(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 121143113 A
(43)申请公布日 2025.12.16

(21)申请号 202511204239.5

(22)申请日 2025.08.26

(71)申请人 清华大学
地址 100084 北京市海淀区清华园

(72)发明人 封硕 杨敬轩 张毅 王子航 姬浩元 姚丹亚

(74)专利代理机构 北京安信方达知识产权代理有限公司 11262
专利代理师 魏文佳 解婷婷

(51)Int.Cl.
G05B 19/042(2006.01)
G06N 3/0475(2023.01)
G06N 3/092(2023.01)
G06N 3/084(2023.01)
G06N 3/094(2023.01)

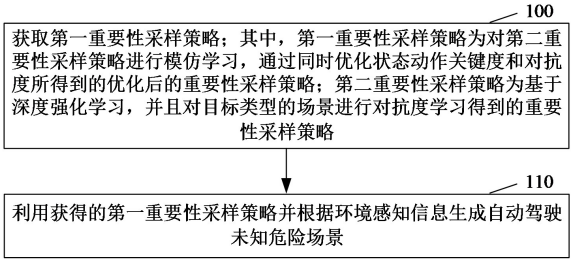
权利要求书3页 说明书15页 附图6页

(54)发明名称

一种自动驾驶未知危险场景的生成方法及电子设备

(57)摘要

一种自动驾驶未知危险场景的生成方法及电子设备，该方法包括获取第一重要性采样策略；其中，所述第一重要性采样策略为对第二重要性采样策略进行模仿学习，通过同时优化状态动作关键度和对抗度所得到的优化后的重要性采样策略；所述第二重要性采样策略为基于深度强化学习，并且对目标类型的场景进行对抗度学习得到的重要性采样策略；利用获得的第一重要性采样策略并根据环境感知信息生成自动驾驶未知危险场景，因此同时结合关键度与对抗度实现了采样策略的优化，提高了对目标类型场景的生成能力。



CN 121143113 A

1.一种自动驾驶未知危险场景的生成方法,其特征在于,包括:

获取第一重要性采样策略;其中,所述第一重要性采样策略为对第二重要性采样策略进行模仿学习,通过同时优化状态动作关键度和对抗度所得到的优化后的重要性采样策略;所述第二重要性采样策略为基于深度强化学习,并且对目标类型的场景进行对抗度学习得到的重要性采样策略;

利用获得的第一重要性采样策略并根据环境感知信息生成自动驾驶未知危险场景。

2.根据权利要求1所述的方法,其特征在于,所述第一重要性采样策略通过以下方式获得:

对模仿学习策略进行初始化;其中,所述模仿学习策略用于模仿学习所述第二重要性采样策略,并且所述模仿学习策略和所述第二重要性采样策略均包括:关键度和对抗度;

构建第一奖励函数;其中,所述第一奖励函数用于奖励所述模仿学习策略中的关键度逼近所述第二重要性采样策略中的关键度,且所述模仿学习策略中的对抗度逼近所述第二重要性采样策略中的对抗度;

从状态空间中获取多个初始状态作为第一初始状态S_{at},利用获得的多个第一初始状态和所构建的第一奖励函数更新所述模仿学习策略,得到第三重要性采样策略;

根据获得的第三重要性采样策略得到所述第一重要性采样策略。

3.根据权利要求2所述的方法,其特征在于,所述根据获得的第三重要性采样策略得到所述第一重要性采样策略,包括:

将获得的第三重要性采样策略作为所述第一重要性采样策略;

或者,

对获得的第三重要性采样策略再进行训练,并将经过训练得到的第三重要性采样策略作为所述第一重要性采样策略。

4.根据权利要求2所述的方法,其特征在于,所述利用获得的多个第一初始状态和所构建的第一奖励函数更新所述模仿学习策略,得到第三重要性采样策略,包括:

从获得的多个第一初始状态中任意选择一个第一初始状态,将经过初始化的模仿学习策略作为所选择的第一初始状态下的模仿学习策略,并执行以下第一策略更新操作;

所述第一策略更新操作包括:

根据第一奖励函数不断对所选择的第一初始状态下的模仿学习策略进行迭代更新,以获得所选择的第一初始状态下最终更新的模仿学习策略;在最终更新的模仿学习策略收敛的情况下,将所选择的第一初始状态下获得的最终更新的模仿学习策略作为所述第三重要性采样策略,在最终更新的模仿学习策略未收敛的情况下,从获得的多个第一初始状态中另外选择一个第一初始状态,并将之前所选择的第一初始状态下获得的最终更新的模仿学习策略,作为新选择的第一初始状态下的模仿学习策略,继续执行所述第一策略更新操作,直到新选择的第一初始状态下获得的更新的模仿学习策略收敛,将新选择的第一初始状态下获得的最终更新的模仿学习策略作为所述第三重要性采样策略。

5.根据权利要求4所述的方法,其特征在于,所述根据第一奖励函数不断对所选择的第一初始状态下的模仿学习策略进行迭代更新,以获得所选择的第一初始状态下最终更新的模仿学习策略,包括:

将所选择的第一初始状态下的初始时刻作为当前第一初始状态下的当前时刻,将所选

择的第一初始状态下的模仿学习策略作为所选择的第一初始状态下当前时刻的模仿学习策略,并执行以下第二策略更新操作;

所述第二策略更新操作包括:

根据所选择的第一初始状态下当前时刻的模仿学习策略从动作空间中采样动作并执行,以将所选择的第一初始状态从当前时刻转移至下一时刻,计算本次状态转移过程的第一奖励函数的奖励值;对获得的本次状态转移过程的第一奖励函数的奖励值进行优化,并基于优化后的第一奖励函数的奖励值反向更新当前时刻的模仿学习策略中的状态动作关键度和对抗度,得到更新的模仿学习策略;将下一时刻替代当前时刻作为新的当前时刻,将更新的模仿学习策略作为所选择的第一初始状态下新的当前时刻的模仿学习策略,继续执行所述第二策略更新操作,直到更新的模仿学习策略已收敛或当前时刻已是最大时刻。

6.根据权利要求4或5所述的方法,其特征在于,所述第一奖励函数设计如下:

$$R_1(s_t) = -\| Q_1(s_t, a_t) - Q_{D2RL}(s, a) \|_1 - \| \epsilon_1(s_t) - \epsilon_{D2RL}(s) \|_1;$$

其中,$R(s_t)$为t时刻的第一奖励函数,$Q_{D2RL}(s,a)$和$\epsilon_{D2RL}(s)$分别为所述第二重要性采样策略中的状态动作关键度和对抗度,$Q_1(s_t,a_t)$和$\epsilon_1(s_t)$分别为t时刻的模仿学习策略中的状态动作关键度和对抗度。

7.根据权利要求3所述的方法,其特征在于,所述对获得的第三重要性采样策略再进行训练,并将经过训练得到的第三重要性采样策略作为所述第一重要性采样策略,包括:

构建第二奖励函数;其中,所述第二奖励函数用于奖励第三重要性采样策略快速完成收敛;

从状态空间中获取多个初始状态作为第二初始状态S_{ct};

从获得的多个第二初始状态中任意选择一个第二初始状态,将第三重要性采样策略作为所选择的第二初始状态下的策略,并执行以下第三策略更新操作;

所述第三策略更新操作包括:

根据第二奖励函数不断对所选择的第二初始状态下的第三重要性采样策略进行迭代更新,以获得所选择的第二初始状态下最终更新的第三重要性采样策略;在最终更新的第三重要性采样策略收敛的情况下,将所选择的第一初始状态下获得的并且用第二初始状态S_{ct}更新后的第三重要性采样策略作为所述第一重要性采样策略,在最终更新的第三重要性采样策略未收敛的情况下,从获得的多个第二初始状态中另外选择一个第二初始状态,并将之前所选择的第二初始状态下获得的最终更新的第三重要性采样策略,作为新选择的第二初始状态下更新的第三重要性采样策略,继续执行所述第三策略更新操作,直到新选择的第二初始状态下获得的更新的第三重要性采样策略收敛,将新选择的第一初始状态下获得的最终更新的第三策略更新操作作为所述第一重要性采样策略。

8.根据权利要求7所述的方法,其特征在于,所述根据第二奖励函数不断对所选择的第二初始状态下的第三重要性采样策略进行迭代更新,以获得所选择的第二初始状态下最终更新的第三重要性采样策略,包括:

将所选择的第二初始状态下的初始时刻作为当前第二初始状态下的当前时刻,将所选择的第二初始状态下的第三重要性采样策略作为所选择的第二初始状态下当前时刻的第

三重要性采样策略,并执行以下第四策略更新操作;

所述第四策略更新操作包括:

根据所选择的第二初始状态下当前时刻的第三重要性采样策略从动作空间采样动作并执行,以将所选择的第二初始状态从当前时刻转移至下一时刻,计算本次状态转移过程的第二奖励函数的奖励值;对获得的本次状态转移过程的第二奖励函数的奖励值进行优化,并基于优化后的第二奖励函数的奖励值反向更新当前时刻的第三重要性采样策略中的状态动作关键度和对抗度,得到更新的第三重要性采样策略;将下一时刻替代当前时刻作为新的当前时刻,并将更新的第三重要性采样策略作为所选择的第二初始状态下新的当前时刻的第三重要性采样策略,继续执行所述第四策略更新操作,直到更新的第三重要性采样策略已收敛或当前时刻已是最大时刻。

9.根据权利要求7或8所述的方法,其特征在于,所述第二奖励函数设计如下:

$$R_2(s_t)=\begin{cases}max\{100-50000I_F(x)w(x),-100\},if\ t=T或者碰撞\\0,otherwise\end{cases};$$

$$w(x)=\prod_{t=0}^{T-1}w_t,且\ w_t=\frac{\pi_{base}(a_t|s_t)}{\pi_2(a_t|s_t)}=\frac{V(s_t)}{[1-\epsilon_2(s_t)]V(s_t)+\epsilon_2(s_t)Q_2(s_t,a_t)},t=0,\ldots,T;$$

$$V(s_t)=\sum_a Q_2(s_t,a_t)\pi_{base}(a_t\,|\,s_t),\forall s\in 状态空间\mathcal{S},a\in 动作空间\mathcal{A};$$

其中,$R_2(s_t)$为t时刻的第二奖励函数,$Q_2(s_t,a_t)$、$\epsilon_2(s_t)$分别为t时刻的第一重要性采样策略中的状态动作关键度和对抗度,$\pi_2(a_t\,|\,s_t)$为t时刻的第一重要性采样策略,$\pi_{base}(a_t\,|\,s_t)$为t时刻的第三重要性采样策略,$V(s_t)$为t时刻的状态价值函数,w_t为t时刻第三重要性采样的权重。

10.一种电子设备,其特征在于,包括:存储器和处理器;

所述存储器与所述处理器连接,用于存储程序;

所述处理器,用于通过运行所述存储器中的程序,实现如权利要求1-9任一项所述的自动驾驶未知危险场景的生成方法。

一种自动驾驶未知危险场景的生成方法及电子设备

技术领域

[0001] 本文涉及自动驾驶技术,尤指一种自动驾驶未知危险场景的生成方法及电子设备。

背景技术

[0002] 当前自动驾驶汽车在复杂驾驶环境下仍面临安全性问题,传统的安全评估方法难以全面覆盖潜在危险场景,导致事故率预测的准确性受到限制。因此,如何高效发现和评估自动驾驶系统可能遇到的危险场景,成为自动驾驶汽车大规模应用前的重要挑战。

[0003] 相关技术中,在基于深度强化学习的测试场景生成中,自动驾驶系统的安全验证可以通过生成高效率的关键测试场景得以优化。深度强化学习利用近端策略优化算法训练对抗度,从而动态调整重要性采样策略,从而提升关键场景的采样概率,同时保持策略稳定性。

[0004] 然而,现有的基于深度强化学习的场景生成方法只涉及对抗度的学习,在面对未知危险场景时,关键度往往趋近于零,无法有效反映场景的重要性。

发明内容

[0005] 本申请实施例提供了一种自动驾驶未知危险场景的生成方法及电子设备,能够基于深度强化学习的关键度与对抗度联合学习,从而有效反映场景的重要性。

[0006] 本申请实施例提供了一种自动驾驶未知危险场景的生成方法,所述方法包括:

获取第一重要性采样策略;其中,所述第一重要性采样策略为对第二重要性采样策略进行模仿学习,通过同时优化状态动作关键度和对抗度所得到的优化后的重要性采样策略;所述第二重要性采样策略为基于深度强化学习,并且对目标类型的场景进行对抗度学习得到的重要性采样策略;

利用获得的第一重要性采样策略并根据环境感知信息生成自动驾驶未知危险场景。

[0007] 本申请实施例还提供了一种电子设备,包括:存储器和处理器;

所述存储器与所述处理器连接,用于存储程序;

所述处理器,用于通过运行所述存储器中的程序,实现如上所述的自动驾驶未知危险场景的生成方法。

[0008] 本申请实施例包括获取第一重要性采样策略;其中,所述第一重要性采样策略为对第二重要性采样策略进行模仿学习,通过同时优化状态动作关键度和对抗度所得到的优化后的重要性采样策略;所述第二重要性采样策略为基于深度强化学习,并且对目标类型的场景进行对抗度学习得到的重要性采样策略;利用获得的第一重要性采样策略并根据环境感知信息生成自动驾驶未知危险场景。由于采用基于深度强化学习的关键度与对抗度联合学习方法,因此同时结合关键度与对抗度实现了采样策略的优化,提高了对目标类型场景的生成能力。

[0009] 本申请的其它特征和优点将在随后的说明书中阐述,并且,部分地从说明书中变得明显的,或者通过实施本申请而了解。本申请的其他优点可通过在说明书以及附图中所描述的方案来实现和获得。

附图说明

[0010] 附图用来提供对本申请技术方案的理解,并且构成说明书的一部分,与本申请的实施例一起用于解释本申请的技术方案,并不构成对本申请技术方案的限制。

[0011] 图1为本申请实施例一种自动驾驶未知危险场景的生成方法的流程示意图;

图2为相关技术中一种对抗度的学习过程示意图;

图3为本申请实施例一种对抗度和关键度的学习过程示意图;

图4为本申请实施例一种第一重要性采样策略的获取过程示意图;

图5为本申请实施例另一种第一重要性采样策略的获取过程示意图;

图6为本申请实施例一种自动驾驶未知危险场景的生成装置的结构示意图;

图7为本申请实施例一种电子设备的结构示意图。

具体实施方式

[0012] 本申请描述了多个实施例,但是该描述是示例性的,而不是限制性的,并且对于本领域的普通技术人员来说明显的是,在本申请所描述的实施例包含的范围内可以有更多的实施例和实现方案。尽管在附图中示出了许多可能的特征组合,并在具体实施方式中进行了讨论,但是所公开的特征的许多其它组合方式也是可能的。除非特意加以限制的情况以外,任何实施例的任何特征或元件可以与任何其它实施例中的任何其他特征或元件结合使用,或可以替代任何其它实施例中的任何其他特征或元件。

[0013] 本申请包括并设想了与本领域普通技术人员已知的特征和元件的组合。本申请已经公开的实施例、特征和元件也可以与任何常规特征或元件组合,以形成独特的发明方案。任何实施例的任何特征或元件也可以与来自其它发明方案的特征或元件组合,以形成另一个独特的发明方案。因此,应当理解,在本申请中示出和/或讨论的任何特征可以单独地或以任何适当的组合来实现。因此,除了根据所附权利要求及其等同替换所做的限制以外,实施例不受其它限制。此外,可以在所附权利要求的保护范围内进行各种修改和改变。

[0014] 此外,在描述具有代表性的实施例时,说明书可能已经将方法和/或过程呈现为特定的步骤序列。然而,在该方法或过程不依赖于本文所述步骤的特定顺序的程度上,该方法或过程不应限于所述的特定顺序的步骤。如本领域普通技术人员将理解的,其它的步骤顺序也是可能的。因此,说明书中阐述的步骤的特定顺序不应被解释为对权利要求的限制。此外,针对该方法和/或过程的权利要求不应限于按照所写顺序执行它们的步骤,本领域技术人员可以容易地理解,这些顺序可以变化,并且仍然保持在本申请实施例的精神和范围内。

[0015] 当前自动驾驶汽车在复杂驾驶环境下仍面临安全性问题,传统的安全评估方法难以全面覆盖潜在危险场景,导致事故率预测的准确性受到限制。因此,如何高效发现和评估自动驾驶系统可能遇到的危险场景,成为自动驾驶汽车大规模应用前的重要挑战。

[0016] 由于自然驾驶环境中的安全关键场景,尤其是未知危险场景极为罕见,基于蒙特卡洛采样的传统测试方法难以高效收集事故样本,同时,现有的强化学习方法难以有效生

成未知危险场景,影响自动驾驶安全性测试的可靠性。因此,如何有效预测并生成未知危险场景,成为当前自动驾驶安全测试领域的核心挑战。

[0017] 为此,本公开实施例提供一种自动驾驶未知危险场景的生成方法,如图1所示,包括:

步骤100、获取第一重要性采样策略;其中,第一重要性采样策略为对第二重要性采样策略进行模仿学习,通过同时优化状态动作关键度和对抗度所得到的优化后的重要性采样策略;第二重要性采样策略为基于深度强化学习,并且对目标类型的场景进行对抗度学习得到的重要性采样策略。

[0018] 步骤110、利用获得的第一重要性采样策略并根据环境感知信息生成自动驾驶未知危险场景。

[0019] 强化学习的框架通常包括五个基本要素:智能体、环境、状态、动作和奖励。智能体通过与环境的交互来学习最佳策略。它在每个时间步从环境中获取当前的状态,然后根据策略选择一个动作且执行,并在执行后,环境会根据智能体的动作反馈一个奖励,并更新状态。智能体的目标是通过试错和优化策略,最大化累积奖励。强化学习的核心挑战在于如何平衡探索(即尝试新动作)与利用(选择已有的最佳动作),并在动态的环境中持续改进策略。

[0020] 密集深度强化学习是强化学习的一种变体,它通常指在大规模、复杂环境中训练智能体时,使用密集的奖励信号来引导学习过程。在这种方法中,智能体可以通过频繁的反馈获得更多的信息,从而加速学习和优化决策策略。通过结合深度神经网络,密集深度强化学习能够处理更高维度的状态空间和复杂的环境任务。密集深度强化学习在自动驾驶领域已有取得显著成果的应用。

[0021] 在自动驾驶领域,记自动驾驶汽车的测试指标(例如:事故率)为$\mu = \mathrm{P}(F) = \mathrm{E}_p[I_F(X)]$,其中$F$表示兴趣事件(例如:车辆之间的碰撞事故),$p$表示测试场景$X$的自然概率分布,$I_F(X)$为事件$F$的指示函数,$I_F(X) = \begin{cases} 1, & X \in F, \\ 0, & X \notin F. \end{cases}$。蒙特卡洛采样是一种基于随机抽样的数值计算方法,其原理是通过随机生成大量样本点,然后通过统计分析这些样本的结果来近似求解问题。对于一个随机变量Y的期望值$E[Y]$,蒙特卡洛方法的估算公式是:

$$E[Y] = \frac{1}{N}\sum_{i=1}^{N} Y_i \tag{1}$$

其中,Y_i是从Y的概率分布中独立随机抽取的样本。基于蒙特卡洛采样原理,测试指标可以估计为:

$$\hat{\mu}_p = \frac{1}{n}\sum_{i=1}^{n} I_F(X_i), X_i \sim p \tag{2}$$

其中,$X_i, i = 1,\ldots,n$为独立同概率分布p的测试场景。自动驾驶汽车的事故事件发生概率极低,导致基于蒙特卡洛采样的测试方法需要非常多的测试次数才可以得到事故率的估计值,而大量测试所消耗的时间成本和经济成本是不可接受的。

[0022] 重要性采样方法可以缓解这一问题,其基本思想是提高那些容易发生事故的关键

场景的采样概率,从而提高测试效率。但是重要性采样方法面临维度灾难问题,无法应用于高维测试环境中。因此,稀疏对抗采样方法被提出,其只对测试场景中包含的关键变量进行重要性采样,而其他变量遵循原本的自然概率分布不变。基于稀疏对抗采样方法,测试指标可以估计为:

$$\hat{\mu}_q = \frac{1}{n}\sum_{i=1}^{n}\frac{I_F(X_i)p(X_i)}{q(X_i)}, X_i \sim q \quad (3)$$

其中,q为稀疏对抗采样方法构造的重要性采样函数,定义为:

$$q(x) := \rho(s_0)\prod_{t=0}^{T-1}\pi(a_t \mid s_t)P_\pi(s_{t+1} \mid s_t, a_t), \forall x \in X$$

$\pi(a \mid s)$为重要性采样后的更新策略,ρ是初始状态分布,P_π是与自动驾驶车辆策略π相关的状态转移概率。场景x可以定义为:

$$x := (s_0, a_0, \ldots, s_{T-1}, a_{T-1}, s_T) \in X$$

其中$s_t \in S$表示在时间t的状态,$a_t \in A$是在时间t的自动驾驶车辆动作,T是时间跨度,X是所有可行场景的集合。

[0023] 在基于密集深度强化学习的对抗度学习中,针对关键状态,重要性采样策略为:

$$\pi(a|s) = [1-\epsilon(s)]\pi_{base}(a|s) + \epsilon(s)\frac{Q(s,a)\pi_{base}(a|s)}{V(s)} \quad (4)$$

其中,a为动作,s为状态,$\pi_{base}(a \mid s)$为基础采样策略,$V(s) = \sum_a Q(s,a)\pi_{base}(a \mid s)$为状态价值函数,$\epsilon(s)$为对抗度,$Q(s,a)$为关键度。基于代理模型训练得到状态动作关键度$Q(s,a)$,采用密集深度强化学习对对抗度$\epsilon(s)$进行训练。具体而言,该方法使用基于密集深度强化学习的近端策略优化算法训练对抗度$\epsilon(s)$,其网络结构可以如图2所示:

在基于密集深度强化学习的测试场景生成中,自动驾驶系统的安全验证可以通过生成高效率的关键测试场景得以优化。密集深度强化学习利用近端策略优化算法训练对抗度$\epsilon(s)$,从而动态调整重要性采样策略$\pi(a \mid s)$,从而提升关键场景的采样概率,同时保持策略稳定性。

[0024] 然而,现有的基于密集深度强化学习的场景生成方法只涉及对抗度的学习,在面对未知危险场景时,关键度往往趋近于零,无法有效反映场景的重要性,仅依赖对抗度的学习难以有效指导场景挖掘与评估,而由于未考虑关键度$Q(s,a)$的非最优性,这可能会影响重要性采样策略$\pi(a \mid s)$的优化效果,从而降低加速测试的效率。

[0025] 针对此问题,本申请实施例自动驾驶未知危险场景的生成方法中提出基于密集深度强化学习的关键度与对抗度联合学习方法,同时优化$Q(s,a)$和$\pi(a \mid s)$,如图3所示。其核心思想是:(1)关键度学习:识别哪些状态-动作对 $Q(s,a)$最关键,能够触发事故;(2)对抗度学习:构造最具挑战性的测试场景,以提高测试的有效性。在基于密集深度强化学习的

对抗度学习方法中，对抗度$\epsilon(s)$度量一个状态s在测试中的对抗性。基于重要性采样策略，可以对测试场景的采样分布进行优化，以提高测试效率并减少灾难性失效。

[0026] 本申请实施例提供的自动驾驶未知危险场景的生成方法，获取第一重要性采样策略；其中，第一重要性采样策略为对第二重要性采样策略进行模仿学习，通过同时优化状态动作关键度和对抗度所得到的优化后的重要性采样策略；第二重要性采样策略为基于深度强化学习，并且对目标类型的场景进行对抗度学习得到的重要性采样策略；利用获得的第一重要性采样策略并根据环境感知信息生成自动驾驶未知危险场景。由于采用基于深度强化学习的关键度与对抗度联合学习方法，因此同时结合关键度与对抗度实现了采样策略的优化，优化了样本分布，显著提高了测试效率，降低了计算成本，提高了对目标类型场景的生成能力。

[0027] 现有的基于密集深度强化学习的场景生成方法只涉及对抗度的学习，然而，在面对未知危险场景时，关键度往往趋近于零，无法有效反映场景的重要性，仅依赖对抗度的学习难以有效指导场景挖掘与评估。而本申请实施例提供的自动驾驶未知危险场景的生成方法，实际提出了一种基于密集深度强化学习的未知危险场景生成方法，采用关键度与对抗度联合学习，提高了训练数据的密集性，进而优化了事故率估计。此外，结合防御型重要性采样，有效提升了自动驾驶测试的效率和准确性，为自动驾驶汽车的安全性评估提供了一种可行的加速测试方案。

[0028] 一种示例性实例中，如图4所示，第一重要性采样策略通过以下方式获得：

步骤200、对模仿学习策略进行初始化；其中，模仿学习策略用于模仿学习第二重要性采样策略，并且模仿学习策略和第二重要性采样策略均包括：关键度和对抗度；

步骤210、构建第一奖励函数；其中，第一奖励函数用于奖励模仿学习策略中的关键度逼近第二重要性采样策略中的关键度，且模仿学习策略中的对抗度逼近第二重要性采样策略中的对抗度；

步骤220、从状态空间中获取多个初始状态作为第一初始状态S_{at}，利用获得的多个第一初始状态和所构建的第一奖励函数更新模仿学习策略，得到第三重要性采样策略；

步骤230、根据获得的第三重要性采样策略得到第一重要性采样策略。

[0029] 在实际应用时，可以从状态空间中随机获取多个初始状态作为第一初始状态。

[0030] 一种示例性实例中，根据获得的第三重要性采样策略得到第一重要性采样策略，包括：

将获得的第三重要性采样策略作为第一重要性采样策略；

或者，

对获得的第三重要性采样策略再进行训练，并将经过训练得到的第三重要性采样策略作为第一重要性采样策略。

[0031] 根据获得的第三重要性采样策略得到第一重要性采样策略可以包括两种方法，一种是直接将获得的第三重要性采样策略作为第一重要性采样策略，另一种是对获得的第三重要性采样策略再进行训练以得到第一重要性采样策略。

[0032] 一种示例性实例中，利用获得的多个第一初始状态和所构建的第一奖励函数更新模仿学习策略，得到第三重要性采样策略，包括：

从获得的多个第一初始状态中任意选择一个第一初始状态，将经过初始化的模仿

学习策略作为所选择的第一初始状态下的模仿学习策略,并执行以下第一策略更新操作;

第一策略更新操作包括:

根据第一奖励函数不断对所选择的第一初始状态下的模仿学习策略进行迭代更新,以获得所选择的第一初始状态下最终更新的模仿学习策略;在最终更新的模仿学习策略收敛的情况下,将所选择的第一初始状态下获得的最终更新的模仿学习策略作为第三重要性采样策略,在最终更新的模仿学习策略未收敛的情况下,从获得的多个第一初始状态中另外选择一个第一初始状态,并将之前所选择的第一初始状态下获得的最终更新的模仿学习策略,作为新选择的第一初始状态下的模仿学习策略,继续执行第一策略更新操作,直到新选择的第一初始状态下获得的更新的模仿学习策略收敛,将新选择的第一初始状态下获得的最终更新的模仿学习策略作为第三重要性采样策略。

[0033] 获得的所选择的第一初始状态下最终更新的模仿学习策略是指利用所选择的第一初始状态在不同时刻下不断更新的模仿学习策略,所得到的最终的模仿学习策略。

[0034] 在利用所选择的第一初始状态获得最终更新的模仿学习策略收敛的情况下,不再利用其他第一初始状态更新模仿学习策略,而在利用所选择的第一初始状态获得最终更新的模仿学习策略未收敛的情况下,需要选择新的第一初始状态,并利用新的所选择的第一初始状态继续更新模仿学习策略。

[0035] 一种示例性实例中,根据第一奖励函数不断对所选择的第一初始状态下的模仿学习策略进行迭代更新,以获得所选择的第一初始状态下最终更新的模仿学习策略,包括:

将所选择的第一初始状态下的初始时刻作为当前第一初始状态下的当前时刻,将所选择的第一初始状态下的模仿学习策略作为所选择的第一初始状态下当前时刻的模仿学习策略,并执行以下第二策略更新操作;

第二策略更新操作包括:

根据所选择的第一初始状态下当前时刻的模仿学习策略从动作空间中采样动作并执行,以将所选择的第一初始状态从当前时刻转移至下一时刻,计算本次状态转移过程的第一奖励函数的奖励值;对获得的本次状态转移过程的第一奖励函数的奖励值进行优化,并基于优化后的第一奖励函数的奖励值反向更新当前时刻的模仿学习策略中的状态动作关键度和对抗度,得到更新的模仿学习策略;将下一时刻替代当前时刻作为新的当前时刻,将更新的模仿学习策略作为所选择的第一初始状态下新的当前时刻的模仿学习策略,继续执行第二策略更新操作,直到更新的模仿学习策略已收敛或当前时刻已是最大时刻。

[0036] 一种示例性实例中,第一奖励函数设计如下:

$$R_1(s_t) = -\| Q_1(s_t, a_t) - Q_{D2RL}(s, a) \|_1 - \| \epsilon_1(s_t) - \epsilon_{D2RL}(s) \|_1 ;$$

其中,$R(s_t)$为t时刻的第一奖励函数,$Q_{D2RL}(s,a)$和$\epsilon_{D2RL}(s)$分别为第二重要性采样策略中的状态动作关键度和对抗度,$Q_1(s_t,a_t)$和$\epsilon_1(s_t)$分别为t时刻的模仿学习策略中的状态动作关键度和对抗度。1代表1-范数。

[0037] 一种示例性实例中,对获得的第三重要性采样策略再进行训练,并将经过训练得到的第三重要性采样策略作为第一重要性采样策略,如图5所示,包括:

步骤300、构建第二奖励函数;其中,第二奖励函数用于奖励第三重要性采样策略快速完成收敛;

步骤310、从状态空间中获取多个初始状态作为第二初始状态S_{ct}；

步骤320、从获得的多个第二初始状态中任意选择一个第二初始状态，将第三重要性采样策略作为所选择的第二初始状态下的策略，并执行以下第三策略更新操作；

第三策略更新操作包括：

根据第二奖励函数不断对所选择的第二初始状态下的第三重要性采样策略进行迭代更新，以获得所选择的第二初始状态下最终更新的第三重要性采样策略；在最终更新的第三重要性采样策略收敛的情况下，将所选择的第一初始状态下获得的并且用第二初始状态S_{ct}更新后的第三重要性采样策略作为第一重要性采样策略，在最终更新的第三重要性采样策略未收敛的情况下，从获得的多个第二初始状态中另外选择一个第二初始状态，并将之前所选择的第二初始状态下获得的最终更新的第三重要性采样策略，作为新选择的第二初始状态下更新的第三重要性采样策略，继续执行第三策略更新操作，直到新选择的第二初始状态下获得的更新的第三重要性采样策略收敛，将新选择的第一初始状态下获得的最终更新的第三策略更新操作作为第一重要性采样策略。

[0038] 在实际应用时，可以从状态空间中随机获取多个初始状态作为第二初始状态。

[0039] 获得的所选择的第二初始状态下最终更新的模仿学习策略是指利用所选择的第二初始状态在不同时刻下不断更新的模仿学习策略，所得到的最终的模仿学习策略。

[0040] 在利用所选择的第二初始状态获得最终更新的模仿学习策略收敛的情况下，不再利用其他第二初始状态更新模仿学习策略，而在利用所选择的第二初始状态获得最终更新的模仿学习策略未收敛的情况下，需要选择新的第二初始状态，并利用新的所选择的第二初始状态继续更新模仿学习策略。

[0041] 一种示例性实例中，根据第二奖励函数不断对所选择的第二初始状态下的第三重要性采样策略进行迭代更新，以获得所选择的第二初始状态下最终更新的第三重要性采样策略，包括：

将所选择的第二初始状态下的初始时刻作为当前第二初始状态下的当前时刻，将所选择的第二初始状态下的第三重要性采样策略作为所选择的第二初始状态下当前时刻的第三重要性采样策略，并执行以下第四策略更新操作；

第四策略更新操作包括：

根据所选择的第二初始状态下当前时刻的第三重要性采样策略从动作空间采样动作并执行，以将所选择的第二初始状态从当前时刻转移至下一时刻，计算本次状态转移过程的第二奖励函数的奖励值；对获得的本次状态转移过程的第二奖励函数的奖励值进行优化，并基于优化后的第二奖励函数的奖励值反向更新当前时刻的第三重要性采样策略中的状态动作关键度和对抗度，得到更新的第三重要性采样策略；将下一时刻替代当前时刻作为新的当前时刻，并将更新的第三重要性采样策略作为所选择的第二初始状态下新的当前时刻的第三重要性采样策略，继续执行第四策略更新操作，直到更新的第三重要性采样策略已收敛或当前时刻已是最大时刻。

[0042] 一种示例性实例中，第二奖励函数设计如下：

$$R_2(s_t)=\begin{cases}max\{100-50000I_F(x)w(x),-100\},if\ t=T\text{或者碰撞}\\0,otherwise\end{cases};$$

$$w(x)=\prod_{t=0}^{T-1}w_t,\text{且}w_t=\frac{\pi_{base}(a_t|s_t)}{\pi_2(a_t|s_t)}=\frac{V(s_t)}{[1-\epsilon_2(s_t)]V(s_t)+\epsilon_2(s_t)Q_2(s_t,a_t)},t=0,\ldots,T;$$

$$V(s_t)=\sum_a Q_2(s_t,a_t)\pi_{base}(a_t\,|\,s_t),\forall s\in\text{状态空间}\mathcal{S},a\in\text{动作空间}\mathcal{A};$$

其中，$R_2(s_t)$为t时刻的第二奖励函数，$Q_2(s_t,a_t)$、$\epsilon_2(s_t)$分别为t时刻的第一重要性采样策略中的状态动作关键度和对抗度，$\pi_2(a_t\,|\,s_t)$为t时刻的第一重要性采样策略，$\pi_{base}(a_t\,|\,s_t)$为t时刻的第三重要性采样策略，$V(s_t)$为t时刻的状态价值函数，W_t为t时刻第三重要性采样的权重。

[0043] $Q_2(s_t,a_t)$、$\epsilon_2(s_t)$是$\pi_2(a_t\,|\,s_t)$这一神经网络的输出。

[0044] 本公开实施例提供的自动驾驶未知危险场景的生成方法所涉及的模型训练过程分为两个阶段：

阶段一：模仿学习(Imitation Learning, IL)

对基于密集深度强化学习的对抗度学习方法得到的重要性采样策略进行模仿学习。因此，设定奖励函数为：

$$R_1(s_t)=-\|\,Q_1(s_t,a_t)-Q_{D2RL}(s,a)\,\|_1-\|\,\epsilon_1(s_t)-\epsilon_{D2RL}(s)\,\|_1$$

其中，Q_{D2RL}和ϵ_{D2RL}为基于密集深度强化学习的对抗度学习方法中的状态动作关键度和对抗度。

[0045] 阶段二：重要性采样策略优化

对阶段一训练得到的重要性采样策略进行进一步训练，以提高加速测试的效果。因此，设定奖励函数为：

$$R_2(s_t)=\begin{cases}max\{100-50000I_F(x)w(x),-100\},if\ t=T\text{或者碰撞}\\0,otherwise\end{cases};$$

$$w(x)=\prod_{t=0}^{T-1}w_t,\text{且}w_t=\frac{\pi_{base}(a_t|s_t)}{\pi_2(a_t|s_t)}=\frac{V(s_t)}{[1-\epsilon_2(s_t)]V(s_t)+\epsilon_2(s_t)Q_2(s_t,a_t)},t=0,\ldots,T;$$

$$V(s_t)=\sum_a Q_2(s_t,a_t)\pi_{base}(a_t\,|\,s_t),\forall s\in\text{状态空间}\mathcal{S},\ a\in\text{动作空间}\mathcal{A};$$

其中，$R_2(s_t)$为t时刻的第二奖励函数，$Q_2(s_t,a_t)$、$\epsilon_2(s_t)$分别为t时刻的第一重要性采样策略中的状态动作关键度和对抗度，$\pi_2(a_t\,|\,s_t)$为t时刻的第一重要性采样策略，$\pi_{base}(a_t\,|\,s_t)$为t时刻的第三重要性采样策略，$V(s_t)$为t时刻的状态价值函数，W_t为t时刻第三重要性采样的权重。

[0046] 这样可以确保强化学习过程中对关键状态和对抗状态的有效训练，提高测试效率。

[0047] 本申请实施例提供的自动驾驶未知危险场景的生成方法，面向未知危险场景生成的基于密集深度强化学习的关键度与对抗度联合学习方法，使得测试数据中的关键信息更集中，避免相关方法中的数据稀疏问题。同时结合防御型重要性采样，优化样本分布，显著提高测试效率，降低计算成本。

[0048] 本申请实施例提供的自动驾驶未知危险场景的生成方法，面向未知危险场景生成

的基于密集深度强化学习的关键度与对抗度学习方法实现步骤如下：

1、获取驾驶环境的自然概率分布$\pi_{base}(a_t \mid s_t)$，$Q_{D2RL}(s,a)$(table形式)，记状态空间S，动作空间A，最长测试时间为T。

[0049] 2、收集训练数据：全部状态转移数组的集合S_{at}，以及关键状态的状态转移数组的集合S_{ct}

2.1、构造收集数据的行为策略(控制自动驾驶汽车的驾驶行为)：

$$\pi_b(a \mid s) = [1-\epsilon]\pi_{base}(a \mid s) + \epsilon\frac{Q_{D2RL}(s,a)\pi_{base}(a \mid s)}{V(s)}$$

其中，$\epsilon = 0.99$。

[0050] 2.2、设定奖励函数为：

$$r(s_t) = \begin{cases} max\{100 - 50000\mathbb{I}_F(x)w(x), -100\}, if\ t = T\ or\ 发生碰撞, \\ 0, otherwise, \end{cases}$$

其中$w(x) = \prod_{t=0}^{T-1} w_t$表示重要性采样的权重，且

$$w_t = \frac{\pi_{base}(a_t|s_t)}{\pi_b(a_t|s_t)} = \frac{V(s_t)}{[1-\epsilon]V(s_t) + \epsilon Q_{D2RL}(s_t,a_t)}, t = 0, \ldots, T.$$

2.3、for $i = 1,2,\ldots N$ （N为最大测试次数)：

2.3.1、采样初始状态s_0；

2.3.2、for $t = 0,1,2,\ldots T-1$(T为最长测试时间)：

a、根据行为策略$\pi_b(\cdot \mid s_t)$采样动作a_t；

b、执行动作a_t，更新下一步状态s_t+1，计算并记录状态的奖励$r_t = r(s_{t+1})$；

c、判断s_t是否是关键状态，若是，则将状态转移数组(s_t,a_t,r_t,s_{t+1})同时记录在S_{at}和S_{ct}，否则只记录在S_{at}中；

d、判断s_t+1是否是事故状态，若是，则跳出循环(break)；

2.3.3、最终得到全部状态转移数组的集合S_{at}，以及关键状态的状态转移数组的集合S_{ct}。

[0051] 3、基于密集深度强化学习训练$\epsilon_\theta(s)$：

随机初始化对抗度神经网络$\epsilon_\theta(s)$，设定最大训练轮数 $\text{epoch} = N$；每个批次(batch)设定批数据大小B，将S_{at}中的数据每B个划分为一个batch，最后一个batch 可以不足B个。

[0052] for epoch=1, 2…, N:

for batch(每个batch中有B个状态转移数组) in S_{ct}：

3.1、根据每一步的奖励r_t，计算优势函数$\hat{A}_t$；

3.2、计算近端策略优化算法(Proximal Policy Optimization,PPO)的损失：

$$L^{PPO}(\theta,\phi) = -L^{CLIP}(\theta) + c_1 L^{VF}(\phi) - c_2 L^{S}(\theta)$$

其中，$L^{CLIP}(\theta) = E_t\left[min\left(r_t(\theta)\hat{A}_t, clip(r_t(\theta),1-\varepsilon,1+\varepsilon)\hat{A}_t \right)\right]$是策略损失，$L^{VF}(\phi) = E_t\left[\left(V_\phi(s_t) - \hat{R}_t\right)^2\right]$是值函数损失，$L^{S}(\theta) = E_t[H[\pi_\theta(\cdot \mid s_t)]]$是熵奖励，$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$是新旧策略的概率比，$\varepsilon$是指概率比值的裁剪范围(即clip范围)，即通过参数ε限制新旧策略概率比值的上下限，防止新策略更新步长过大而造成策略性能的不稳定，$V_\phi(s_t)$是值函数估计，$\hat{R}_t = \sum_{l=0}^{\infty}\gamma^l r_{t+l}$是估算的回报，$c_1$、$c_2$是超参数；

3.3、通过梯度下降方法优化策略参数$\theta_{old} \leftarrow \theta$。

[0053] 4、输出$\epsilon_{D2RL}(s) = \epsilon_\theta(s)$。

[0054] 5、对训练后的重要性采样策略($Q_{D2RL}(s,a)$、$\epsilon_{D2RL}(s)$)进行模仿学习：

随机初始化模仿学习策略神经网络π_θ^{imit}($Q(s,a)$、$\epsilon(s)$均为神经网络的输出)，设定模仿学习的奖励函数为(其中1代表1-范数)：

$$R(s_t) = -\| Q(s_t,a_t) - Q_{D2RL}(s_t,a_t) \|_1 - \| \epsilon(s_t) - \epsilon_{D2RL}(s_t) \|_1$$

设定批数据大小B，将S_{at}中的数据每B个划分为一个batch，最后一个batch 可以不足B个；

while$R(s_t)$未收敛(收敛：足够趋于0)：

for batch(每个batch中有B个状态转移数组) in S_{at}：

5.1、观察策略网络π_θ^{imit}的输出$Q(s_t,a_t)$、$\epsilon(s_t)$；

5.2、计算相应的$R(s_t)$；

5.3、根据每一步的奖励$R(s_t)$，计算优势函数$\hat{A}_t$；

5.4、计算近端策略优化算法(Proximal Policy Optimization，PPO)的损失(同上第3步)；

5.5、判断策略网络奖励函数$R(s_t)$是否已收敛；

5.6、通过梯度下降方法优化策略参数$\theta_{old} \leftarrow \theta$；

5.7、输出模仿学习得到的策略网络π_θ^{imit}。

[0055] 6、对训练得到的重要性采样策略π_{imit}进行进一步训练：

设定批数据大小B，将S_{ct}中的数据每B个划分为一个batch，最后一个batch可以不足B个。

[0056] for epoch=1,2…, N：

for batch(每个batch中有B个状态转移数组) in S_{ct}：

6.1根据每一步的奖励r_t，计算优势函数$\hat{A}_t$；

6.2、计算近端策略优化算法(PPO)的损失(同上第3步)；

6.3、通过梯度下降方法优化策略参数$\theta_{old}\leftarrow\theta$。

[0057] 7、输出最终的策略网络π_{θ}^{imit}。

[0058] 本申请实施例提出面向未知危险场景生成的基于密集深度强化学习的关键度与对抗度联合学习方法,提高对未知危险场景的生成能力,减少对未知危险场景的低估风险,优化采样策略,提升测试效率,降低稀疏度灾难性带来的影响。

[0059] 本申请实施例提出了一种基于密集深度强化学习的关键度与对抗度联合学习方法,用于自动驾驶安全性测试领域未知危险场景的生成。该方法通过优化重要性采样策略,降低测试中的稀疏度灾难性的影响,以提高测试效率并减少测试成本,使自动驾驶安全性评估更加高效、稳健和精准。

[0060] 与上述的自动驾驶未知危险场景的生成方法对应的,本申请实施例还提供了一种自动驾驶未知危险场景的生成装置。图6是本申请实施例提供的一种自动驾驶未知危险场景的生成装置的结构示意图。如图6所示,本申请实施例提供的自动驾驶未知危险场景的生成装置包括:

策略获取单元400,用于获取第一重要性采样策略;其中,第一重要性采样策略为对第二重要性采样策略进行模仿学习,通过同时优化状态动作关键度和对抗度所得到的优化后的重要性采样策略;第二重要性采样策略为基于深度强化学习,并且对目标类型的场景进行对抗度学习得到的重要性采样策略;

场景生成单元410,用于利用获得的第一重要性采样策略并根据环境感知信息生成自动驾驶未知危险场景。

[0061] 在一种示例性实例中,本申请实施例提供的自动驾驶未知危险场景的生成装置还包括:策略生成单元420,策略生成单元420用于通过以下方式获得第一重要性采样策略:

对模仿学习策略进行初始化;其中,模仿学习策略用于模仿学习第二重要性采样策略,并且模仿学习策略和第二重要性采样策略均包括:关键度和对抗度;

构建第一奖励函数;其中,第一奖励函数用于奖励模仿学习策略中的关键度逼近第二重要性采样策略中的关键度,且模仿学习策略中的对抗度逼近第二重要性采样策略中的对抗度;

从状态空间中获取多个初始状态作为第一初始状态S_{at},利用获得的多个第一初始状态和所构建的第一奖励函数更新模仿学习策略,得到第三重要性采样策略;

根据获得的第三重要性采样策略得到第一重要性采样策略。

[0062] 在一种示例性实例中,策略生成单元420用于:

将获得的第三重要性采样策略作为第一重要性采样策略;

或者,

对获得的第三重要性采样策略再进行训练,并将经过训练得到的第三重要性采样策略作为第一重要性采样策略。

[0063] 在一种示例性实例中,策略生成单元420用于:

从获得的多个第一初始状态中任意选择一个第一初始状态,将经过初始化的模仿学习策略作为所选择的第一初始状态下的模仿学习策略,并执行以下第一策略更新操作;

第一策略更新操作包括:

根据第一奖励函数不断对所选择的第一初始状态下的模仿学习策略进行迭代更

新,以获得所选择的第一初始状态下最终更新的模仿学习策略;在最终更新的模仿学习策略收敛的情况下,将所选择的第一初始状态下获得的最终更新的模仿学习策略作为第三重要性采样策略,在最终更新的模仿学习策略未收敛的情况下,从获得的多个第一初始状态中另外选择一个第一初始状态,并将之前所选择的第一初始状态下获得的最终更新的模仿学习策略,作为新选择的第一初始状态下的模仿学习策略,继续执行第一策略更新操作,直到新选择的第一初始状态下获得的更新的模仿学习策略收敛,将新选择的第一初始状态下获得的最终更新的模仿学习策略作为第三重要性采样策略。

[0064] 在一种示例性实例中,策略生成单元420用于:

将所选择的第一初始状态下的初始时刻作为当前第一初始状态下的当前时刻,将所选择的第一初始状态下的模仿学习策略作为所选择的第一初始状态下当前时刻的模仿学习策略,并执行以下第二策略更新操作;

第二策略更新操作包括:

根据所选择的第一初始状态下当前时刻的模仿学习策略从动作空间中采样动作并执行,以将所选择的第一初始状态从当前时刻转移至下一时刻,计算本次状态转移过程的第一奖励函数的奖励值;对获得的本次状态转移过程的第一奖励函数的奖励值进行优化,并基于优化后的第一奖励函数的奖励值反向更新当前时刻的模仿学习策略中的状态动作关键度和对抗度,得到更新的模仿学习策略;将下一时刻替代当前时刻作为新的当前时刻,将更新的模仿学习策略作为所选择的第一初始状态下新的当前时刻的模仿学习策略,继续执行第二策略更新操作,直到更新的模仿学习策略已收敛或当前时刻已是最大时刻。

[0065] 在一种示例性实例中,第一奖励函数设计如下:

$$R_1(s_t) = -\| Q_1(s_t, a_t) - Q_{D2RL}(s, a) \|_1 - \| \epsilon_1(s_t) - \epsilon_{D2RL}(s) \|_1;$$

其中,$R(s_t)$为t时刻的第一奖励函数,$Q_{D2RL}(s,a)$和$\epsilon_{D2RL}(s)$分别为第二重要性采样策略中的状态动作关键度和对抗度,$Q_1(s_t,a_t)$和$\epsilon_1(s_t)$分别为t时刻的模仿学习策略中的状态动作关键度和对抗度。

[0066] 在一种示例性实例中,策略生成单元420用于:

构建第二奖励函数;其中,第二奖励函数用于奖励第三重要性采样策略快速完成收敛;

从状态空间中获取多个初始状态作为第二初始状态S_{ct};

从获得的多个第二初始状态中任意选择一个第二初始状态,将第三重要性采样策略作为所选择的第二初始状态下的策略,并执行以下第三策略更新操作;

第三策略更新操作包括:

根据第二奖励函数不断对所选择的第二初始状态下的第三重要性采样策略进行迭代更新,以获得所选择的第二初始状态下最终更新的第三重要性采样策略;在最终更新的第三重要性采样策略收敛的情况下,将所选择的第一初始状态下获得的并且用第二初始状态S_{ct}更新后的第三重要性采样策略作为第一重要性采样策略,在最终更新的第三重要性采样策略未收敛的情况下,从获得的多个第二初始状态中另外选择一个第二初始状态,并将之前所选择的第二初始状态下获得的最终更新的第三重要性采样策略,作为新选择的第二初始状态下更新的第三重要性采样策略,继续执行第三策略更新操作,直到新选择的

第二初始状态下获得的更新的第三重要性采样策略收敛,将新选择的第一初始状态下获得的最终更新的第三策略更新操作作为第一重要性采样策略。

[0067] 在一种示例性实例中,策略生成单元420用于:

将所选择的第二初始状态下的初始时刻作为当前第二初始状态下的当前时刻,将所选择的第二初始状态下的第三重要性采样策略作为所选择的第二初始状态下当前时刻的第三重要性采样策略,并执行以下第四策略更新操作;

第四策略更新操作包括:

根据所选择的第二初始状态下当前时刻的第三重要性采样策略从动作空间采样动作并执行,以将所选择的第二初始状态从当前时刻转移至下一时刻,计算本次状态转移过程的第二奖励函数的奖励值;对获得的本次状态转移过程的第二奖励函数的奖励值进行优化,并基于优化后的第二奖励函数的奖励值反向更新当前时刻的第三重要性采样策略中的状态动作关键度和对抗度,得到更新的第三重要性采样策略;将下一时刻替代当前时刻作为新的当前时刻,并将更新的第三重要性采样策略作为所选择的第二初始状态下新的当前时刻的第三重要性采样策略,继续执行第四策略更新操作,直到更新的第三重要性采样策略已收敛或当前时刻已是最大时刻。

[0068] 在一种示例性实例中,第二奖励函数设计如下:

$$R_2(s_t)=\begin{cases}max\{100-50000I_F(x)w(x),-100\},if\ t=T\\0,otherwise\end{cases};$$

$$w(x)=\prod_{t=0}^{T-1}w_t,\text{且}\ w_t=\frac{\pi_{base}(a_t|s_t)}{\pi_2(a_t|s_t)}=\frac{V(s_t)}{[1-\epsilon_2(s_t)]V(s_t)+\epsilon_2(s_t)Q_2(s_t,a_t)},t=0,\dots,T;$$

$$V(s_t)=\sum_a Q_2(s_t,a_t)\pi_{base}(a_t\,|\,s_t),\forall s\in\text{状态空间}\mathcal{S},a\in\text{动作空间}\mathcal{A};$$

其中,$R_2(s_t)$为t时刻的第二奖励函数,$Q_2(s_t,a_t)$、$\epsilon_2(s_t)$分别为t时刻的第一重要性采样策略中的状态动作关键度和对抗度,$\pi_2(a_t\,|\,s_t)$为t时刻的第一重要性采样策略,$\pi_{base}(a_t\,|\,s_t)$为t时刻的第三重要性采样策略,$V(s_t)$为t时刻的状态价值函数。

[0069] 本实施例提供的自动驾驶未知危险场景的生成装置与本申请上述实施例所提供的自动驾驶未知危险场景的生成方法属于同一申请构思,可执行本申请上述任意实施例所提供的自动驾驶未知危险场景的生成方法,具备执行自动驾驶未知危险场景的生成方法相应的功能模块和有益效果。未在本实施例中详尽描述的技术细节,可参见本申请上述实施例提供的自动驾驶未知危险场景的生成方法的具体处理内容,此处不再加以赘述。

[0070] 本申请实施例还提供一种电子设备,如图7所示,包括:存储器500和处理器510;

所述存储器500与所述处理器510连接,用于存储程序;

所述处理器510用于通过运行所述存储器500中的程序,实现上述任一实施例描述的自动驾驶未知危险场景的生成方法。

[0071] 具体的,上述电子设备还可以包括:总线、通信接口520、输入设备530和输出设备540。

[0072] 处理器510、存储器500、通信接口520、输入设备530和输出设备540通过总线相互连接。其中:

总线可包括一通路,在计算机系统各个部件之间传送信息。

[0073] 处理器510可以是通用处理器,例如通用中央处理器(CPU)、微处理器等,也可以是特定应用集成电路(application-specific integrated circuit,ASIC),或一个或多个用于控制本发明方案程序执行的集成电路。还可以是数字信号处理器(DSP)、专用集成电路(ASIC)、现成可编程门阵列(FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件。

[0074] 处理器510可包括主处理器,还可包括基带芯片、调制解调器等。

[0075] 存储器500中保存有执行本发明技术方案的程序,还可以保存有操作系统和其他关键业务。具体地,程序可以包括程序代码,程序代码包括计算机操作指令。更具体的,存储器500可以包括只读存储器(read-only memory,ROM)、可存储静态信息和指令的其他类型的静态存储设备、随机存取存储器(random access memory,RAM)、可存储信息和指令的其他类型的动态存储设备、磁盘存储器、flash等等。

[0076] 输入设备530可包括接收用户输入的数据和信息的装置,例如键盘、鼠标、摄像头、扫描仪、光笔、语音输入装置、触摸屏、计步器或重力感应器等。

[0077] 输出设备540可包括允许输出信息给用户的装置,例如展示屏、打印机、扬声器等。

[0078] 通信接口520可包括使用任何收发器一类的装置,以便与其他设备或通信网络通信,如以太网,无线接入网(RAN),无线局域网(WLAN)等。

[0079] 处理器510执行存储器500中所存放的程序,以及调用其他设备,可用于实现本申请上述实施例所提供的任意一种自动驾驶未知危险场景的生成方法的各个步骤。

[0080] 除了上述方法和设备以外,本申请的实施例还可以是计算机程序产品,其包括计算机程序指令,所述计算机程序指令在被处理器运行时使得所述处理器执行本说明书上述任意实施例中描述的根据本申请各种实施例的自动驾驶未知危险场景的生成方法中的步骤。

[0081] 所述计算机程序产品可以以一种或多种程序设计语言的任意组合来编写用于执行本申请实施例操作的程序代码,所述程序设计语言包括面向对象的程序设计语言,诸如Java、C++等,还包括常规的过程式程序设计语言,诸如“C”语言或类似的程序设计语言。程序代码可以完全地在用户计算设备上执行、部分地在用户设备上执行、作为一个独立的软件包执行、部分在用户计算设备上部分在远程计算设备上执行、或者完全在远程计算设备或服务器上执行。

[0082] 此外,本申请实施例还提供一种存储介质,所述存储介质上存储有计算机程序,所述计算机程序被处理器运行时,实现上述任一实施例描述的自动驾驶未知危险场景的生成方法。

[0083] 本领域普通技术人员可以理解,上文中所公开方法中的全部或某些步骤、系统、装置中的功能模块/单元可以被实施为软件、固件、硬件及其适当的组合。在硬件实施方式中,在以上描述中提及的功能模块/单元之间的划分不一定对应于物理组件的划分;例如,一个物理组件可以具有多个功能,或者一个功能或步骤可以由若干物理组件合作执行。某些组件或所有组件可以被实施为由处理器,如数字信号处理器或微处理器执行的软件,或者被实施为硬件,或者被实施为集成电路,如专用集成电路。这样的软件可以分布在计算机可读介质上,计算机可读介质可以包括计算机存储介质(或非暂时性介质)和通信介质(或暂时性介质)。如本领域普通技术人员公知的,术语“计算机存储介质”包括在用于存储信息(诸

如计算机可读指令、数据结构、程序模块或其他数据)的任何方法或技术中实施的易失性和非易失性、可移除和不可移除介质。计算机存储介质包括但不限于RAM、ROM、EEPROM、闪存或其他存储器技术、CD-ROM、数字多功能盘(DVD)或其他光盘存储、磁盒、磁带、磁盘存储或其他磁存储装置、或者可以用于存储期望的信息并且可以被计算机访问的任何其他的介质。此外,本领域普通技术人员公知的是,通信介质通常包含计算机可读指令、数据结构、程序模块或者诸如载波或其他传输机制之类的调制数据信号中的其他数据,并且可包括任何信息递送介质。

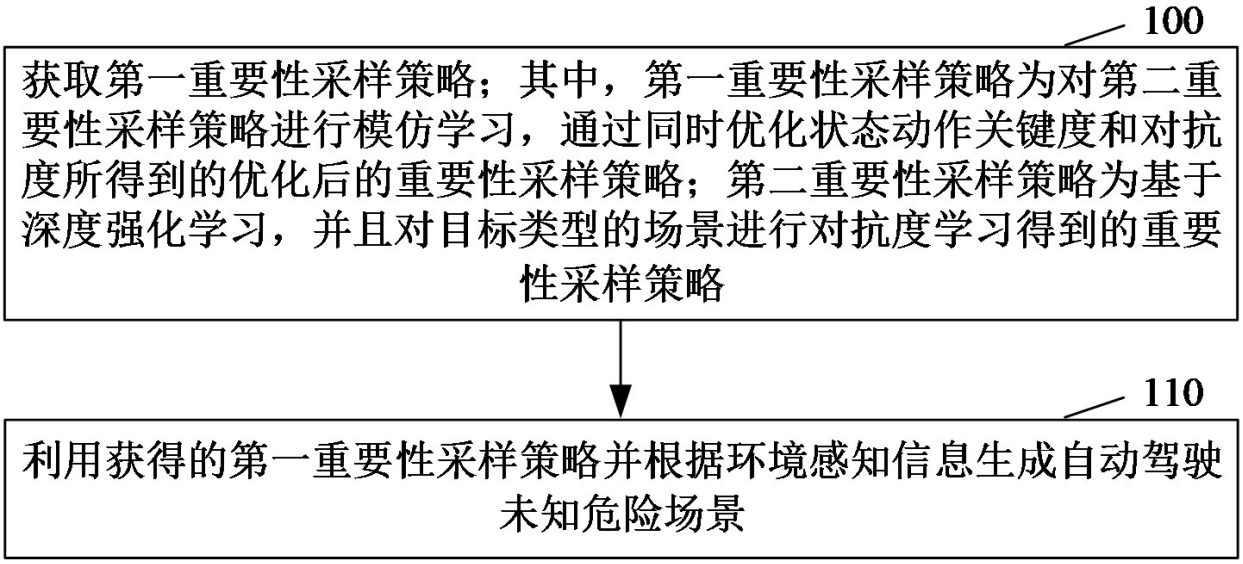

100
获取第一重要性采样策略；其中，第一重要性采样策略为对第二重要性采样策略进行模仿学习，通过同时优化状态动作关键度和对抗度所得到的优化后的重要性采样策略；第二重要性采样策略为基于深度强化学习，并且对目标类型的场景进行对抗度学习得到的重要性采样策略
110
利用获得的第一重要性采样策略并根据环境感知信息生成自动驾驶未知危险场景

图1

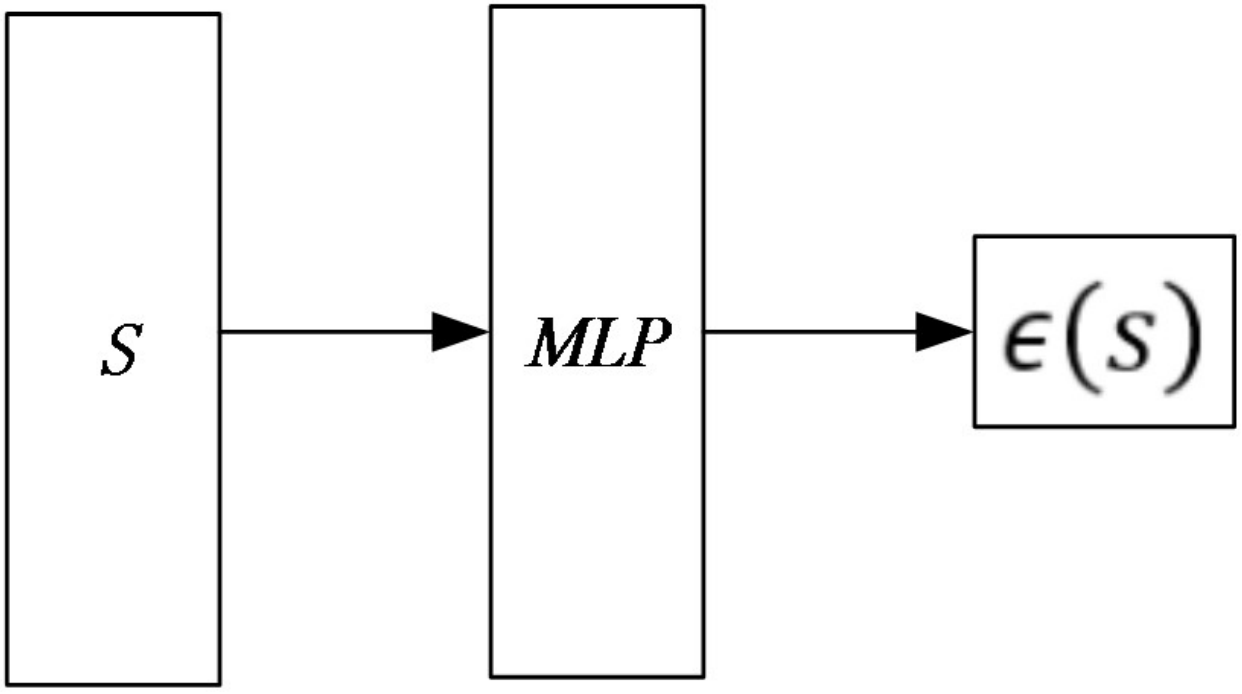

S
MLP
ϵ(S)

图2

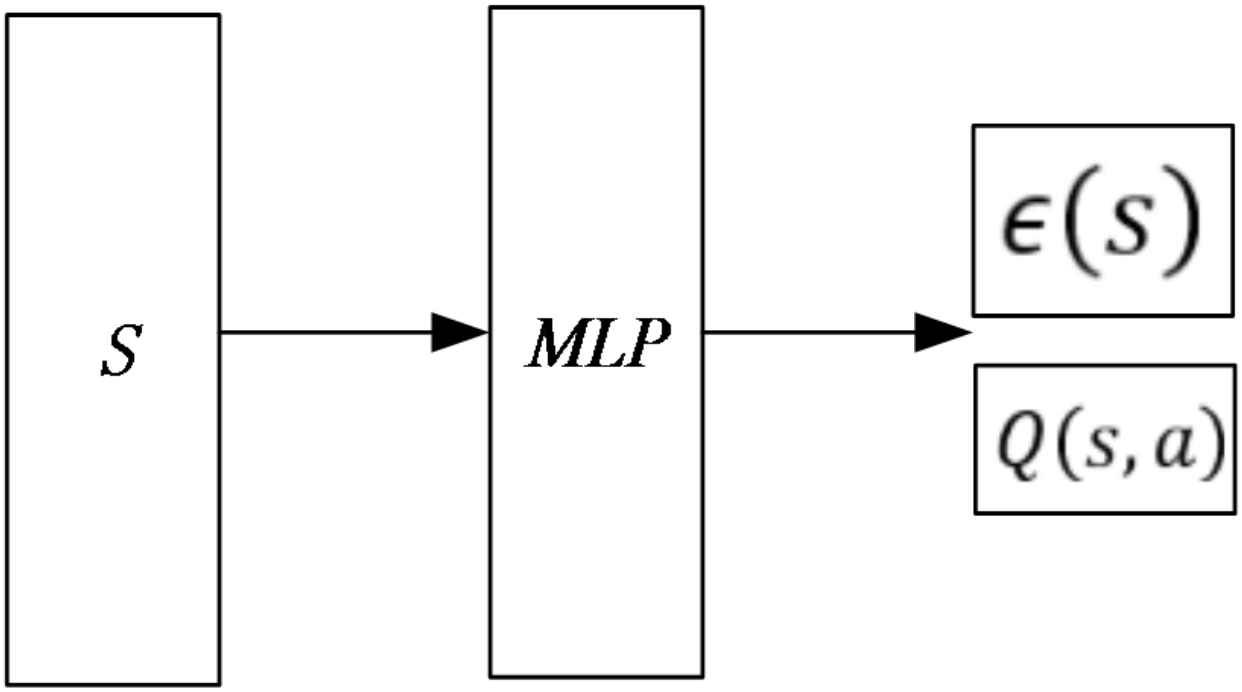


图3

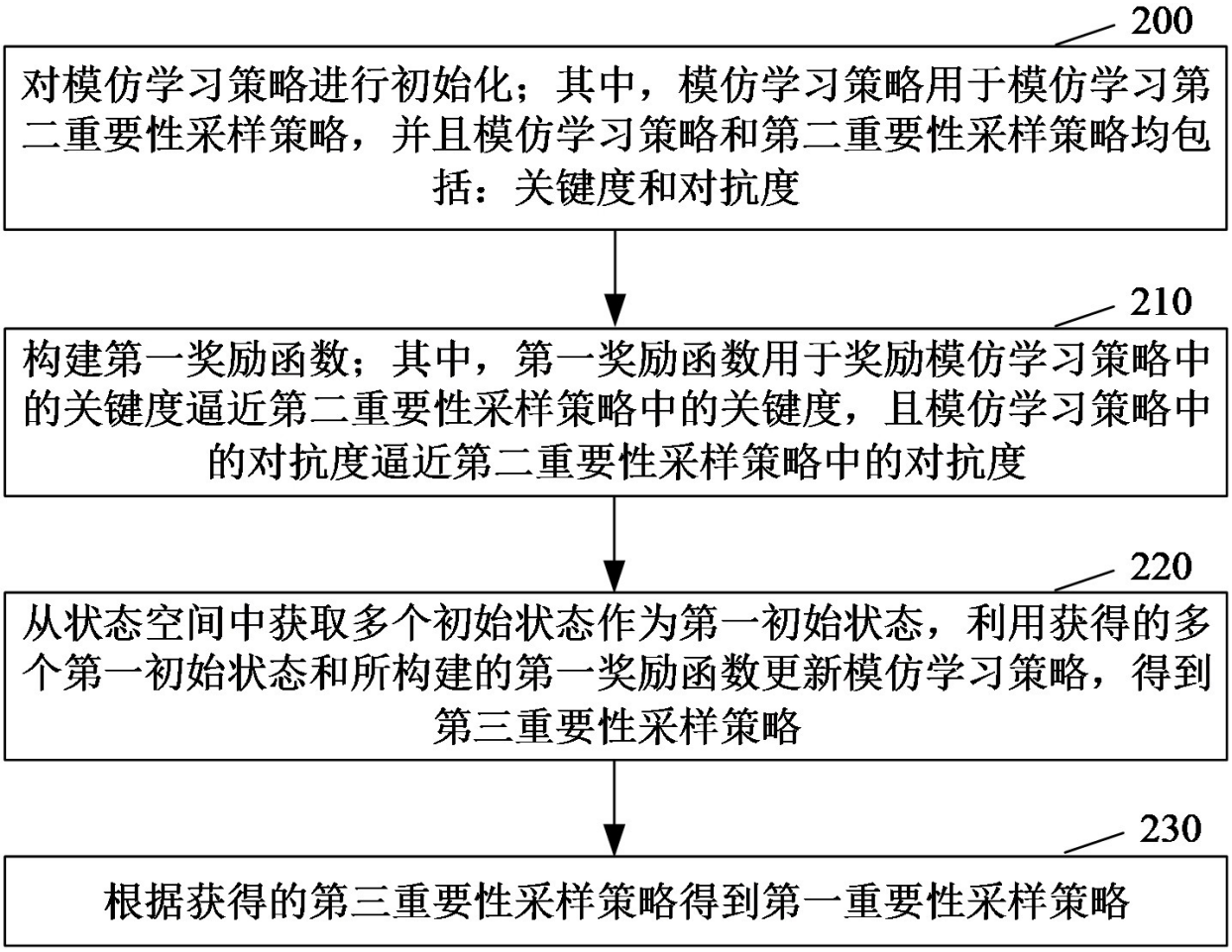

200
对模仿学习策略进行初始化；其中，模仿学习策略用于模仿学习第二重要性采样策略，并且模仿学习策略和第二重要性采样策略均包括：关键度和对抗度
210
构建第一奖励函数；其中，第一奖励函数用于奖励模仿学习策略中的关键度逼近第二重要性采样策略中的关键度，且模仿学习策略中的对抗度逼近第二重要性采样策略中的对抗度
220
从状态空间中获取多个初始状态作为第一初始状态，利用获得的多个第一初始状态和所构建的第一奖励函数更新模仿学习策略，得到第三重要性采样策略
230
根据获得的第三重要性采样策略得到第一重要性采样策略

图4

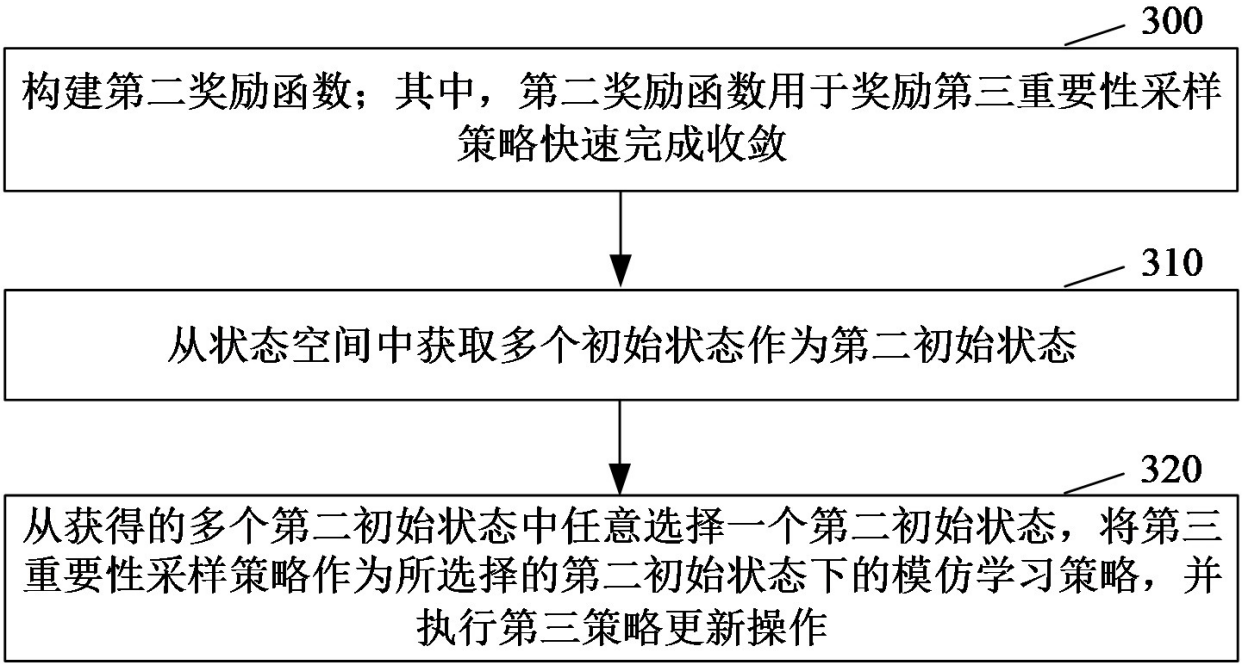

300
构建第二奖励函数；其中，第二奖励函数用于奖励第三重要性采样策略快速完成收敛
310
从状态空间中获取多个初始状态作为第二初始状态
320
从获得的多个第二初始状态中任意选择一个第二初始状态，将第三重要性采样策略作为所选择的第二初始状态下的模仿学习策略，并执行第三策略更新操作

图5

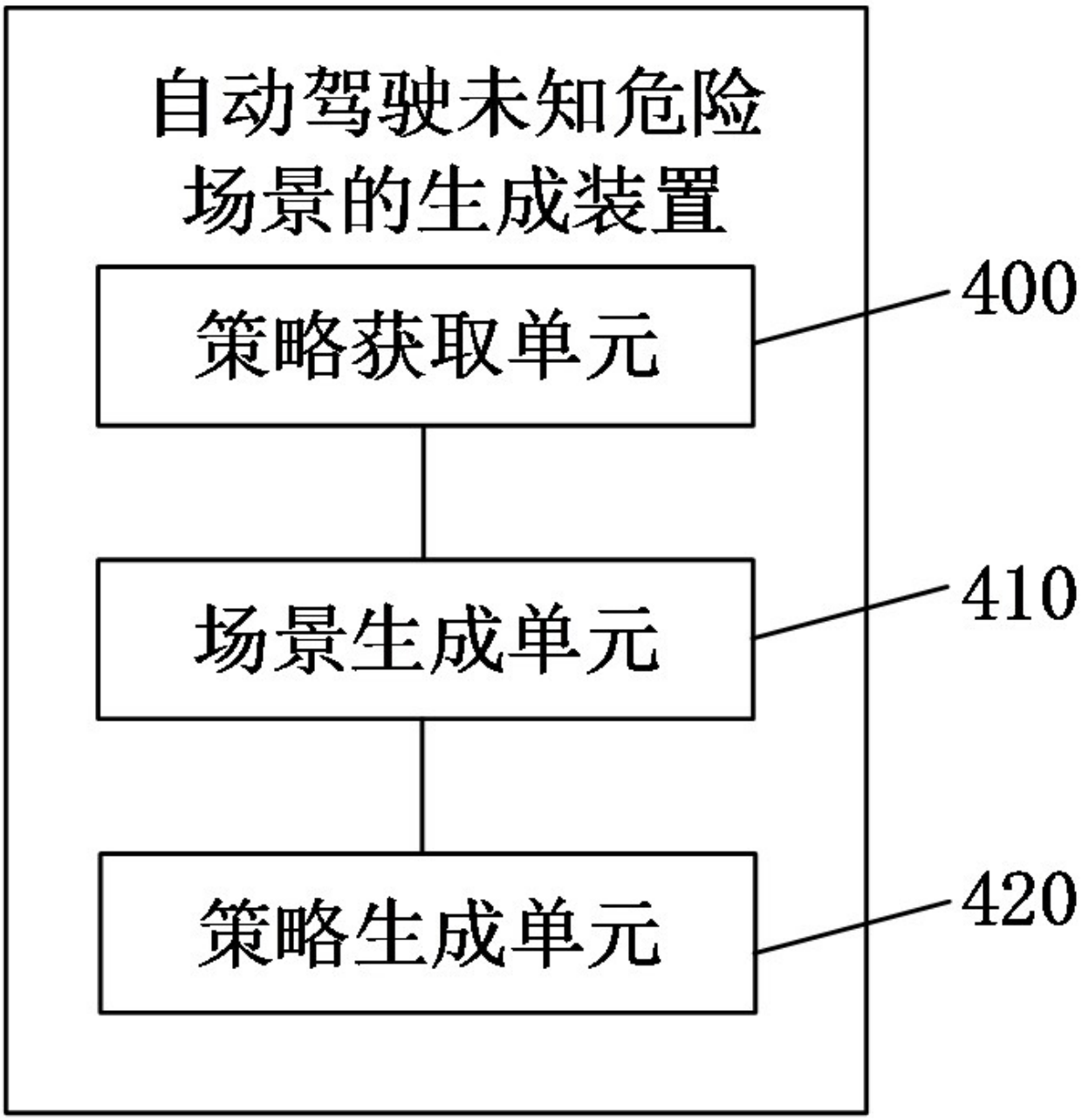


图6

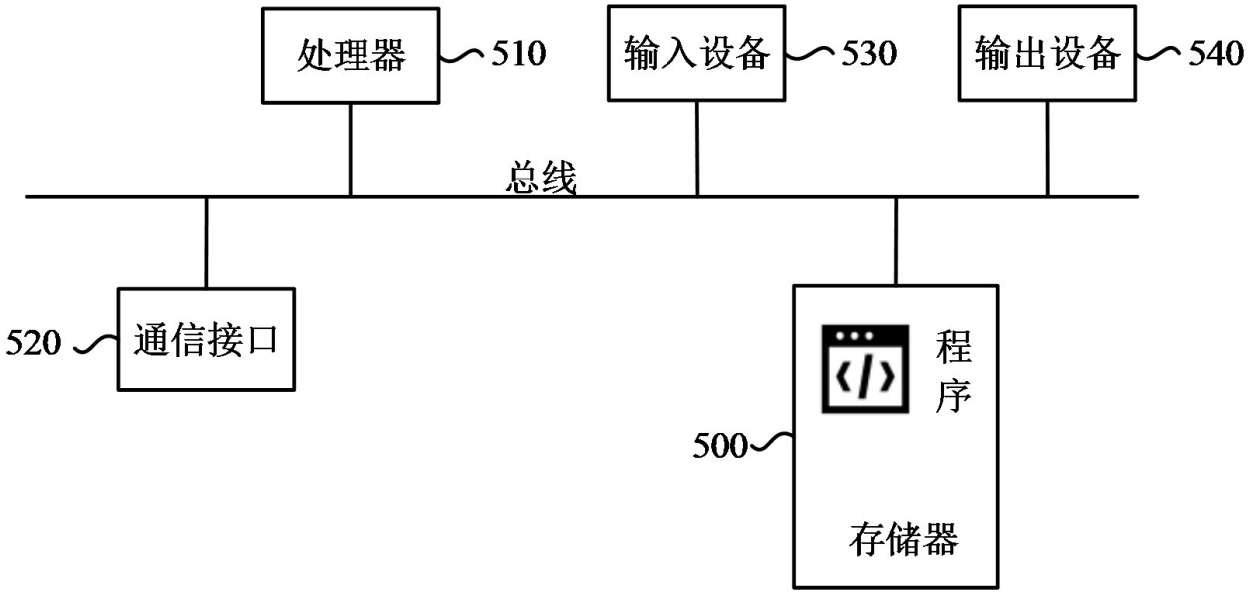


图7