公开的技术方案。

[0031] 图1示出根据本公开实施例的神经网络训练方法的流程图;

[0032] 图2示出根据本公开实施例的神经网络训练方法的应用示意图;

[0033] 图3示出根据本公开实施例的神经网络训练装置的框图;

[0034] 图4示出根据本公开实施例的一种电子设备的框图;

[0035] 图5示出根据本公开实施例的一种电子设备的框图。

具体实施方式

[0036] 以下将参考附图详细说明本公开的各种示例性实施例、特征和方面。附图中相同的附图标记表示功能相同或相似的元件。尽管在附图中示出了实施例的各种方面,但是除非特别指出,不必按比例绘制附图。

[0037] 在这里专用的词“示例性”意为“用作例子、实施例或说明性”。这里作为“示例性”所说明的任何实施例不必解释为优于或好于其它实施例。

[0038] 本文中术语“和/或”,仅仅是一种描述关联对象的关联关系,表示可以存在三种关系,例如,A和/或B,可以表示:单独存在A,同时存在A和B,单独存在B这三种情况。另外,本文中术语“至少一种”表示多种中的任意一种或多种中的至少两种的任意组合,例如,包括A、B、C中的至少一种,可以表示包括从A、B和C构成的集合中选择的任意一个或多个元素。

[0039] 另外,为了更好地说明本公开,在下文的具体实施方式中给出了众多的具体细节。本领域技术人员应当理解,没有某些具体细节,本公开同样可以实施。在一些实例中,对于本领域技术人员熟知的方法、手段、元件和电路未作详细描述,以便于凸显本公开的主旨。

[0040] 图1示出根据本公开实施例的神经网络训练方法的流程图,如图1所示,所述神经网络训练方法包括:

[0041] 在步骤S11中,将当前周期的环境观测信息以及上一个周期的动作信息输入神经网络,确定当前周期的状态信息,所述环境观测信息用于描述环境,所述环境包括机器人的运行环境,所述动作信息用于作用于当前周期的环境中,对环境造成改变,所述动作信息包括用于控制机器人动作的控制信息,所述状态信息用于描述机器人的控制状态;

[0042] 在步骤S12中,根据所述当前周期的状态信息以及历史周期的状态信息,预测未来多个周期的状态信息;

[0043] 在步骤S13中,根据所述当前周期的状态信息及所述未来多个周期的状态信息,确定当前周期的动作信息以及未来多个周期的动作信息;

[0044] 在步骤S14中,根据所述当前周期的动作信息及历史周期的动作信息,以及所述当前周期的状态信息、所述历史周期的状态信息及下一个周期的状态信息,确定下一个周期的奖励信息;

[0045] 在步骤S15中,根据所述当前周期的动作信息及所述未来多个周期的动作信息、所述当前周期的状态信息及所述未来多个周期的状态信息以及所述下一个周期的奖励信息,对所述神经网络进行训练。

[0046] 根据本公开的实施例的神经网络训练方法,可通过当前周期的状态信息以及历史周期的状态信息,预测未来多个周期的状态信息,因而在预测未来周期的状态信息时,考虑历史周期的影响,即,考虑了时间尺度上的长程依赖、空间尺度上的相互作用,使得动作信息不易陷入单一的模式。并且,可基于未来多个周期的动作信息以及未来多个周期的状态信息来训练神经网络,在训练过程中使神经网络获得对训练过程的全局认知,提高训练效率,且易于适应新环境。

[0047] 在一种可能的实现方式中,所述神经网络训练方法可以由终端设备或服务器等电子设备执行,终端设备可以为用户设备(User Equipment, UE)、移动设备、用户终端、终端、蜂窝电话、无绳电话、个人数字助理(Personal Digital Assistant, PDA)、手持设备、计算设备、车载设备、可穿戴设备等,所述方法可以通过处理器调用存储器中存储的计算机可读指令的方式来实现。或者,可通过服务器执行所述方法。

[0048] 在一种可能的实现方式中,在机器人的控制过程中,机器人可通过对环境的观测,产生对应的控制信息,以控制机器人进行动作,并对环境产生作用,以改变环境。进一步地,可依据改变后的环境获得奖励信息,例如,如果环境与目标状态之间的差距缩小,则产生正的奖励信息,反之,则产生负的奖励信息,通过奖励信息可进一步确定下一个周期机器人的动作信息。在基于对环境的观测以及奖励信息产生动作信息的过程中,可通过增强学习神经网络来进行处理,本公开对神经网络的类型不做限制。

[0049] 在一种可能的实现方式中,为了将真实环境中的复杂动态在时间尺度上的长程依赖和空间尺度上的相互作用等因素(例如,历史周期中的多种因素)加入机器人的控制策略中,可基于当前周期的环境观测信息 o_t (即,当前周期对外部高维度环境进行观测获得的低

纬度的描述信息,例如,特征信息),以及上一个周期的动作信息 a_{t-1} (即,上一个周期的用于控制机器人动作的控制信息,例如,控制指令等)输入神经网络,神经网络可确定当前周期的状态信息 s_t (例如,描述机器人控制状态的信息), t 为大于1的正整数。

[0050] 在一种可能的实现方式中,步骤S11可包括:根据所述当前周期的环境观测信息以及所述上一个周期的动作信息,确定隐状态;根据所述隐状态确定历史周期中与所述隐状态特征距离最近的第一隐状态值;根据所述第一隐状态值和所述隐状态,确定所述当前周期的状态信息。

[0051] 在一种可能的实现方式中,所述神经网络可包括时间序列与空间信息结合的时空记忆神经网络,可通过该时空记忆神经网络来确定所述隐状态。所述时空记忆神经网络可以是包括关键值存储和边索引两个数据结构的图结构 $\Gamma = \{ (K, V), E \}$,其中, $V = \{v_i\}$ 为隐状态值, $K = \{k_i\}$ 是图结构的全局索引, V 和 K 可表示隐空间坐标, $E = \{(v_i, v_j)\}$ 是图结构的边索引,代表由隐状态值 v_i 转换为隐状态值 v_j 状态间的空间依存, i, j 均为正整数。

[0052] 在一种可能的实现方式中,可将当前周期的环境观测信息 o_t 和上一个周期的动作信息 a_{t-1} 输入时空记忆神经网络进行编码,获得隐状态 z_t 。

[0053] 在一种可能的实现方式中,可对时空记忆神经网络的图结构进行初始化,在示例中,可通过以下公式(1)进行初始化:

$$\begin{aligned} V &= \mathcal{N}(0,1) \\ [0054] \quad K &= \text{QueryNet}(V) \\ E &= \text{KNN}(K) \end{aligned} \quad (1)$$

[0055] 其中, $\mathcal{N}(0,1)$ 为均值为0,方差为1的正态分布,即,图结构中的多个隐状态值 $\{v_i\}$ 服从正态分布。 $\text{QueryNet}(V)$ 为多层感知器,可通过多层感知器分别对各隐状态值进行处理,确定图结构中各隐状态值的坐标 k_i 。 KNN 为K-Nearest Neighbors(K-近邻)聚类算法,可通过隐状态值的坐标 k_i 确定图结构的边索引。

[0056] 在一种可能的实现方式中,可通过时空记忆神经网络对隐状态 z_t 进行转换,获得隐空间中的坐标 k_t 。进一步地,可确定隐空间中与坐标 k_t 的特征距离最近的隐状态值 v_t 的坐标 k_{t_1} 。其中,隐状态值 v_t 可被确定为第一隐状态值。

[0057] 在一种可能的实现方式中,可通过以下公式(2)确定状态传递函数:

$$[0058] \quad \text{Message}(v_t) = \text{Aggregate}(v_j) \text{ s.t. } (v_t, v_j) \in E \quad (2)$$

[0059] 其中, $\text{Aggregate}(v_j)$ 为隐空间中多个隐状态值的平均值, $\text{Message}(v_t)$ 表示以 z_t 为中心的局部隐空间中的历史隐状态。

[0060] 在一种可能的实现方式中,可通过第一隐状态值和隐状态,确定当前周期的状态信息,例如,可通过以下公式(3)确定当前周期的状态信息 s_t :

$$[0061] \quad s_t = z_t + \text{Message}(v_t) \quad (3)$$

[0062] 通过这种方式,可利用多个历史隐状态来求解当前周期的状态信息,可在确定状态信息时考虑了历史周期中状态信息的影响,可在确定状态信息时考虑真实环境中的复杂动态在时间尺度上的长程依赖和空间尺度上的相互作用等因素,提升状态信息的准确性。

[0063] 在一种可能的实现方式中,在确定当前周期的状态信息 s_t 后,所述方法还包括:根据所述当前周期的状态信息,对历史周期中的第一隐状态值进行更新,获得当前周期的第一隐状态值。即,利用获得的当前周期的状态信息 s_t ,更新隐空间中的隐状态值的坐标,即,

使 $v_t = s_t$, 并使 $k_t = \text{QueryNet}(v_t)$ 。以使得在后续周期中, 可将当前周期的隐状态作为历史隐状态, 来求解后续周期的隐状态, 从而使得后续周期的状态信息的求解过程中, 可利用更准确的历史隐状态来进行求解。

[0064] 在一种可能的实现方式中, 在获得当前周期的状态信息后, 可对未来周期的状态信息进行预测, 获得未来多个周期的状态信息。再此预测的未来多个周期的状态信息为预测值, 与未来多个周期真实的状态信息之间可能存在误差。步骤S12可包括: 根据所述当前周期的状态信息, 确定当前周期的状态信息转移至下一个周期的状态信息的马尔科夫转移概率; 根据所述当前周期的状态信息转移至下一个周期的状态信息的马尔科夫转移概率, 以及所述当前周期的状态信息, 确定下一个周期的状态信息; 根据所述下一个周期的状态信息, 确定所述未来多个周期的状态信息。

[0065] 在一种可能的实现方式中, 可通过循环状态空间模型 (Recurrent State Space Model, RSSM) 来确定估计当前周期的状态信息 s_t 转移至下一个周期的状态信息 s_{t+1} 的马尔科夫转移概率 $p(s_{t+1}|s_t)$ 。基于马尔科夫转移概率 $p(s_{t+1}|s_t)$, 神经网络可预测下一个周期的状态信息 s_{t+1} 。

[0066] 在一种可能的实现方式中, 所述预测过程可通过以下公式 (4) 来预测未来多个周期的状态信息:

$$[0067] \quad \text{STI}(s_{t+h+1}|s_{t+h}, s_{t+h-1}) = \begin{cases} \text{RSSM}(s_{t+h}, s_{t+h-1}) & K \neq 0 \\ \text{RSSM}(\text{STM}(s_{t+h}), s_{t+h-1}) & K = 0 \end{cases} \quad (4)$$

[0068] 其中, $h \geq 0$, 且为整数。 K 为每个周期的时间步 (即, 时间间隔), 通过以上公式 (4), 可通过上一个周期 (历史周期) 的状态信息和当前周期的状态信息, 确定马尔科夫转移概率 $p(s_{t+1}|s_t)$, 进而确定下一个周期的状态信息 s_{t+1} 。可迭代执行上述过程, 确定 s_{t+2} 、 s_{t+3} 等未来多个周期的状态信息。

[0069] 通过这种方式, 可利用历史周期的状态信息以及当前周期的状态信息来预测未来周期的状态信息, 可在预测未来周期的状态信息时考虑了历史周期中状态信息的影响, 可在预测状态信息时考虑真实环境中的复杂动态在时间尺度上的长程依赖和空间尺度上的相互作用等因素, 提升预测状态信息的准确性。

[0070] 在一种可能的实现方式中, 在步骤S13中, 在获得当前周期的状态信息 s_t , 以及预测的未来多个周期的状态信息 s_{t+1} 、 s_{t+2} ... s_{t+h} ...后, 可基于上述状态信息, 获得当前周期的动作信息 a_t 以及预测的未来多个周期的动作信息 a_{t+1} 、 a_{t+2} ... a_{t+h} ...。在示例中, 可将当前周期的状态信息 s_t 输入上述神经网络, 可获得当前周期的状态信息 s_t , 可将预测的下一个周期的状态信息 s_{t+1} 输入上述神经网络, 可获得预测的下一个周期的动作信息 a_{t+1} ...即, 可通过神经网络分别获得当前周期以及预测的未来多个周期的动作信息。

[0071] 在一种可能的实现方式中, 在步骤S14中, 可确定未来多个周期的奖励信息, 例如, 所述神经网络包括贝叶斯网络, 可将当前周期的状态信息 s_t 和当前周期的动作信息 a_t 输入贝叶斯网络, 获得下一个周期的奖励信息 \tilde{r}_{t+1} 。例如, 可通过贝叶斯网络 \tilde{R} , 对状态信息 s_t 和动作信息 a_t 进行处理, 获得奖励信息 \tilde{r}_{t+1} 所服从的概率分布 $p(\tilde{r}_{t+1}|s_t, a_t)$, 并基于该概率分布确定奖励信息 \tilde{r}_{t+1} 。

[0072] 在一种可能的实现方式中, 还可进一步预测未来多个周期的奖励信息 \tilde{r}_{t+2} 、 \tilde{r}_{t+3} ...

步骤S14可包括：根据所述当前周期的动作信息及历史周期的动作信息，以及所述当前周期的状态信息、所述历史周期的状态信息及下一个周期的状态信息，确定相对熵；根据所述相对熵，确定所述确定未来多个周期的奖励信息。

[0073] 在一种可能的实现方式中，为了让机器人能够自行探索，以更快适应新环境，可将奖励信息进行分解，例如，分解为环境奖励信息和自驱力奖励信息，自驱力奖励信息可使机器人从动作信息和状态信息中获取奖励信息（例如，可确定动作信息和状态信息或是否可带来正的收益，即，是否可使环境与目标状态的差距缩小）。在示例中，奖励信息可通过以下公式（5）确定：

$$[0074] \quad r = r_{\text{env}} - \text{KL}[p(s, a) \parallel \tau(s, a)] \quad (5)$$

[0075] 其中， r 为未来任意周期的奖励信息， r_{env} 为未来任意周期的环境奖励信息， $\text{KL}[p(s, a) \parallel \tau(s, a)]$ 为未来任意周期的自驱力奖励信息，可用实际概率分布 $p(s, a)$ 与目标概率分布 $\tau(s, a)$ 之间的相对熵表示。 s 为未来任意周期的状态信息， a 为未来任意周期的动作信息，实际概率分布 $p(s, a)$ 为在状态信息为 s ，动作信息为 a 的情况下，外部环境服从的概率分布，目标概率分布为在状态信息为 s ，动作信息为 a 的情况下，目标状态所服从的概率分布。

[0076] 在一种可能的实现方式中，可通过训练，使得相对熵 $\text{KL}[p(s, a) \parallel \tau(s, a)]$ 最小化，使得奖励信息最大化，以实现设置自驱力奖励信息的目的，同时，在使相对熵最小化的过程中，可使机器人的神经网络对状态信息和动作信息进行探索，以更快适应新环境。以下对相对熵进行分解，以求解所述相对熵。

[0077] 在一种可能的实现方式中，根据所述当前周期的动作信息及历史周期的动作信息，以及所述当前周期的状态信息、所述历史周期的状态信息及下一个周期的状态信息，确定相对熵，包括：确定所述当前周期的动作信息及历史周期的动作信息组成的动作序列，以及所述当前周期的状态信息、所述历史周期的状态信息及下一个周期的状态信息组成的状态序列之间的第一互信息；根据所述第一互信息确定所述相对熵。

[0078] 在一种可能的实现方式中，当前周期的动作信息及历史周期的动作信息组成的动作序列为 $a_1, a_2 \cdots a_t$ ，当前周期的状态信息、历史周期的状态信息及下一个周期的状态信息组成的状态序列为 $s_1, s_2 \cdots s_t, s_{t+1}$ ，可确定动作序列和状态序列之间的互信息 $J(s_{t+1}; a_t | s_t)$ ，例如，可确定上述两个序列之间的相对熵的表达式，在该相对熵最大的情况下，可确定该最大的相对熵为所述第一互信息。在示例中，在这种情况下，实际概率分布 $p(s, a)$ 和目标概率分布 $\tau(s, a)$ 可通过以下公式（6）表示：

$$[0079] \quad \begin{aligned} p(s, a) &= \prod_t p(s_{t+1} | s_t, a_t) \pi(a_t | s_t) \\ \tau(s, a) &\propto \prod_t \tau(a_t | s_{t+1}, s_t, a_{t-1}) \end{aligned} \quad (6)$$

[0080] 其中， π 为通过神经网络的处理过程， p 为第 t 个周期的实际概率分布。 τ 为第 t 个周期的目标概率分布。

[0081] 进一步地，可基于公式（6）所示的实际概率分布 $p(s, a)$ 与目标概率分布 $\tau(s, a)$ ，来确定二者之间的相对熵 $\text{KL}[p(s, a) \parallel \tau(s, a)]$ 。进而可根据相对熵 $\text{KL}[p(s, a) \parallel \tau(s, a)]$ 和公式（5），确定未来各周期的奖励信息。

[0082] 在一种可能的实现方式中，根据所述当前周期的动作信息及历史周期的动作信息，以及所述当前周期的状态信息、所述历史周期的状态信息及下一个周期的状态信息，确定相对熵，包括：根据所述当前周期的动作信息及历史周期的动作信息，以及所述当前周期

的状态信息及所述历史周期的状态信息,确定技能序列;确定所述技能序列,和所述当前周期的动作信息及历史周期的动作信息组成的动作序列之间的第二互信息;根据所述第二互信息确定所述相对熵。

[0083] 在一种可能的实现方式中,各周期的动作信息 a 的获得方式为神经网络对各周期的状态信息进行处理,获得动作信息 a 的概率分布,并基于 a 的概率分布获得动作信息 a ,在所述动作信息 a 的概率分布中,获得动作信息 a 的先验概率为 $p(z)$,因此,可基于个周期的动作信息以及状态信息,获得先验概率的序列,即,技能序列 $z_1, z_2 \cdots z_t$ 。

[0084] 在一种可能的实现方式中,可确定技能序列 $z_1, z_2 \cdots z_t$ 与动作序列 $a_1, a_2 \cdots a_t$ 之间的第二互信息 $\mathcal{J}(s_{t+1}; z_t | s_t)$ 。例如,可确定上述两个序列之间的相对熵的表达式,在该相对熵最大的情况下,可确定该最大的相对熵为所述第二互信息。在这种情况下,相对熵 $KL[p(s, a) || \tau(s, a)]$ 可最小化, $KL[p(s, a) || \tau(s, a)]$ 可通过以下公式(7)表示:

$$[0085] \quad KL[p(s, a) || \tau(s, a)] = KL[p(s|a) || \tau(s)] - KL[\pi(a|s, z) || \tau(a)] - \mathbb{E}[\ln \tau(z|s) - \ln p(z)] \quad (7)$$

[0086] 将 $KL[p(s, a) || \tau(s, a)]$ 最小化的过程可通过DIAYN (Diversity is All You Need) 等无监督学习算法来进行,在最小化过程完成后,实际概率分布 $p(s, a)$ 和目标概率分布 $\tau(s, a)$ 可通过以下公式(8)表示:

$$[0087] \quad \begin{aligned} p(s, a, z) &= \prod_{k=1}^{t/k} p(z_k) \prod_t \pi(a_t | s_t, z_k) p(s_t | s_{t-1}, a_{t-1}) \\ \tau(s, a, z) &\propto \prod_{k=1}^{t/k} \tau(z_k | s) \prod_t \tau(a_t) \end{aligned} \quad (8)$$

[0088] 进一步地,可基于公式(8)所示的实际概率分布 $p(s, a)$ 与目标概率分布 $\tau(s, a)$,来确定二者之间的相对熵 $KL[p(s, a) || \tau(s, a)]$ 。进而可根据相对熵 $KL[p(s, a) || \tau(s, a)]$ 和公式(5),确定未来各周期的奖励信息。

[0089] 在一种可能的实现方式中,根据所述当前周期的动作信息及历史周期的动作信息,以及所述当前周期的状态信息和所述历史周期的状态信息,确定相对熵,包括:根据所述当前周期的状态信息、所述历史周期的状态信息及下一个周期的状态信息,确定全局特征;根据所述全局特征、所述当前周期的动作信息及历史周期的动作信息,以及所述当前周期的状态信息、所述历史周期的状态信息及下一个周期的状态信息,确定所述相对熵。

[0090] 在一种可能的实现方式中,各周期的状态信息可组成状态信息序列 $s_1, s_2 \cdots s_t$,可通过神经网络来提取状态信息序列 $s_1, s_2 \cdots s_t$ 的全局特征 w ,例如,可通过神经网络获得全局特征的概率分布 $p(w)$,进而基于全局特征的概率分布 $p(w)$ 确定全局特征 w 。全局特征为当前和历史状态的特征,可影响未来的状态,例如,在确定马尔科夫转移概率时,表示历史状态的全局特征可影响当前的状态,表示当前和历史状态的全局特征,可影响未来的状态等。

[0091] 在一种可能的实现方式中,可基于全局特征来优化 $KL[p(s, a) || \tau(s, a)]$,即,使 $KL[p(s, a) || \tau(s, a)]$ 最小化,可通过以下公式(9)来表示 $KL[p(s, a) || \tau(s, a)]$ 的优化过程的边界条件:

$$[0092] \quad \begin{aligned} KL[p(s, a) || \tau(s, a)] &\leq KL[p(w|s_t) || \\ \tau(w)] &- \mathbb{E}[\ln \tau(s_t | w) - \ln p(s_t)] + KL[p(s_{t+1} | s_t) || \\ \tau(s_{t+1} | s_t, w)] &- -\mathbb{E}[\ln \tau(w | s_{t+1}, s_t) - \ln p(w | s_t)] \end{aligned} \quad (9)$$

[0093] 在最小化过程完成后,实际概率分布 $p(s, a)$ 和目标概率分布 $\tau(s, a)$ 可通过以下公

式(10)表示:

$$\begin{aligned} p(s, w) &= p(w) \prod_t p(s_t | s_{t-1}) \\ \tau(s, w) &= \tau(w) \prod_t \tau(s_t | s_{t-1}, w) \end{aligned} \quad (10)$$

[0095] 进一步地,可基于公式(10)所示的实际概率分布 $p(s, a)$ 与目标概率分布 $\tau(s, a)$,来确定二者之间的相对熵 $\text{KL}[p(s, a) || \tau(s, a)]$ 。进而可根据相对熵 $\text{KL}[p(s, a) || \tau(s, a)]$ 和公式(5),确定未来各周期的奖励信息。

[0096] 在一种可能的实现方式中,在步骤S15中,可基于各周期的动作信息、各周期的状态信息以及各周期的奖励信息,来训练神经网络。步骤S15可包括:根据所述未来多个周期的奖励信息、所述当前周期的动作信息及所述未来多个周期的动作信息,以及所述当前周期的状态信息及所述未来多个周期的状态信息,确定奖励价值信息;根据所述奖励价值信息,训练所述神经网络。

[0097] 在一种可能的实现方式中,可基于各周期的动作信息、各周期的状态信息以及各周期的奖励信息确定长期规划,例如,可通过以下公式(10)确定长期规划LTP(Long-Term Planning):

$$\begin{aligned} S &= \{s_{t+H}, \dots, s_{t+1}\} \sim p(s_{t+H}, \dots, s_{t+1} | s_t, s_{t-1}) \\ A &= \{a_{t+H-1}, \dots, a_t\} \sim \prod_{h=0}^H \pi(a_{t+h} | s_{t+h}) \\ R &= \{\tilde{r}_{t+H}, \dots, \tilde{r}_{t+1}\} \sim \prod_{h=0}^H p(\tilde{r}_{t+h+1} | s_{t+h}, a_{t+h}) \\ \text{LTP}(s_t, s_{t-1}) &= S \cup A \cup R \end{aligned} \quad (10)$$

[0099] 其中,H为正整数,S为包括未来多个周期的状态信息的集合,A为包括未来多个周期的动作信息的集合,R为包括未来多个周期的奖励信息的集合,长期规划LTP为包括上述集合的并集的集合,即,包括多个周期的状态信息、动作信息和奖励信息。

[0100] 在一种可能的实现方式中,可基于长期规划LTP来确定奖励价值信息,例如,可根据以下公式(11)确定奖励价值信息 V_π :

$$V_\pi = \sum_t \pi(a_t | s_t) \mathbb{E}[\tilde{r}_{t+1} + \gamma V(s_{t+1}) | s_t] \quad (11)$$

[0102] 其中,V为通过具有多层感知机制的神经网络对状态信息进行处理获得价值信息的函数, γ 为折现系数。可通过长期规划LTP中的各项参数来计算奖励价值信息 V_π 。在对神经网络进行优化的过程中,可使神经网络按照奖励价值信息 V_π 最大化为目标来优化神经网络的参数。在经过多个周期的训练后,可获得能够使奖励价值信息符合训练要求的神经网络,例如,训练要求可包括使奖励价值信息大于或等于预设阈值等,本公开对训练要求不做限制。

[0103] 在完成训练后,可使用训练后的神经网络来进行机器人的控制,使机器人在一定的环境中运行,并进行多个周期的动作,以对环境产生影响,并使得环境逐步接近目标状态。例如,使机器人能够在工业生产环境中运行,并产生动作来影响工业生产的进度,使工业生产环境逐步接近生产完成的状态。

[0104] 在一种可能的实现方式中,本公开提供了一种机器人控制方法,其特征在于,包括:将当前周期的环境观测信息、上一个周期的动作信息、上一个周期的状态信息和当前周期的奖励信息输入上述的神经网络训练方法训练后的神经网络,获得当前周期的动作信息。

[0105] 在示例中,可将当前周期的环境观测信息 o_t ,上一个周期的动作信息 a_{t-1} ,上一个

周期的状态信息 s_{t-1} 和当前周期的奖励信息 \tilde{r}_t 输入上述训练后的神经网络,可获得当前周期的动作信息 a_t ,该动作信息可对当前周期的环境产生影响,使得环境发生变化,基于该变化与目标状态,可获得下一个周期的 \tilde{r}_{t+1} 。进一步地,可获得变化后的环境观测信息 o_{t+1} ,以及状态信息 s_t ,并可将动作信息 a_t ,下一个周期的奖励信息 \tilde{r}_{t+1} 输入神经网络,以输出下一个周期的动作信息 a_{t+1} ,并可对环境产生影响……最终可使得环境达到或接近目标状态。

[0106] 根据本公开的实施例的神经网络训练方法,可利用多个历史隐状态来求解当前周期的状态信息,可在确定状态信息时考虑了历史周期中状态信息的影响。并可可通过当前周期的状态信息以及历史周期的状态信息,预测未来多个周期的状态信息,因而在预测未来周期的状态信息时,考虑历史周期的影响,即,考虑了时间尺度上的长程依赖、空间尺度上的相互作用,使得动作信息不易陷入单一的模式。并且,可基于未来多个周期的动作信息以及未来多个周期的状态信息来训练神经网络,在训练过程中使神经网络获得对训练过程的全局认知,提高训练效率,且易于适应新环境。

[0107] 图2示出根据本公开实施例的神经网络训练方法的应用示意图,如图2所示,可通过时空记忆神经网络对当前周期的环境观测信息 o_t 以及上一个周期的动作信息 a_{t-1} 进行处理,获得隐状态 z_t ,并基于公式(3)可确定当前周期的状态信息 s_t 。

[0108] 在一种可能的实现方式中,可通过公式(4),即循环状态空间模型,来估计周期的状态信息 s_t 转移至下一个周期的状态信息 s_{t+1} 的马尔科夫转移概率,并迭代执行该过程,获得未来多个周期的状态信息。

[0109] 在一种可能的实现方式中,可将当前及未来多个周期的状态信息输入神经网络,获得预测的未来多个周期的动作信息 a_{t+1} 、 a_{t+2} … a_{t+h} …。

[0110] 在一种可能的实现方式中,可将当前周期的状态信息 s_t 和当前周期的动作信息 a_t 输入贝叶斯网络,获得下一个周期的奖励信息 \tilde{r}_{t+1} 。并可基于多个历史周期的状态信息和动作信息确定相对熵 $KL[p(s,a) || \tau(s,a)]$,例如,可基于公式(6)、(8)或(10)获得 $p(s,a)$ 和 $\tau(s,a)$,并基于公式(5)确定未来多个周期的奖励信息。

[0111] 进一步地,可基于未来多个周期的动作信息、未来多个周期的奖励信息和未来多个周期的状态信息获得长期规划LTP,并基于LTP和奖励价值信息,即,公式(11)来优化神经网络。即,按照使奖励价值信息最大化的方向调节神经网络的网络参数,在符合训练条件后,获得训练后的神经网络。

[0112] 在一种可能的实现方式中,训练后的神经网络可用于机器人的控制中,即,基于当前周期的环境观测信息,上一个周期的动作信息,上一个周期的状态信息和当前周期的奖励信息生成当前周期的动作信息。当前周期的动作信息可对环境产生影响,并使环境逐步接近目标状态。例如,可通过神经网络控制工业生产中的机器人,是机器人产生动作对工业生产环境造成影响,工业生产环境逐步接近生产完成的状态。本公开对神经网络训练方法的应用领域不做限制。

[0113] 图3示出根据本公开实施例的神经网络训练装置的框图,如图3所示,所述装置包括:环境确定模块11,用于将当前周期的环境观测信息以及上一个周期的动作信息输入神经网络,确定当前周期的状态信息,所述环境观测信息用于描述环境,所述环境包括机器人的运行环境,所述动作信息用于作用于当前周期的环境中,对环境造成改变,所述动作信息

包括用于控制机器人动作的控制信息,所述状态信息用于描述机器人的控制状态;环境预测模块12,用于根据所述当前周期的状态信息以及历史周期的状态信息,预测未来多个周期的状态信息;动作确定模块13,用于根据所述当前周期的状态信息及所述未来多个周期的状态信息,确定当前周期的动作信息以及未来多个周期的动作信息;奖励确定模块14,用于根据所述当前周期的动作信息及历史周期的动作信息,以及所述当前周期的状态信息、所述历史周期的状态信息及下一个周期的状态信息,确定未来多个周期的奖励信息;训练模块15,用于根据所述当前周期的动作信息及所述未来多个周期的动作信息、所述当前周期的状态信息及所述未来多个周期的状态信息,以及未来多个周期的奖励信息,对所述神经网络进行训练。

[0114] 在一种可能的实现方式中,所述环境确定模块进一步用于根据所述当前周期的环境观测信息以及所述上一个周期的动作信息,确定隐状态;根据所述隐状态确定历史周期中与所述隐状态特征距离最近的第一隐状态值;根据所述第一隐状态值和所述隐状态,确定所述当前周期的状态信息。

[0115] 在一种可能的实现方式中,所述装置还包括:更新模块,用于根据所述当前周期的状态信息,对历史周期中的第一隐状态值进行更新,获得当前周期的第一隐状态值。

[0116] 在一种可能的实现方式中,所述环境预测模块进一步用于:根据所述当前周期的状态信息和历史周期的状态信息,确定当前周期的状态信息转移至下一个周期的状态信息的马尔科夫转移概率;根据所述当前周期的状态信息转移至下一个周期的状态信息的马尔科夫转移概率,以及所述当前周期的状态信息,确定下一个周期的状态信息;根据所述下一个周期的状态信息,确定所述未来多个周期的状态信息。

[0117] 在一种可能的实现方式中,所述奖励确定模块进一步用于:根据所述当前周期的动作信息及历史周期的动作信息,以及所述当前周期的状态信息、所述历史周期的状态信息及下一个周期的状态信息,确定相对熵;根据所述相对熵,确定所述确定未来多个周期的奖励信息。

[0118] 在一种可能的实现方式中,所述奖励确定模块进一步用于:确定所述当前周期的动作信息及历史周期的动作信息组成的动作序列,以及所述当前周期的状态信息、所述历史周期的状态信息及下一个周期的状态信息组成的状态序列之间的第一互信息;根据所述第一互信息确定所述相对熵。

[0119] 在一种可能的实现方式中,所述奖励确定模块进一步用于:根据所述当前周期的动作信息及历史周期的动作信息,以及所述当前周期的状态信息及所述历史周期的状态信息,确定技能序列;确定所述技能序列,和所述当前周期的动作信息及历史周期的动作信息组成的动作序列之间的第二互信息;根据所述第二互信息确定所述相对熵。

[0120] 在一种可能的实现方式中,所述奖励确定模块进一步用于:根据所述当前周期的状态信息和所述历史周期的状态信息,确定全局特征;根据所述全局特征、所述当前周期的动作信息及历史周期的动作信息,以及所述当前周期的状态信息、所述历史周期的状态信息及下一个周期的状态信息,确定所述相对熵。

[0121] 在一种可能的实现方式中,所述训练模块进一步用于:根据所述未来多个周期的奖励信息、所述当前周期的动作信息及所述未来多个周期的动作信息,以及所述当前周期的状态信息及所述未来多个周期的状态信息,确定奖励价值信息;根据所述奖励价值信息,

训练所述神经网络。

[0122] 一种机器人控制装置,包括:控制模块,用于将当前周期的环境观测信息、上一个周期的动作信息、上一个周期的状态信息和当前周期的奖励信息输入根据所述神经网络训练装置训练后的神经网络,获得当前周期的动作信息。

[0123] 可以理解,本公开提及的上述各个方法实施例,在不违背原理逻辑的情况下,均可以彼此相互结合形成结合后的实施例,限于篇幅,本公开不再赘述。本领域技术人员可以理解,在具体实施方式的上述方法中,各步骤的具体执行顺序应当以其功能和可能的内在逻辑确定。

[0124] 此外,本公开还提供了神经网络训练装置、电子设备、计算机可读存储介质、程序,上述均可用来实现本公开提供的任一种神经网络训练方法,相应技术方案和描述和参见方法部分的相应记载,不再赘述。

[0125] 在一些实施例中,本公开实施例提供的装置具有的功能或包含的模块可以用于执行上文方法实施例描述的方法,其具体实现可以参照上文方法实施例的描述,为了简洁,这里不再赘述。

[0126] 本公开实施例还提出一种计算机可读存储介质,其上存储有计算机程序指令,所述计算机程序指令被处理器执行时实现上述方法。计算机可读存储介质可以是非易失性计算机可读存储介质。

[0127] 本公开实施例还提出一种电子设备,包括:处理器;用于存储处理器可执行指令的存储器;其中,所述处理器被配置为调用所述存储器存储的指令,以执行上述方法。

[0128] 本公开实施例还提供了一种计算机程序产品,包括计算机可读代码,当计算机可读代码在设备上运行时,设备中的处理器执行用于实现如上任一实施例提供的神经网络训练方法的指令。

[0129] 本公开实施例还提供了另一种计算机程序产品,用于存储计算机可读指令,指令被执行时使得计算机执行上述任一实施例提供的神经网络训练方法的操作。

[0130] 电子设备可以被提供为终端、服务器或其它形态的设备。

[0131] 图4示出根据本公开实施例的一种电子设备800的框图。例如,电子设备800可以是移动电话,计算机,数字广播终端,消息收发设备,游戏控制台,平板设备,医疗设备,健身设备,个人数字助理等终端。

[0132] 参照图4,电子设备800可以包括以下一个或多个组件:处理组件802,存储器804,电源组件806,多媒体组件808,音频组件810,输入/输出(I/O)的接口812,传感器组件814,以及通信组件816。

[0133] 处理组件802通常控制电子设备800的整体操作,诸如与显示,电话呼叫,数据通信,相机操作和记录操作相关联的操作。处理组件802可以包括一个或多个处理器820来执行指令,以完成上述的方法的全部或部分步骤。此外,处理组件802可以包括一个或多个模块,便于处理组件802和其他组件之间的交互。例如,处理组件802可以包括多媒体模块,以方便多媒体组件808和处理组件802之间的交互。

[0134] 存储器804被配置为存储各种类型的数据以支持在电子设备800的操作。这些数据的示例包括用于在电子设备800上操作的任何应用程序或方法的指令,联系人数据,电话簿数据,消息,图片,视频等。存储器804可以由任何类型的易失性或非易失性存储设备或者它

们的组合实现,如静态随机存取存储器(SRAM),电可擦除可编程只读存储器(EEPROM),可擦除可编程只读存储器(EPROM),可编程只读存储器(PROM),只读存储器(ROM),磁存储器,快闪存储器,磁盘或光盘。

[0135] 电源组件806为电子设备800的各种组件提供电力。电源组件806可以包括电源管理系统,一个或多个电源,及其他与为电子设备800生成、管理和分配电力相关联的组件。

[0136] 多媒体组件808包括在所述电子设备800和用户之间的提供一个输出接口的屏幕。在一些实施例中,屏幕可以包括液晶显示器(LCD)和触摸面板(TP)。如果屏幕包括触摸面板,屏幕可以被实现为触摸屏,以接收来自用户的输入信号。触摸面板包括一个或多个触摸传感器以感测触摸、滑动和触摸面板上的手势。所述触摸传感器可以不仅感测触摸或滑动动作的边缘,而且还检测与所述触摸或滑动操作相关的持续时间和压力。在一些实施例中,多媒体组件808包括一个前置摄像头和/或后置摄像头。当电子设备800处于操作模式,如拍摄模式或视频模式时,前置摄像头和/或后置摄像头可以接收外部的多媒体数据。每个前置摄像头和后置摄像头可以是一个固定的光学透镜系统或具有焦距和光学变焦能力。

[0137] 音频组件810被配置为输出和/或输入音频信号。例如,音频组件810包括一个麦克风(MIC),当电子设备800处于操作模式,如呼叫模式、记录模式和语音识别模式时,麦克风被配置为接收外部音频信号。所接收的音频信号可以被进一步存储在存储器804或经由通信组件816发送。在一些实施例中,音频组件810还包括一个扬声器,用于输出音频信号。

[0138] I/O接口812为处理组件802和外围接口模块之间提供接口,上述外围接口模块可以是键盘,点击轮,按钮等。这些按钮可包括但不限于:主页按钮、音量按钮、启动按钮和锁定按钮。

[0139] 传感器组件814包括一个或多个传感器,用于为电子设备800提供各个方面的状态评估。例如,传感器组件814可以检测到电子设备800的打开/关闭状态,组件的相对定位,例如所述组件为电子设备800的显示器和小键盘,传感器组件814还可以检测电子设备800或电子设备800一个组件的位置改变,用户与电子设备800接触的存在或不存在,电子设备800方位或加速/减速和电子设备800的温度变化。传感器组件814可以包括接近传感器,被配置用来在没有任何的物理接触时检测附近物体的存在。传感器组件814还可以包括光传感器,如CMOS或CCD图像传感器,用于在成像应用中使用。在一些实施例中,该传感器组件814还可以包括加速度传感器,陀螺仪传感器,磁传感器,压力传感器或温度传感器。

[0140] 通信组件816被配置为便于电子设备800和其他设备之间有线或无线方式的通信。电子设备800可以接入基于通信标准的无线网络,如WiFi,2G或3G,或它们的组合。在一个示例性实施例中,通信组件816经由广播信道接收来自外部广播管理系统的广播信号或广播相关信息。在一个示例性实施例中,所述通信组件816还包括近场通信(NFC)模块,以促进短程通信。例如,在NFC模块可基于射频识别(RFID)技术,红外数据协会(IrDA)技术,超宽带(UWB)技术,蓝牙(BT)技术和其他技术来实现。

[0141] 在示例性实施例中,电子设备800可以被一个或多个应用专用集成电路(ASIC)、数字信号处理器(DSP)、数字信号处理设备(DSPD)、可编程逻辑器件(PLD)、现场可编程门阵列(FPGA)、控制器、微控制器、微处理器或其他电子元件实现,用于执行上述方法。

[0142] 在示例性实施例中,还提供了一种非易失性计算机可读存储介质,例如包括计算机程序指令的存储器804,上述计算机程序指令可由电子设备800的处理器820执行以完成

上述方法。

[0143] 图5示出根据本公开实施例的一种电子设备1900的框图。例如,电子设备1900可以被提供为一服务器。参照图5,电子设备1900包括处理组件1922,其进一步包括一个或多个处理器,以及由存储器1932所代表的存储器资源,用于存储可由处理组件1922的执行的指令,例如应用程序。存储器1932中存储的应用程序可以包括一个或一个以上的每一个对应于一组指令的模块。此外,处理组件1922被配置为执行指令,以执行上述方法。

[0144] 电子设备1900还可以包括一个电源组件1926被配置为执行电子设备1900的电源管理,一个有线或无线网络接口1950被配置为将电子设备1900连接到网络,和一个输入输出(I/O)接口1958。电子设备1900可以操作基于存储在存储器1932的操作系统,例如Windows Server™,Mac OS X™,Unix™,Linux™,FreeBSD™或类似。

[0145] 在示例性实施例中,还提供了一种非易失性计算机可读存储介质,例如包括计算机程序指令的存储器1932,上述计算机程序指令可由电子设备1900的处理组件1922执行以完成上述方法。

[0146] 本公开可以是系统、方法和/或计算机程序产品。计算机程序产品可以包括计算机可读存储介质,其上载有用于使处理器实现本公开的各个方面的计算机可读程序指令。

[0147] 计算机可读存储介质可以是可以保持和存储由指令执行设备使用的指令的有形设备。计算机可读存储介质例如可以是一一但不限于一一电存储设备、磁存储设备、光存储设备、电磁存储设备、半导体存储设备或者上述的任意合适的组合。计算机可读存储介质的更具体的例子(非穷举的列表)包括:便携式计算机盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、静态随机存取存储器(SRAM)、便携式压缩盘只读存储器(CD-ROM)、数字多功能盘(DVD)、记忆棒、软盘、机械编码设备、例如其上存储有指令的打孔卡或凹槽内凸起结构、以及上述的任意合适的组合。这里所使用的计算机可读存储介质不被解释为瞬时信号本身,诸如无线电波或者其他自由传播的电磁波、通过波导或其他传输媒介传播的电磁波(例如,通过光纤电缆的光脉冲)、或者通过电线传输的电信号。

[0148] 这里所描述的计算机可读程序指令可以从计算机可读存储介质下载到各个计算/处理设备,或者通过网络、例如因特网、局域网、广域网和/或无线网下载到外部计算机或外部存储设备。网络可以包括铜传输电缆、光纤传输、无线传输、路由器、防火墙、交换机、网关计算机和/或边缘服务器。每个计算/处理设备中的网络适配卡或者网络接口从网络接收计算机可读程序指令,并转发该计算机可读程序指令,以供存储在各个计算/处理设备中的计算机可读存储介质中。

[0149] 用于执行本公开操作的计算机程序指令可以是汇编指令、指令集架构(ISA)指令、机器指令、机器相关指令、微代码、固件指令、状态设置数据、或者以一种或多种编程语言的任意组合编写的源代码或目标代码,所述编程语言包括面向对象的编程语言—诸如Smalltalk、C++等,以及常规的过程式编程语言—诸如“C”语言或类似的编程语言。计算机可读程序指令可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中,远程计算机可以通过任意种类的网络—包括局域网(LAN)或广域网(WAN)—连接到用户计算机,或者,可以连接到外部计算机(例如利

用因特网服务提供商来通过因特网连接)。在一些实施例中,通过利用计算机可读程序指令的状态信息来个性化定制电子电路,例如可编程逻辑电路、现场可编程门阵列(FPGA)或可编程逻辑阵列(PLA),该电子电路可以执行计算机可读程序指令,从而实现本公开的各个方面。

[0150] 这里参照根据本公开实施例的方法、装置(系统)和计算机程序产品的流程图和/或框图描述了本公开的各个方面。应当理解,流程图和/或框图的每个方框以及流程图和/或框图中各方框的组合,都可以由计算机可读程序指令实现。

[0151] 这些计算机可读程序指令可以提供给通用计算机、专用计算机或其它可编程数据处理装置的处理器,从而生产出一种机器,使得这些指令在通过计算机或其它可编程数据处理装置的处理器执行时,产生了实现流程图和/或框图中的一个或多个方框中规定的功能/动作的装置。也可以把这些计算机可读程序指令存储在计算机可读存储介质中,这些指令使得计算机、可编程数据处理装置和/或其他设备以特定方式工作,从而,存储有指令的计算机可读介质则包括一个制品,其包括实现流程图和/或框图中的一个或多个方框中规定的功能/动作的各个方面的指令。

[0152] 也可以把计算机可读程序指令加载到计算机、其它可编程数据处理装置、或其它设备上,使得在计算机、其它可编程数据处理装置或其它设备上执行一系列操作步骤,以产生计算机实现的过程,从而使得在计算机、其它可编程数据处理装置、或其它设备上执行的指令实现流程图和/或框图中的一个或多个方框中规定的功能/动作。

[0153] 附图中的流程图和框图显示了根据本公开的多个实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上,流程图或框图中的每个方框可以代表一个模块、程序段或指令的一部分,所述模块、程序段或指令的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。在有些作为替换的实现中,方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如,两个连续的方框实际上可以基本并行地执行,它们有时也可以按相反的顺序执行,这依所涉及的功能而定。也要注意的,框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合,可以用执行规定的功能或动作的专用的基于硬件的系统来实现,或者可以用专用硬件与计算机指令的组合来实现。

[0154] 该计算机程序产品可以具体通过硬件、软件或其结合的方式实现。在一个可选实施例中,所述计算机程序产品具体体现为计算机存储介质,在另一个可选实施例中,计算机程序产品具体体现为软件产品,例如软件开发包(Software Development Kit, SDK)等等。

[0155] 以上已经描述了本公开的各实施例,上述说明是示例性的,并非穷尽性的,并且也不限于所披露的各实施例。在不偏离所说明的各实施例的范围和精神的情况下,对于本技术领域的普通技术人员来说许多修改和变更都是显而易见的。本文中所用术语的选择,旨在最好地解释各实施例的原理、实际应用或对市场中的技术的改进,或者使本技术领域的其它普通技术人员能理解本文披露的各实施例。

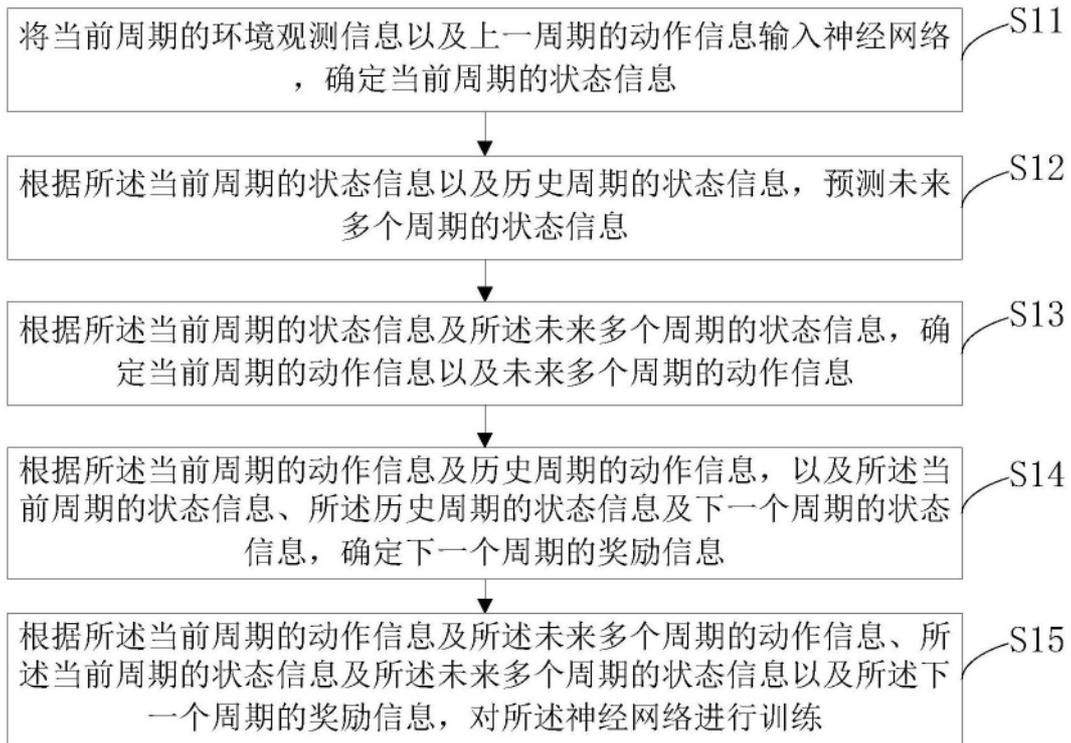


图1

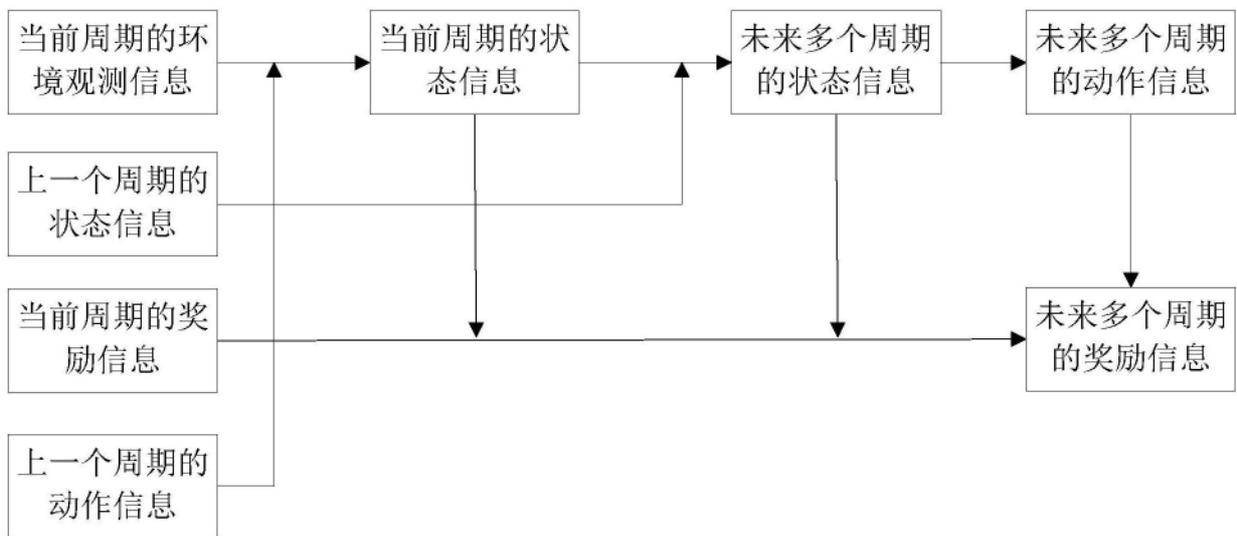


图2

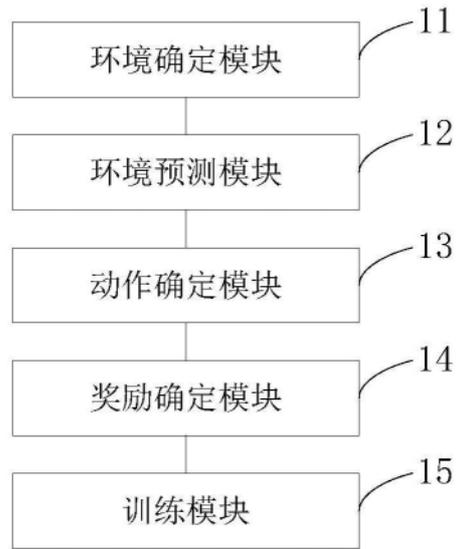


图3

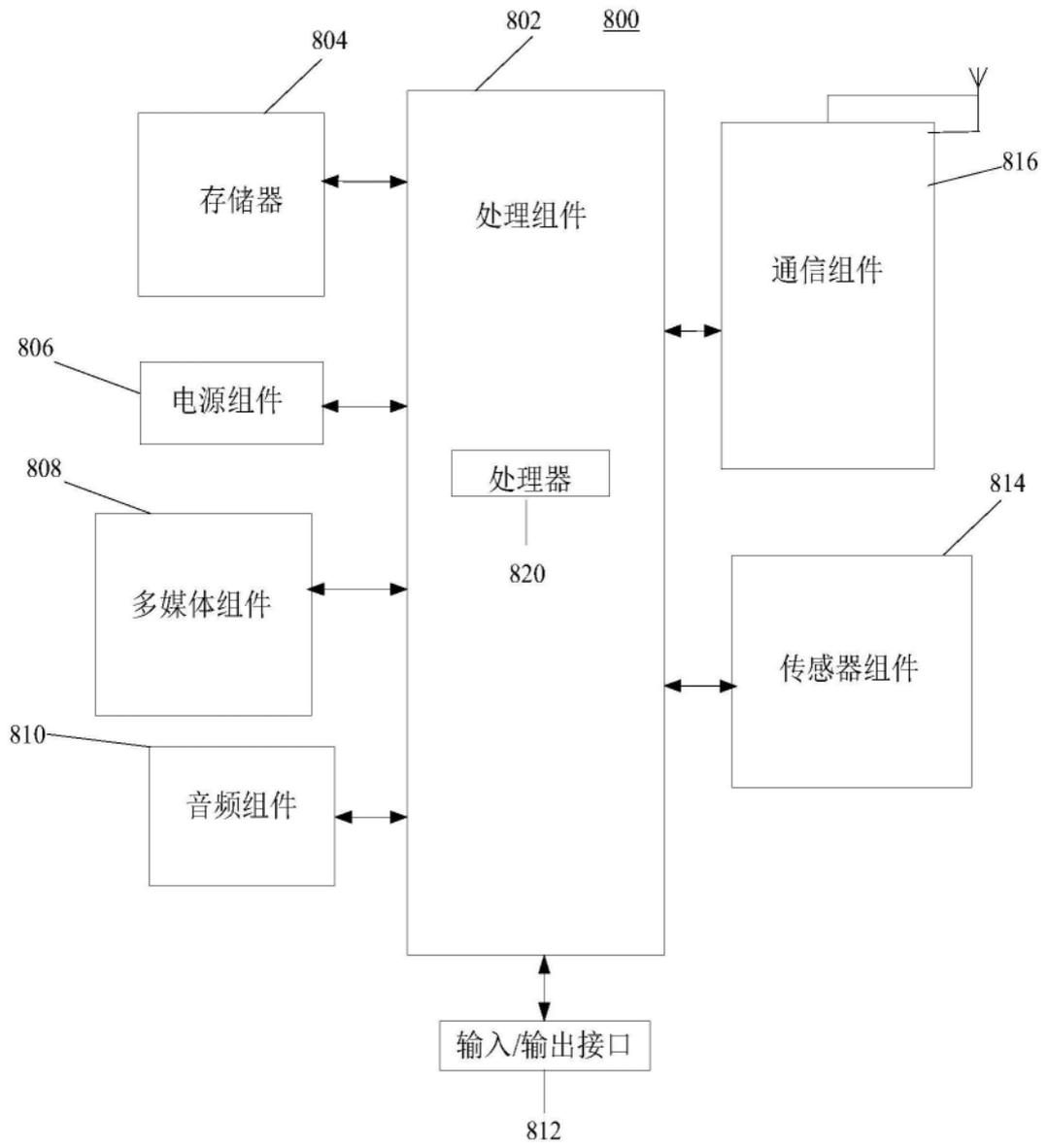


图4

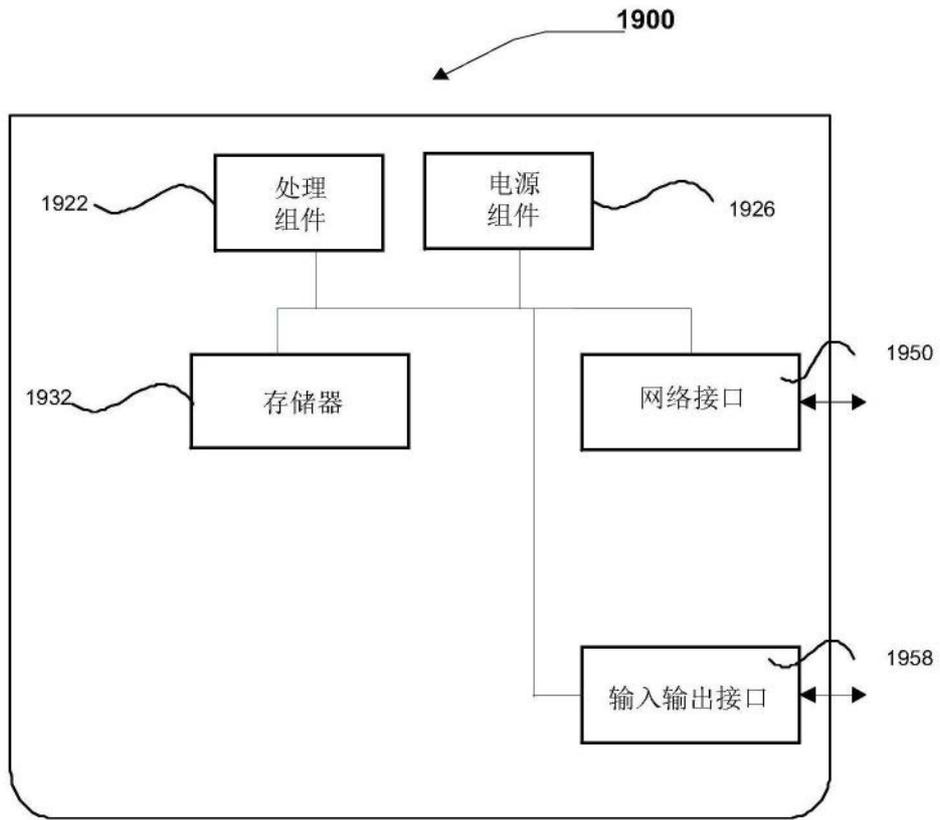


图5