(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 114404976 B
(45)授权公告日 2024.08.20

(21)申请号 202210067453.0

(22)申请日 2022.01.20

(65)同一申请的已公布的文献号
申请公布号 CN 114404976 A

(43)申请公布日 2022.04.29

(73)专利权人 腾讯科技(深圳)有限公司
地址 518057 广东省深圳市南山区高新区科技中一路腾讯大厦35层

(72)发明人 刘若尘 曹琪扬 廖詩飚 张良鹏 曾政文

(74)专利代理机构 北京三高永信知识产权代理有限责任公司 11138
专利代理师 李文静

(51)Int.Cl.
A63F 13/55(2014.01)
A63F 13/67(2014.01)
A63F 13/833(2014.01)
G06N 20/00(2019.01)
G06N 5/01(2023.01)

(56)对比文件
CN 109091868 A,2018.12.28

审查员 张鹏

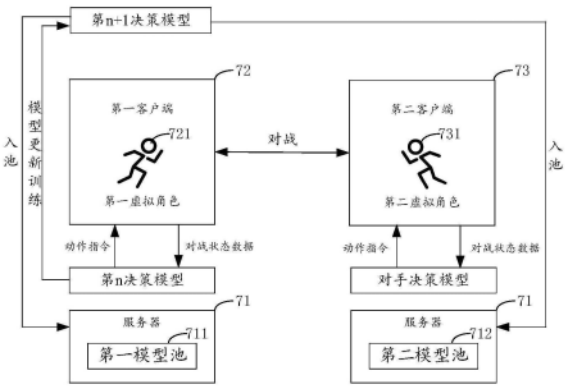
权利要求书3页 说明书21页 附图9页

(54)发明名称

决策模型的训练方法、装置、计算机设备及存储介质

(57)摘要

本申请实施例公开了一种决策模型的训练方法、装置、计算机设备及存储介质，属于人工智能领域。该方法包括：获取第一模型池，第一模型池中包含至少一个决策模型，决策模型用于指示虚拟角色在对战时采用的对战策略；在第n轮迭代过程中，将第i对战状态数据输入第n决策模型，得到第n决策模型输出的动作指令，动作指令用于指示第一客户端控制第一虚拟角色执行动作，并反馈第一虚拟角色执行动作后的收集到的第i+1对战状态数据；基于对战过程中的对战状态数据训练第n决策模型，得到第n+1决策模型；将第n+1决策模型添加至第一模型池。通过对战过程中的对战状态数据训练决策模型，提高决策模型的对战拟人性。



CN 114404976 B

1.一种决策模型的训练方法,其特征在于,所述方法包括:

获取第一模型池,所述第一模型池中包含至少一个决策模型,所述决策模型用于指示虚拟角色在对战时采用的对战策略;

在第n轮迭代过程中,将第i对战状态数据输入第n决策模型,得到所述第n决策模型输出的动作指令,所述第n决策模型为上一轮迭代过程训练得到的模型,所述第i对战状态数据用于表征对战过程中第i对战时刻时的对战状态,所述动作指令用于指示第一客户端控制第一虚拟角色执行动作,并反馈所述第一虚拟角色执行动作后的收集到的第i+1对战状态数据,所述动作指令包括移动指令和技能使用指令中的至少一种,所述第i+1对战状态数据包括所述第一虚拟角色执行动作后所述第一虚拟角色以及敌对虚拟角色的状态数据,所述状态数据包括位置状态数据、属性状态数据和技能状态数据中的至少一种;

基于对战过程中的对战状态数据训练所述第n决策模型,得到第n+1决策模型;

将所述第n+1决策模型添加至所述第一模型池。

2.根据权利要求1所述的方法,其特征在于,所述基于对战过程中的对战状态数据训练所述第n决策模型,得到第n+1决策模型,包括:

基于相邻对战时刻对应的所述对战状态数据,确定第一奖励值;

基于对战结果确定第二奖励值,所述对战结果用于指示对战的胜负结果;

基于所述第一奖励值和所述第二奖励值训练所述第n决策模型,得到所述第n+1决策模型。

3.根据权利要求2所述的方法,其特征在于,所述基于相邻对战时刻对应的所述对战状态数据,确定第一奖励值,包括:

获取所述第i对战状态数据中的第i角色属性值,以及所述第i+1对战状态数据中的第i+1角色属性值,其中,角色属性值包括所述第一虚拟角色以及与第二虚拟角色的属性值,所述第二虚拟角色是与所述第一虚拟角色对战的虚拟角色;

基于所述第i角色属性值与所述第i+1角色属性值的属性值变化情况,确定所述第一奖励值。

4.根据权利要求2所述的方法,其特征在于,所述基于所述第一奖励值和所述第二奖励值训练所述第n决策模型,得到所述第n+1决策模型,包括:

在迭代轮数达到轮数阈值的情况下,基于所述第一奖励值和所述第二奖励值训练所述第n决策模型,得到所述第n+1决策模型。

5.根据权利要求4所述的方法,其特征在于,所述基于对战过程中的对战状态数据训练所述第n决策模型,得到第n+1决策模型,还包括:

在所述迭代轮数未达到所述轮数阈值的情况下,基于所述对战状态数据以及所述动作指令,确定第三奖励值;

基于所述第一奖励值、所述第二奖励值以及所述第三奖励值训练所述第n决策模型,得到所述第n+1决策模型。

6.根据权利要求5所述的方法,其特征在于,所述在所述迭代轮数未达到所述轮数阈值的情况下,基于所述对战状态数据以及所述动作指令,确定第三奖励值,包括:

在所述迭代轮数未达到所述轮数阈值的情况下,基于所述对战状态数据以及动作指令确定合理性参数,所述合理性参数用于表征在所述对战状态数据所表征对战状态下执行所

述动作指令所表征动作的合理性；

基于所述合理性参数确定所述第三奖励值，所述第三奖励值与所述合理性参数呈正相关关系。

7.根据权利要求5所述的方法，其特征在于，所述基于所述第一奖励值、所述第二奖励值以及所述第三奖励值训练所述第n决策模型，得到所述第n+1决策模型，包括：

基于所述第一奖励值、所述第二奖励值以及基础奖励权重，确定基础奖励值；

基于所述第三奖励值和辅助奖励权重，确定辅助奖励值，所述辅助奖励权重小于所述基础奖励权重，且所述辅助奖励权重与所述迭代轮数呈负相关关系；

基于基础奖励值和所述辅助奖励值训练所述第n决策模型，得到所述第n+1决策模型。

8.根据权利要求1至7任一所述的方法，其特征在于，所述方法还包括：

从第二模型池中采样得到对手决策模型，所述第二模型池中包含历史迭代过程中训练得到的至少一个所述决策模型；

在第n轮迭代过程中，将第i对手对战状态数据输入所述对手决策模型，得到所述对手决策模型输出的对手动作指令，所述对手动作指令用于指示第二客户端控制第二虚拟角色执行动作，并反馈所述第二虚拟角色执行动作后的收集到的第i+1对手对战状态数据，所述第二虚拟角色是与所述第一虚拟角色对战的虚拟角色。

9.根据权利要求8所述的方法，其特征在于，所述从第二模型池中采样得到对手决策模型，包括：

基于所述第二模型池中各个决策模型对应的历史胜率，从所述第二模型池中采样得到所述对手决策模型，其中，所述决策模型的采样率与所述历史胜率呈正相关关系；

所述方法还包括：

基于所述第一虚拟角色与所述第二虚拟角色的对战结果更新所述对手决策模型的所述历史胜率。

10.根据权利要求8所述的方法，其特征在于，所述方法还包括：

在采用所述对手决策模型进行对战的对战次数达到次数阈值的情况下，重新从所述第二模型池中采样得到所述对手决策模型。

11.根据权利要求8所述的方法，其特征在于，所述方法还包括：

将所述第n+1决策模型添加至所述第二模型池。

12.根据权利要求1至7任一所述的方法，其特征在于，所述第i对战状态数据包含基础数据、角色增益数据以及元件数据；

所述将第i对战状态数据输入第n决策模型，得到所述第n决策模型输出的动作指令，包括：

将所述基础数据输入所述第n决策模型的第一全连接网络；

将所述角色增益数据输入所述第n决策模型的第二全连接网络；

将所述元件数据输入所述第n决策模型的第三全连接网络；

拼接所述第一全连接网络、所述第二全连接网络以及所述第三全连接网络的全连接处理结果，并输入所述第n决策模型的第四全连接网络，得到所述第四全连接网络输出的动作采样概率；

基于所述动作采样概率进行动作采样，得到所述动作指令。

13.根据权利要求12所述的方法,其特征在于,所述方法还包括:

基于所述第i对战状态数据确定动作屏蔽信息,所述动作屏蔽信息用于指示在当前对战状态下所述第一虚拟角色无法执行的动作;

基于所述动作屏蔽信息对所述动作采样概率进行概率屏蔽处理,所述概率屏蔽处理用于屏蔽所述第一虚拟角色无法执行的动作的采样概率;

所述基于所述动作采样概率进行动作采样,得到所述动作指令,包括:

基于概率屏蔽处理后的所述动作采样概率进行动作采样,得到所述动作指令。

14.根据权利要求1至7任一所述的方法,其特征在于,所述方法包括:

在满足迭代训练结束条件的情况下,将所述第一模型池中最后一轮训练得到的决策模型确定为目标决策模型。

15.一种决策模型的训练装置,其特征在于,所述装置包括:

获取模块,用于获取第一模型池,所述第一模型池中包含至少一个决策模型,所述决策模型用于指示虚拟角色在对战时采用的对战策略;

第一输出模块,用于在第n轮迭代过程中,将第i对战状态数据输入第n决策模型,得到所述第n决策模型输出的动作指令,所述第n决策模型为上一轮迭代过程训练得到的模型,所述第i对战状态数据用于表征对战过程中第i对战时刻时的对战状态,所述动作指令用于指示第一客户端控制第一虚拟角色执行动作,并反馈所述第一虚拟角色执行动作后的收集到的第i+1对战状态数据,所述动作指令包括移动指令和技能使用指令中的至少一种,所述第i+1对战状态数据包括所述第一虚拟角色执行动作后所述第一虚拟角色以及敌对虚拟角色的状态数据,所述状态数据包括位置状态数据、属性状态数据和技能状态数据中的至少一种;

训练模块,用于基于对战过程中的对战状态数据训练所述第n决策模型,得到第n+1决策模型;

第一添加模块,用于将所述第n+1决策模型添加至所述第一模型池。

16.一种计算机设备,其特征在于,所述计算机设备包括处理器和存储器,所述存储器中存储有至少一条指令、至少一段程序、代码集或指令集,所述至少一条指令、所述至少一段程序、所述代码集或指令集由所述处理器加载并执行以实现如权利要求1至14任一所述的决策模型的训练方法。

17.一种计算机可读存储介质,其特征在于,所述计算机可读存储介质中存储有至少一条指令、至少一段程序、代码集或指令集,所述至少一条指令、所述至少一段程序、所述代码集或指令集由处理器加载并执行以实现如权利要求1至14任一所述的决策模型的训练方法。

18.一种计算机程序产品,其特征在于,所述计算机程序产品包括计算机指令,所述计算机指令被处理器执行时实现如权利要求1至14任一所述的决策模型的训练方法。

决策模型的训练方法、装置、计算机设备及存储介质

技术领域

[0001] 本申请实施例涉及人工智能领域,特别涉及一种决策模型的训练方法、装置、计算机设备及存储介质。

背景技术

[0002] 格斗游戏中,在新手教程、人机模式等智能场景中,玩家控制的虚拟角色可以与具有一定策略和决策能力的游戏AI(Artificial Intelligence,人工智能)进行对战。

[0003] 相关技术中,通常采用监督学习的方法训练游戏AI,即利用玩家在游戏过程中产生的对战数据对游戏AI进行训练学习。但是在训练学习过程中,由于玩家的对战数据可靠性较差,容易产生过拟合问题,导致游戏AI的实际表现和预期值相差较大,使得游戏AI的拟人化效果差。

发明内容

[0004] 本申请实施例提供了一种决策模型的训练方法、装置、计算机设备及存储介质,能够提高决策模型的对战拟人性,所述技术方案如下:

[0005] 一方面,本申请实施例提供了一种决策模型的训练方法,所述方法包括:

[0006] 获取第一模型池,所述第一模型池中包含至少一个决策模型,所述决策模型用于指示虚拟角色在对战时采用的对战策略;

[0007] 在第n轮迭代过程中,将第i对战状态数据输入第n决策模型,得到所述第n决策模型输出的动作指令,所述第n决策模型为上一轮迭代过程训练得到的模型,所述第i对战状态数据用于表征对战过程中第i对战时刻时的对战状态,所述动作指令用于指示第一客户端控制第一虚拟角色执行动作,并反馈所述第一虚拟角色执行动作后的收集到的第i+1对战状态数据;

[0008] 基于对战过程中的对战状态数据训练所述第n决策模型,得到第n+1决策模型;

[0009] 将所述第n+1决策模型添加至所述第一模型池。

[0010] 另一方面,本申请实施例提供了一种决策模型的训练装置,所述装置包括:

[0011] 获取模块,用于获取第一模型池,所述第一模型池中包含至少一个决策模型,所述决策模型用于指示虚拟角色在对战时采用的对战策略;

[0012] 第一输出模块,用于在第n轮迭代过程中,将第i对战状态数据输入第n决策模型,得到所述第n决策模型输出的动作指令,所述第n决策模型为上一轮迭代过程训练得到的模型,所述第i对战状态数据用于表征对战过程中第i对战时刻时的对战状态,所述动作指令用于指示第一客户端控制第一虚拟角色执行动作,并反馈所述第一虚拟角色执行动作后的收集到的第i+1对战状态数据;

[0013] 训练模块,用于基于对战过程中的对战状态数据训练所述第n决策模型,得到第n+1决策模型;

[0014] 第一添加模块,用于将所述第n+1决策模型添加至所述第一模型池。

[0015] 另一方面,本申请实施例提供了一种计算机设备,所述计算机设备包括处理器和存储器,所述存储器中存储有至少一条指令、至少一段程序、代码集或指令集,所述至少一条指令、所述至少一段程序、所述代码集或指令集由所述处理器加载并执行以实现如上述方面所述的决策模型的训练方法。

[0016] 另一方面,本申请实施例提供了一种计算机可读存储介质,所述计算机可读存储介质中存储有至少一条指令、至少一段程序、代码集或指令集,所述至少一条指令、所述至少一段程序、所述代码集或指令集由处理器加载并执行以实现如上述方面所述的决策模型的训练方法。

[0017] 另一方面,本申请实施例提供了一种计算机程序产品或计算机程序,该计算机程序产品或计算机程序包括计算机指令,该计算机指令存储在计算机可读存储介质中。计算机设备的处理器从计算机可读存储介质读取该计算机指令,处理器执行该计算机指令,使得该计算机设备执行上述方面的各种可选实现方式中提供的决策模型的训练方法。

[0018] 本申请实施例提供的技术方案的有益效果至少包括:

[0019] 每轮迭代中,将第i对战状态数据输入第n决策模型,得到第n决策模型输出的动作指令,该动作指令用于指示第一客户端控制第一虚拟角色执行动作。第一客户端基于第一虚拟角色执行动作后,反馈收集到的第i+1对战状态数据,从而学习建立了从一个状态映射到一个动作的映射关系,进而通过对战过程中的对战状态数据训练第n决策模型,得到第n+1决策模型。通过对战过程中的对战状态数据训练决策模型,提高决策模型的决策能力,使其更加接近真实玩家的表现,进而增强决策模型的对战拟人性。

附图说明

[0020] 图1示出了本申请一个示例性实施例提供的决策模型的训练方法的原理示意图;

[0021] 图2示出了本申请一个示例性实施例提供的实施环境的示意图;

[0022] 图3示出了本申请一个示例性实施例提供的决策模型的训练方法的流程图;

[0023] 图4示出了本申请一个示例性实施例提供的游戏对战时客户端与服务器的交互示意图;

[0024] 图5示出了本申请另一个示例性实施例提供的决策模型的训练方法的流程图;

[0025] 图6示出了本申请另一个示例性实施例提供的决策模型的训练方法的流程图;

[0026] 图7示出了本申请一个示例性实施例提供的决策模型的训练方法的过程示意图;

[0027] 图8示出了本申请另一个示例性实施例提供的决策模型的训练方法的流程图;

[0028] 图9示出了本申请一个示例性实施例提供的决策模型基于对战状态数据输出动作指令过程的流程图;

[0029] 图10示出了本申请一个示例性实施例提供的决策模型基于对战状态数据输出动作指令过程的示意图;

[0030] 图11示出了本申请一个示例性实施例提供的决策模型的训练装置的结构框图;

[0031] 图12示出了本申请一个示例性实施例提供的服务器的结构框图。

具体实施方式

[0032] 为使本申请的目的、技术方案和优点更加清楚,下面将结合附图对本申请实施方

式作进一步地详细描述。

[0033] 在本文中提及的“多个”是指两个或两个以上。“和/或”,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。字符“/”一般表示前后关联对象是一种“或”的关系。

[0034] 首先,对本申请实施例中涉及的名词进行介绍:

[0035] 虚拟环境:是应用程序在客户端上运行时显示(或提供)的虚拟环境。该虚拟环境可以是对真实世界的仿真环境,也可以是半仿真半虚构的环境,还可以是纯虚构的环境。虚拟环境可以是二维虚拟环境、2.5维虚拟环境和三维虚拟环境中的任意一种,本申请对此不加以限定。下述实施例以虚拟环境是三维虚拟环境为例进行说明。

[0036] 虚拟角色:是指虚拟环境中的可活动对象。该可活动对象可以是虚拟人物、虚拟动物、动漫人物等,比如:在三维虚拟环境中显示的人物、动物。可选地,虚拟角色是基于动画骨骼技术创建的三维立体模型。每个虚拟角色在三维虚拟环境中具有自身的形状和体积,占据三维虚拟环境中的一部分空间。

[0037] 强化学习:强化学习又称再励学习、评价学习或增强学习,是机器学习的范式和方法论之一,用于描述和解决智能体(agent)与环境的交互过程中通过学习策略以达成回报最大化或实现特定目标的问题。强化学习是智能体(agent)以“试错”的方式进行学习,通过与环境进行交互获得的奖赏指导行为,目标是使智能体获得最大的奖赏。

[0038] 人工智能(Artificial Intelligence,AI)是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能,感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。换句话说,人工智能是计算机科学的一个综合技术,它企图了解智能的实质,并生产出一种新的能以人类智能相似的方式做出反应的智能机器。人工智能也就是研究各种智能机器的设计原理与实现方法,使机器具有感知、推理与决策的功能。

[0039] 人工智能技术是一门综合学科,涉及领域广泛,既有硬件层面的技术也有软件层面的技术。人工智能基础技术一般包括如传感器、专用人工智能芯片、云计算、分布式存储、大数据处理技术、操作/交互系统、机电一体化等技术。人工智能软件技术主要包括计算机视觉技术、语音处理技术、自然语言处理技术以及机器学习/深度学习等几大方向。

[0040] 机器学习(Machine Learning,ML)是一门多领域交叉学科,涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。专门研究计算机怎样模拟或实现人类的学习行为,以获取新的知识或技能,重新组织已有的知识结构使之不断改善自身的性能。机器学习是人工智能的核心,是使计算机具有智能的根本途径,其应用遍及人工智能的各个领域。机器学习和深度学习通常包括人工神经网络、置信网络、强化学习、迁移学习、归纳学习、示教学习等技术。

[0041] 目前,以格斗游戏为例,在新手教程、人机模式等游戏场景中,玩家控制的虚拟角色可以具备一定策略和决策能力的游戏人工智能体,即游戏AI进行对战。相关技术中,通过客户端内置的行为树控制游戏AI的行为,或者利用玩家在游戏过程中产生的对战数据进行模仿学习,训练出游戏状态到输出动作的映射网络,来控制游戏AI和玩家控制的虚拟角色进行对战。但是相关技术中,行为树结构固定,使得游戏AI在游戏表现中单一,缺少变化,无法接近真实玩家的表现。另外,由于玩家的对战数据可靠性较差,容易导致在训练过程中产生过拟合问题,从而导致游戏AI的实际表现和预期值相差较大,游戏AI拟人化效果差。

[0042] 因此在本申请实施例中,通过强化学习方法训练决策模型,相比于相关技术中的方案,提高了决策模型的决策能力,使其更加接近真实玩家的表现,进而使该决策模型的对战拟人性更强。请参考图1,其示出了本申请一个示例性实施例提供的决策模型的训练方法的原理示意图。

[0043] 客户端110运行有游戏应用程序,游戏开局时,客户端110向服务器发送开局请求,由服务器120根据客户端的110的需求返回相应的对战配置数据,客户端获取对战配置数据后,构建对局并生成游戏AI。游戏AI用于控制虚拟角色进行对战。

[0044] 客户端110设置有SDK(Software Development Kit,软件开发工具包),对战开始后,客户端110通过SDK与服务器120进行交互。SDK构建当前对战状态数据并上传至服务器120,服务器120的决策模型基于对战数据输出动作指令,SDK从服务器120处获取动作指令,客户端110根据该动作指令控制虚拟角色111执行动作,同时收集虚拟角色111执行动作后的下一对战状态数据,并将其反馈至SDK。SDK再将下一对战状态数据上传至服务器120。服务器120基于当前对战状态数据以及下一对战状态数据训练决策模型。下面对决策模型的训练方法进行介绍。

[0045] 请参考图2,其示出了本申请一个实施例提供的实施环境的示意图。该实施环境可以包括:第一客户端210、服务器220和第二客户端230。

[0046] 第一客户端210安装和运行有支持虚拟环境的应用程序211,该应用程序211可以是多人在线对战程序。当第一客户端210运行应用程序211时,第一客户端210的屏幕上显示应用程序211的用户界面。该应用程序211可以是多人在线战术竞技游戏(Multiplayer Online Battle Arena Games,MOBA)、射击游戏、模拟战略游戏(Simulation Game,SLG)、格斗游戏中的任意一种。在本实施例中,以该应用程序211是格斗游戏来举例说明。第一客户端210上设置有SDK,第一客户端通过SDK与服务器220进行交互,包括构建对战状态数据并上传至服务器220,从服务器220处获取动作指令。在本申请实施例中,第一客户端210对应第一虚拟角色,示例性的,第一虚拟角色可以是仿真人物角色或动漫人物角色。格斗游戏应用程序从SDK获取到服务器220发出的动作指令,第一虚拟角色执行相应的动作,模拟出人为控制的效果,只是这里的控制并不是通过外部设备(比如触摸屏或摇杆按键)输入,而是由决策模型生成。

[0047] 第二客户端230安装和运行有支持虚拟环境的应用程序231,该应用程序231可以是多人在线对战程序。当第二客户端230运行应用程序231时,第二客户端230的屏幕上显示应用程序231的用户界面。该客户端可以是MOBA游戏、射击游戏、SLG游戏、格斗游戏中的任意一种,在本实施例中,以该应用程序231是格斗游戏来举例说明。同样的,第二客户端230上设置有SDK,第二客户端通过SDK与服务器220进行交互,包括构建对战状态数据并上传至服务器220,从服务器220处获取动作指令。在本申请实施例中,第二客户端230对应第二虚拟角色,示例性的,第二虚拟角色可以是仿真人物角色或动漫人物角色。格斗游戏应用程序从SDK获取到服务器220发出的动作指令,第二虚拟角色执行相应的动作,模拟出人为控制的效果,只是这里的控制并不是通过外部设备(比如触摸屏或摇杆按键)输入,而是由决策模型生成。

[0048] 可选地,第一客户端210和第二客户端230上安装的应用程序是相同的,或两个客户端上安装的应用程序是不同操作系统平台(安卓或IOS)上的同一类型应用程序。第一客

户端210可以泛指多个客户端中的一个,第二客户端230可以泛指多个客户端中的另一个,本实施例仅以第一客户端210和第二客户端230来举例说明。第一客户端210和第二客户端230的设备类型相同或不同,该设备类型包括:智能手机、平板电脑、电子书阅读器、动态影像专家压缩标准音频层面3(Moving Picture Experts Group Audio Layer III,MP3)播放器、动态影像专家压缩标准音频层面4(Moving Picture Experts Group Audio Layer IV,MP4)播放器、膝上型便携计算机和台式计算机中的至少一种。

[0049] 图2中仅示出了两个客户端,但在不同实施例中存在多个其它客户端可以接入服务器220。第一客户端210、第二客户端230以及其它客户端通过无线网络或有线网络与服务器220相连。

[0050] 服务器220包括一台服务器、多台服务器组成的服务器集群、云计算平台和虚拟化中心中的至少一种。可选地,服务器220承担主要计算工作,客户端承担次要计算工作;或者,服务器220承担次要计算工作,客户端承担主要计算工作;或者,服务器220和客户端之间采用分布式计算架构进行协同计算。

[0051] 在本申请实施例中,服务器220包括存储器221、处理器222、账号数据库223、对战服务模块224、输入/输出接口(Input/Output Interface,I/O接口)225。其中,处理器222用于加载服务器220中存储的指令,处理账号数据库223和对战服务模块224中的数据;账号数据库223用于存储第一客户端210、第二客户端230以及其它客户端所使用的账号的数据,比如账号的头像、账号的昵称、账号的等级,账号所在的服务区;对战服务模块224用于提供多个对战房间供游戏AI进行对战,比如1V1对战、3V3对战、5V5对战等;I/O接口225用于通过无线网络或有线网络和第一客户端210和/或第二客户端230建立通信交换数据。

[0052] 在一种可能的实施方式中,在本申请实施例中,服务器220构建第一模型池和第二模型池。第一模型池中至少包含一个决策模型,第二模型池中至少包含历史迭代过程中训练得到的决策模型。在第n轮迭代训练过程中,服务器从第二模型池中采样得到对手决策模型与第n决策模型进行对战,基于对战状态数据训练得到n+1决策模型,并将n+1决策模型分别加入第一模型池和第二模型池,用于下一轮迭代训练。

[0053] 另外,需要说明的是,在本申请实施例中,图2中以训练方法用于服务器为例进行说明,在一种可能的实施方式中,也可以应用于个人计算机或工作站等计算机设备中,下述实施例以决策模型的训练方法用于服务器为例进行示意性说明,但并不对此构成限定。

[0054] 请参考图3,其示出了本申请一个示例性实施例提供的决策模型的训练方法的流程图,该方法包括如下步骤:

[0055] 步骤310,获取第一模型池,第一模型池中包含至少一个决策模型,决策模型用于指示虚拟角色在对战时采用的对战策略。

[0056] 在格斗游戏中,存在敌对双方进行作战,因此在本申请实施例中,在一种可能的实施方式中,由第一客户端控制的第一虚拟角色与敌对虚拟角色进行对战。服务器中设置有第一模型池,第一模型池中的决策模型用于指示第一客户端控制第一虚拟角色执行动作,进而与敌对虚拟角色进行对战。

[0057] 可选地,对战策略可以是虚拟角色的移动方向,例如向上移动、向下移动、向左移动、向右移动等,本申请实施例对此不作限定。

[0058] 可选地,对战策略也可以是虚拟角色的技能输出,例如选择替身、召唤通灵、使用

秘卷等进行作战等,本申请实施例对此不作限定。

[0059] 另外,在本申请实施例中,第一模型池中除了包含至少一个决策模型之外,还包含客户端内置的行为树,用于与初始的决策模型进行对战。

[0060] 步骤320,在第n轮迭代过程中,将第i对战状态数据输入第n决策模型,得到第n决策模型输出的动作指令,第n决策模型为上一轮迭代过程训练得到的模型,第i对战数据用于表征对战过程中第i对战时刻时的对战状态,动作指令用于指示第一客户端控制第一虚拟角色执行动作,并反馈第一虚拟角色执行动作后的收集到的第i+1对战状态数据。

[0061] 其中,对战状态数据用于表征一个对战时刻时的对战状态,每个对战时刻对应一个对战状态。示例性的,对战状态数据包括第一虚拟角色以及敌对虚拟角色对战画面中全部游戏信息。

[0062] 可选地,对战状态数据可以是第一虚拟角色的位置和朝向、敌对虚拟角色的位置和朝向、第一虚拟角色以及敌对虚拟角色的血量、忍术点数等,技能等级、可使用的技能、技能的冷却状态、技能的类型、技能的招式,例如通灵、秘卷、受击、替身状态等,游戏对战时长等,本申请实施例对此不作限定。

[0063] 其中,动作指令包括移动指令和技能使用指令,移动指令可以是上下移动或者左右移动等,技能使用可以是使用技能、替身、通灵等。决策模型输出的动作指令与第一客户端控制第一虚拟角色执行的动作一一对应。动作指令不同,第一客户端控制第一虚拟角色执行不同的动作。

[0064] 在一种可能的实施方式中,避免服务器与客户端交互的频率过于频繁而导致训练出的决策模型表现过于激进,因此客户端周期性地向第n决策模型反馈对战状态数据,也就是说第i对战时刻和第i+1对战时刻间隔一定的时间。

[0065] 可选地,间隔时间可以是100ms、200ms等,本申请实施例对此不作限定。

[0066] 示例性的,如图4所示,客户端每隔3帧游戏帧(即100ms)与服务器进行一次交互。客户端SDK将第1帧到第3帧游戏帧对应的对战状态数据,即第1对战状态数据上传至服务器,服务器将该对战状态数据输入第n决策模型,得到动作指令,并将动作指令反馈至SDK。基于该动作指令,客户端控制虚拟角色执行动作,SDK基于虚拟角色执行的动作收集第4帧到第6帧游戏帧对应的对战状态数据,即第2对战状态数据,并将其上传至服务器。

[0067] 示例性的,在第2轮迭代过程中,服务器将第1对战状态数据输入第2决策模型,得到第2决策模型输出的动作指令,其中第2轮决策模型为第1轮迭代过程中训练得到的模型。服务器将第2决策模型输出的动作指令传输至第一客户端,第一客户端基于第2决策模型输出的动作指令控制第一虚拟角色执行动作,同时客户端将第一虚拟角色执行动作后的收集到的第2对战状态数据,反馈至服务器,用于服务器训练下一轮决策模型。

[0068] 步骤330,基于对战过程中的对战状态数据训练第n决策模型,得到第n+1决策模型。

[0069] 在一种可能的实施方式中,服务器基于第i对战状态数据以及第i+1对战状态数据得到对战状态数据的变化值,基于对战状态数据的变化值训练第n决策模型,得到第n+1决策模型。

[0070] 示例性的,服务器基于第1对战状态数据以及第2对战状态数据得到对战状态数据的变化值,基于该对战状态数据的变化值训练第2决策模型,得到第3决策模型。

[0071] 步骤340,将第n+1决策模型添加至第一模型池。

[0072] 在本申请实施例中,服务器将每轮迭代过程中训练得到的决策模型添加至第一模型池,用于后续决策模型的训练。

[0073] 示例性的,服务器将第3决策模型添加至第一模型池。

[0074] 综上所述,本申请实施例中,每轮迭代中,将第i对战状态数据输入第n决策模型,得到第n决策模型输出的动作指令,该动作指令用于指示第一客户端控制第一虚拟角色执行动作。第一客户端基于第一虚拟角色执行动作后,反馈收集到的第i+1对战状态数据,从而学习建立了从一个状态映射到一个动作的映射关系,进而通过对战过程中的对战状态数据训练第n决策模型,得到第n+1决策模型。通过对战过程中的对战状态数据训练决策模型,提高决策模型的决策能力,使其更加接近真实玩家的表现,进而增强决策模型的对战拟人性。

[0075] 为了增强决策模型的对战拟人性,使得决策模型指示的虚拟角色在对战时采用的对战策略更接近真实玩家控制的虚拟角色在对战时采用的对战策略,在一种可能的实施方式中,引入强化学习的思想,对于决策模型输出的动作指令,需要根据基于相邻对战时刻对应的对战状态数据确定第一奖励值,基于对战结果确定第二奖励值,基于第一奖励值以及第二奖励值训练决策模型。请参考图5,其示出了本申请另一个示例性实施例提供的决策模型的训练方法的流程图。

[0076] 步骤501,获取第一模型池,第一模型池中包含至少一个决策模型,决策模型用于指示虚拟角色在对战时采用的对战策略。

[0077] 本步骤的实施方式请参照步骤310,本申请实施例对此不再赘述。

[0078] 步骤502,在第n轮迭代过程中,将第i对战状态数据输入第n决策模型,得到第n决策模型输出的动作指令,第n决策模型为上一轮迭代过程训练得到的模型,第i对战状态数据用于表征对战过程中第i对战时刻时的对战状态,动作指令用于指示第一客户端控制第一虚拟角色执行动作,并反馈第一虚拟角色执行动作后的收集到的第i+1对战状态数据。

[0079] 本步骤的实施方式请参照步骤320,本申请实施例对此不再赘述。

[0080] 步骤503,获取第i对战状态数据中的第i角色属性值,以及第i+1对战状态数据中的第i+1角色属性值,其中,角色属性值包括第一虚拟角色以及与第二虚拟角色的属性值,第二虚拟角色是与第一虚拟角色对战的虚拟角色。

[0081] 其中,角色属性值是指对战状态数据中影响对局结果的第一虚拟角色以及第二虚拟角色的属性值。

[0082] 可选地,角色属性值可以是第一虚拟角色以及第二虚拟角色的血量、蓝条等,本申请实施例对此不作限定。

[0083] 步骤504,基于第i角色属性值与第i+1角色属性值的属性值变化情况,确定第一奖励值。

[0084] 在一种可能的实施方式中,第一奖励值可以通过下列公式计算得到:

[0085] 第一奖励值=[(第一虚拟角色的第i+1角色属性值)-(第一虚拟角色的第i角色属性值)+(第二虚拟角色的第i+1角色属性值)-(第二虚拟角色的第i角色属性值)]×第一奖励系数。

[0086] 示例性的,角色属性值为第一虚拟角色以及第二虚拟角色的血量。第1对战状态数

据中,第一虚拟角色的血量为100,第二虚拟角色的血量为80,第2对战状态数据中,第一虚拟角色的血量为150,第二虚拟角色的血量为70。其中第一虚拟角色的血量变化为50,第二虚拟角色的血量变化为-10。第一奖励系数为20,则第一奖励值为800。

[0087] 步骤505,基于对战结果确定第二奖励值,对战结果用于指示对战的胜负结果。

[0088] 在一种可能的实施方式中,服务器基于第一虚拟角色胜利的胜率确定第二奖励值。其中,胜率低于50%,第二奖励值为负值,胜率高于50%,第二奖励值为正值。第二奖励值可以通过下列公式计算得到:

[0089] 第二奖励值=第一虚拟角色的胜率×第二奖励系数。

[0090] 示例性的,第一轮迭代过程中,第一虚拟角色与第二虚拟角色共进行100回合对战,第一虚拟角色的胜率为30%,第二奖励系数为30,第二奖励值为-900。

[0091] 步骤506,基于第一奖励值和第二奖励值训练第n决策模型,得到第n+1决策模型。

[0092] 在一种可能的实施方式中,服务器针对第一奖励值和第二奖励值分别设置有不同的权重系数,根据第一奖励值、第二奖励值以及各自的权重系数计算总奖励值。服务器基于总奖励值训练第n决策模型,得到第n+1决策模型。

[0093] 另外,为了以胜利为导向训练决策模型,第一奖励值的权重小于第二奖励值的权重。

[0094] 步骤507,将第n+1决策模型添加至第一模型池。

[0095] 本步骤的实施方式请参照步骤340,本申请实施例对此不再赘述。

[0096] 在本申请实施例中,引入强化学习的思想,基于相邻对战时刻对应的对战状态数据中的角色属性值确定第一奖励值,基于对战结果确定第二奖励值,服务器基于第一奖励值以及第二奖励值训练决策模型,增强决策模型的对战拟人性。

[0097] 在一种可能的实施方式中,如表一所示,从基础奖励值和辅助奖励值两方面进一步优化服务器训练决策模型的奖励值,从而进一步提高决策模型的对战拟人性。

[0098] 表一

[0099]

奖励值主分类	奖励值子分类	奖惩维度
基础奖励值	战斗核心	角色属性值变化
		对战结果
辅助奖励值	替身拟人化	替身使用
		替身反击
		替身剩余量
		替身无敌利用
	技能拟人化	通灵合理使用
		秘卷合理使用
		技能多样性分布
		大招合理使用
	移动拟人化	开局移动
		优势躲避战斗

[0100] 下面结合表一所示的奖励维度,对本申请实施例中决策模型的训练方法进行介绍。请参考图6,其示出了本申请另一个示例性实施例提供的决策模型的训练方法的流程图。

[0101] 步骤601,获取第一模型池,第一模型池中包含至少一个决策模型,决策模型用于指示虚拟角色在对战时采用的对战策略。

[0102] 本步骤的实施方式请参考步骤310,本申请实施例对此不再赘述。

[0103] 步骤602,在第n轮迭代过程中,将第i对战状态数据输入第n决策模型,得到第n决策模型输出的动作指令,第n决策模型为上一轮迭代过程训练得到的模型,第i对战状态数据用于表征对战过程中第i对战时刻时的对战状态,动作指令用于指示第一客户端控制第一虚拟角色执行动作,并反馈第一虚拟角色执行动作后的收集到的第i+1对战状态数据。

[0104] 本步骤的实施方式请参考步骤320,本申请实施例对此不再赘述。

[0105] 步骤603,在迭代轮数未达到轮数阈值的情况下,基于对战状态数据以及动作指令确定合理性参数,合理性参数用于表征在对战状态数据所表征对战状态下执行动作指令所表征动作的合理性。

[0106] 在本申请实施例中,为了引导决策模型在低轮次迭代过程中学习拟人化的基础规则与打法,在迭代轮数未达到轮数阈值的情况下,引入合理性参数作为拟人化的评估标准。而在迭代轮数达到轮数阈值的情况下,侧重训练决策模型的强度,因此并不引入合理参数。

[0107] 在一种可能的实施方式中,服务器预先设置了不同动作指令是否合理的判断标准,并根据该判断标准确定合理性参数区间。根据对战状态数据以及动作指令服务器从合理性参数区间内选择一个合理性参数。

[0108] 可选地,不同动作指令是否合理的判断标准不同。如表二所示,根据动作指令的类型不同,其合理性的标准也不同。

[0109] 表二

[0110]

动作指令类型	合理性的判断标准
替身使用	不使用替身或者使用时机正确
替身反击	在替身后抓住敌人破绽进行伤害输出、硬直控制
替身剩余量	存在1个以上的替身，最多4个替身
替身无敌利用	替身后会有1.5秒的无敌时间，在这段时间内造成伤害、控制敌人、逃出危险区等
通灵/秘卷合理利用	对手处于硬直或对手没有在攻击/释放通灵秘卷动作中
大招合理使用	使用大招斩杀对手并且对手替身在冷却中
技能多样性	使用一个以上的技能对战
开局移动	游戏开局进行走位
优势躲避战斗	获取血量优势后停止移动

[0111] 步骤604,基于合理性参数确定第三奖励值,第三奖励值与合理性参数呈正相关关系。

[0112] 在一种可能的实施方式中,服务器基于对战状态数据以及动作指令从合理性参数区间内选择合理参数,基于合理性参数确定第三奖励值,其中合理性参数越大,第三奖励值越大。

[0113] 示例性的,合理性参数区间范围为0至100,100用于表征动作合理,0用于表征动作不合理,因此合理性参数越接近100,第三奖励值越大。

[0114] 在一种可能的实施方式中,第三奖励值的可以通过下列公式计算得到:

[0115] 第三奖励值=合理性参数×参数系数。

[0116] 示例性的,服务器确定的合理性参数为40,参数系数为10,则第三奖励值为400。

[0117] 步骤605,基于第一奖励值、第二奖励值以及基础奖励权重,确定基础奖励值。

[0118] 在一种可能的实施方式中,基础奖励值可以通过下述的公式进行计算得到:

[0119] 基础奖励值=(第一奖励值+第二奖励值)×基础奖励权重。

[0120] 示例性的,第一奖励值为200,第二奖励值为100,基础奖励权重为3,那么基础奖励值为900。

[0121] 另外,在迭代过程中,基础奖励权重并不是固定不变的,而是根据训练过程中决策模型输出的动作指令以及对战状态数据而调整变化,目的是为了提高模型训练的准确性。

[0122] 例如,在迭代过程中,如果根据对战状态数据中的对战结果可知,第一虚拟角色胜率较低,则增加基础奖励权重,达到以胜利为导向的训练目的。

[0123] 步骤606,基于第三奖励值和辅助奖励权重,确定辅助奖励值,辅助奖励权重小于基础奖励权重,且辅助奖励权重与迭代轮数呈负相关关系。

[0124] 在一种可能的实施方式中,辅助奖励值可以通过下述的公式进行计算得到:

[0125] 辅助奖励值＝第三奖励值×辅助奖励权重。

[0126] 示例性的,第三奖励值为300,辅助奖励值权重为2,那么辅助奖励值为600。

[0127] 另外,在迭代过程中,辅助奖励权重并不是固定的,而是随着迭代轮数的增加而减小,当迭代轮数达到轮数阈值的情况下,辅助奖励权重减小为0。

[0128] 步骤607,基于基础奖励值和辅助奖励值训练第n决策模型,得到第n+1决策模型。

[0129] 在一种可能的实施方式中,当迭代轮数未达到轮数阈值时,服务器基于基础奖励值和辅助奖励值训练第n决策模型,得到第n+1决策模型,目的是为了引导决策模型学习拟人化的基础规则与打法,使得决策模型更接近真实玩家的表现。

[0130] 步骤608,基于相邻对战时刻对应的对战状态数据,确定第一奖励值。

[0131] 本步骤的实施方式请参考步骤503至步骤504,本申请实施例对此不再赘述。

[0132] 步骤609,基于对战结果确定第二奖励值,对战结果用于指示对战的胜负结果。

[0133] 本步骤的实施方式请参考步骤505,本申请实施例对此不再赘述。

[0134] 步骤610,在迭代轮数达到轮数阈值的情况下,基于第一奖励值和第二奖励值训练第n决策模型,得到第n+1决策模型。

[0135] 在本申请实施例中,为了引导决策模型在高轮次迭代过程中提升强度,因此在迭代轮数达到轮数阈值的情况下,基于第一奖励值和第二奖励值训练决策模型。

[0136] 示例性的,当迭代轮数达到1000轮的情况下,服务器基于第一奖励值和第二奖励值训练第2决策模型,得到第3决策模型。

[0137] 步骤611,将第n+1决策模型添加至第一模型池。

[0138] 本步骤的实施方式请参考步骤340,本申请实施例对此不再赘述。

[0139] 步骤612,在满足迭代训练结束条件的情况下,将第一模型池中最后一轮训练得到的决策模型确定为目标决策模型。

[0140] 在一种可能的实施方式中,当第一模型池中最后一轮训练得到的决策模型的胜率和前一轮训练得到的决策模型的胜率差低于训练阈值时,表明决策模型的性能趋于稳定,继续训练对决策模型的性能提醒较小,此时满足迭代训练结束条件,将最后一轮训练得到的决策模型确定为目标决策模型。

[0141] 示例性的,如表三所示,其示出了本申请一个示例性实施例提供的第一模型池中每轮训练得到的决策模型的胜率。

[0142] 表三

[0143]

	行为树1	行为树2	第1决策模型	第2决策模型	第3决策模型	第4决策模型	第5决策模型	第6决策模型

[0144]

行为树1	——	——	——	——	——	——	——	——
行为数2	——	——	——	——	——	——	——	——
第1决策模型	92.5%	93.43%	——	——	——	——	——	——
第2决策模型	——	——	96.51%	——	——	——	——	——
第3决策模型	——	——	95.87%	90.06%	——	——	——	——
第4决策模型	——	——	95%	86.1%	68.74%	——	——	——
第5决策模型	——	——	94.3%	86.2%	66.3%	57.73%	——	——
第6决策模型	——	——	93.77%	80.18%	64.25%	59.68%	56.19%	

[0145] 表三中,百分比用于表示“列模型”对战“行模型”的胜率,其中,“列模型”为第一模型池中的决策模型,“行模型”第二模型池中的对手决策模型。由表二可知,随着迭代轮数的增加,第一模型池中训练得到的决策模型的胜率逐渐趋于平稳,即第6决策模型的胜率和第5决策模型的胜率相比基本没有发生变化,因此可以将第6决策模型作为目标决策模型。

[0146] 在本申请实施例中,在迭代轮数未达到轮数阈值的情况下,基于基础奖励值以及辅助奖励值训练决策模型,使得决策模型在低轮次的迭代过程中学习拟人化的基础规则与打法,从而增强决策模型的对战拟人性,当迭代轮数达到轮数阈值,仅通过第一奖励值以及第二奖励值训练决策模型,从而提高决策模型的强度,进而使得最终训练得到的目标决策模型接近真实玩家的表现。

[0147] 示例性的,以第2轮迭代过程为例,结合图7对决策模型的训练方法进行介绍。服务器71中设置第一模型池711以及第二模型池712,其中第一模型池711至少包含一个决策模型,第二模型池712至少包含一个历史迭代过程中训练得到的决策模型。服务器71从第二模型池712中选择对手决策模型,作为第2决策模型的对手。

[0148] 在第2轮迭代过程中,服务器71将第1对战状态数据输入第2决策模型,得到第2决策模型输出的动作指令。第一客户端72基于该动作指令控制第一虚拟角色721执行动作,与第二客户端73控制的第二虚拟角色731进行对战。同时收集第一虚拟角色721执行动作后的第2对战状态数据,并将其反馈至服务器71。同时,服务器将第1对战状态数据输入对手决策模型,得到对手决策模型输出的动作指令。第二客户端73基于该动作指令控制第二虚拟角色731执行动作,与第一客户端72控制的第一虚拟角色721进行对战。同时收集第二虚拟角色731执行动作后的第2对战状态,并将其反馈至服务器71。服务器71基于第一客户端72以及第二客户端73反馈的第1对战状态数据以及第2对战状态数据训练第2决策模型,得到第3决策模型,并将第3决策模型分别加入第一模型池711以及第二模型池712。

[0149] 在一种可能的实施方式中,服务器基于对手决策模型的历史胜率从第二模型池中采样对手决策模型,进一步提高训练过程中决策模型的强度。请参考图8,其示出了本申请一个示例性实施例提供的决策模型的训练方法的流程图。

[0150] 步骤801,基于第二模型池中各个决策模型对应的历史胜率,从第二模型池中采样得到对手决策模型,其中,决策模型的采样率与历史胜率呈正相关关系。

[0151] 其中,第二模型池中包含历史迭代过程中至少一个训练得到的决策模型。

[0152] 在一种可能的实施方式中,同一服务器分别设置有第一模型池和第二模型池。在另一种可能的实施方式中,不同的服务器分别设置有第一模型池和第二模型池。第一模型池中的决策模型用于指示前述第一客户端控制第一虚拟角色执行动作,第二模型池的中的对手决策模型用于指示第二客户端控制第二虚拟角色执行动作。而第二虚拟角色是与第一虚拟角色对战的虚拟角色,即前述敌对虚拟角色。

[0153] 在一种可能的实施方式中,服务器根据决策模型的历史胜率采样对手决策模型。决策模型的历史胜率越高,该模型的采样率越高,也就是被服务器采样作为对手决策模型的概率越大。通过高胜率的对手决策模型进行训练,有利于提高决策模型的强度,进而使其更加接近真实玩家的表现。

[0154] 另外,需要说明的是,在初次迭代过程中,由于没有历史决策模型,因此将客户端内置的行为数作为对手进行模型训练,进而得到第1决策模型,将训练得到的第1决策模型加入第二模型池后。由于第二模型池中已经存在了初次迭代过程中训练得到的第1决策模型,因此在后续的决策模型训练的过程中,则不再使用客户端内置的行为树作为对手进行模型训练。

[0155] 步骤802,基于第一虚拟角色与第二虚拟角色的对战结果更新对手决策模型的历史胜率。

[0156] 为了保证模型训练的准确性,服务器根据第一虚拟角色与第二虚拟角色的对战结果,更新第二模型池中对手决策模型的历史胜率。

[0157] 在一种可能的实施方式中,当第一虚拟角色与第二虚拟角色的对战结果为第二虚拟角色胜,则提高对手决策模型的历史胜率。

[0158] 在另一种可能的实施方式中,当第一虚拟角色与第二虚拟角色的对战结果为第一虚拟角色胜,则降低对手决策模型的历史胜率。

[0159] 步骤803,在第n轮迭代过程中,将第i对手对战状态数据输入对手决策模型,得到对手决策模型输出的对手动作指令,对手动作指令用于指示第二客户端控制第二虚拟角色执行动作,并反馈第二虚拟角色执行动作后的收集到的第i+1对手对战状态数据,第二虚拟角色是与第一虚拟角色对战的虚拟角色。

[0160] 示例性的,在上述第2轮迭代过程中,服务器将第1对手对战状态数据输入对手决策模型,得到对手决策模型输出的对手动作指令。第二客户端控制的第二虚拟角色根据对手动作指令执行动作,与第一客户端控制的第一虚拟角色进行对战。同时,第二客户端向服务器反馈第二虚拟角色执行动作后收集到的第2对手对战状态数据。

[0161] 步骤804,在采用对手决策模型进行对战的对战次数达到次数阈值的情况下,重新从第二模型池中采样得到对手决策模型。

[0162] 为了保证决策模型训练的全面性,当采用第二模型池中的对手决策模型与第一模

型池中的决策模型对战次数达到次数阈值时,服务器重新从第二模型池中采样对手决策模型。

[0163] 示例性的,当第二模型池中的对手决策模型与第一模型池中的决策模型对战次数达到100次的情况下,服务器重新从第二模型池中采样对手决策模型,重新与第一模型池中的决策模型进行对战。

[0164] 步骤805,将第n+1决策模型添加至第二模型池。

[0165] 服务器将每轮迭代过程中训练得到的决策模型添加至第二模型池。

[0166] 示例性的,服务器将前述第2轮迭代过程中得到的第3决策模型添加至第二模型池。

[0167] 在本申请实施例中,服务器基于决策模型的历史胜率从第二模型池中采样对手决策模型,有利于提高训练过程中决策模型的强度。

[0168] 在一种可能的实施方式中,第i对战状态数据包含基础数据,角色增益数据以及元件数据,将第i对战状态数据输入决策模型,得到动作指令。下面对动作指令的输出过程进行介绍。请参考图9,其示出了本申请一个示例性实施例提供的决策模型基于对战状态数据输出动作指令过程的流程图。

[0169] 步骤901,将基础数据输入第n决策模型的第一全连接网络。

[0170] 其中,基础数据是指构建游戏对局所需要的数据,例如游戏环境数据、虚拟角色的相关数据、虚拟道具的相关数据等。

[0171] 可选地,基础数据包括环境数值,第一虚拟角色以及第二虚拟角色的技能ID、第一虚拟角色以及第二虚拟角色使用的通灵ID、第一虚拟角色以及第二虚拟角色使用的秘卷ID、第一虚拟角色以及第二虚拟角色的角色ID。

[0172] 示例性的,如图10所示,服务器将第一虚拟角色以及第二虚拟角色的技能ID、通灵ID、秘卷ID以及角色ID输入嵌入层(Embedding)1005,进行嵌入处理,将ID数据转化为数值向量,并将技能ID、通灵ID、秘卷ID以及角色ID的嵌入结果进行拼接处理形成第一全连接层1006,同时服务器将环境数值输入第一全连接网络1001的第二全连接层1007。进一步,拼接第一全连接层1006的全连接处理结果以及第二全连接层1007的全连接处理结果,得到第一全连接网络1001的全连接处理结果。

[0173] 可选地,拼接的方法可以是叠加(addition)或者融合(concatenate),本申请实施例对此不作限定。

[0174] 步骤902,将角色增益数据输入第n决策模型的第二全连接网络。

[0175] 可选地,角色增益包括角色增益ID。服务器获取第一虚拟角色以及第二虚拟角色的角色增益ID列表,其中,不同的角色增益对应不同的角色增益ID。

[0176] 示例性的,如图10所示,服务器将角色增益ID输入嵌入层1008,将ID数据转化为数值向量。进一步,区分己方(第一虚拟角色)角色增益数值向量以及敌方角色增益数值向量(第二虚拟角色),将己方角色增益数值向量以及敌方角色增益数值向量分别输入卷积层1009进行卷积处理,再将卷积层1009处理得到的结果分别输入降维层1010,进行降维处理,得到己方角色增益数值向量的降维处理结果以及敌方角色增益数值向量的降维处理结果。将己方角色增益数值向量的降维处理结果以及敌方角色增益数值向量的降维处理结果输入第二全连接网络1002的第四全连接层1011,得到第二全连接网络1002的全连接处理结

果。

[0177] 步骤903,将元件数据输入第n决策模型的第三全连接网络。

[0178] 可选地,元件数据包括元件ID以及元件数值。其中,元件用于表征技能的表现形式。例如技能为发送冲击波,该冲击波在游戏画面中表现为波形的形式,该波形即为元件。元件数值用于表示元件的状态向量。可选地,可以是元件位置、速度等。例如,冲击波的位置、发射的速度等。

[0179] 示例性的,如图10所示,服务器获取元件ID列表,其中元件ID列表中包括元件ID以及对应的元件数值。

[0180] 服务器将元件ID输入嵌入层1012,将ID数据转化为数值向量,进一步,区分己方(第一虚拟角色)元件数值向量以及敌方(第二虚拟角色)元件数值向量。同时,将元件数值区分己方元件数值和敌方元件数值。将前述己方元件数值向量和己方元件数值、敌方元件数值向量以及敌方元件数值分别输入卷积层1013进行卷积处理,再将卷积层1013处理得到的结果分别输入降维层1014,进行降维处理,得到己方元件数据的降维处理结果以及敌方元件数据的降维处理结果。将己方元件数据的降维处理结果以及敌方元件数据的降维处理结果输入第三全连接网络1003的第五全连接层1015,得到第三全连接网络1003的全连接处理结果。

[0181] 步骤904,拼接第一全连接网络、第二全连接网络以及第三全连接网络的全连接处理结果,并输入第n决策模型的第四全连接网络,得到第四全连接网络输出的动作采样概率。

[0182] 由于第一全连接网络、第二全连接网络以及第三全连接网络的全连接处理结果不同,因此服务器首先对其进行拼接处理。

[0183] 示例性的,如图10所示,服务器先对第二全连接网络1002以及第三全连接网络1003的全连接处理结果通过第三全连接层1016进行拼接处理,得到第三全连接层1016对应的全连接处理结果,进一步,再与第一全连接网络1001的全连接处理结果进行拼接,将拼接处理后的结果输入决策模型的第四全连接网络1004,进而得到动作采样概率。

[0184] 其中,动作采样概率用于表征在当前对战状态下,第一虚拟角色可能执行的动作的概率。

[0185] 可选地,拼接的方法可以是叠加(addition)或者融合(concatenate),本申请实施例对此不作限定。

[0186] 为了避免客户端返回无效动作,从而造成计算资源的浪费,示例性的,如图10所示,服务器对动作采样概率进行概率屏蔽处理。服务器基于对战状态数据确定动作屏蔽信息,通过动作屏蔽信息对动作采样概率进行概率屏蔽处理。下面对该方法进行介绍。

[0187] 步骤905,基于第i对战状态数据确定动作屏蔽信息,动作屏蔽信息用于指示在当前对战状态下第一虚拟角色无法执行的动作。

[0188] 在一种可能的实施方式中,服务器根据对战状态数据中技能的冷却状态、可使用的技能、技能的类型、技能等级等确定动作屏蔽信息。

[0189] 示例性的,当前对战状态下,当某一技能还在冷却中,或者使用某一技能的能量不足时,第一虚拟角色无法使用该技能,此时动作屏蔽信息为第一虚拟角色无法使用的技能。

[0190] 示例性的,当前对战状态下,第一虚拟角色受到第二虚拟角色的攻击,无法进行上

下、左右的移动,此时动作屏蔽信息为第一虚拟角色无法执行的移动方向。

[0191] 步骤906,基于动作屏蔽信息对动作采样概率进行概率屏蔽处理,概率屏蔽处理用于屏蔽第一虚拟角色无法执行的动作的采样概率。

[0192] 为了避免客户端返回无效动作,使得虚拟角色无法执行动作而造成计算资源的浪费,服务器根据动作屏蔽信息对动作采样概率进行概率屏蔽处理,即屏蔽第一虚拟性角色无法执行的动作的采样概率。

[0193] 示例性的,技能1的采样概率为50%,技能2的采样概率为10%,技能3的采样概率为10%,技能4的采样概率为30%。服务器根据对战状态数据可知,技能3正在冷却中,因此确定的动作屏蔽信息为技能3,进而对技能3的采样概率进行概率屏蔽处理,即决策模型不输出技能3的动作指令。

[0194] 步骤907,基于概率屏蔽处理后的动作采样概率进行动作采样,得到动作指令。

[0195] 在一种可能的实施方式中,服务器从概率屏蔽处理后的动作采样概率中采样概率最高的动作,得到动作指令。

[0196] 示例性的,前述基于概率屏蔽处理后,技能1的采样概率最高,因此决策模型输出技能1的动作指令,指示第一客户端控制第一虚拟角色执行技能1。

[0197] 在本申请实施例中,通过动作屏蔽信息对决策模型输出得动作采样概率进行概率屏蔽处理,避免向客户端返回无效动作,使得虚拟角色无法执行动作从而造成计算资源的浪费。

[0198] 综上所述,在本申请实施例中,每轮迭代中,将第i对战状态数据输入第n决策模型,得到第n决策模型输出的动作指令,该动作指令用于指示第一客户端控制第一虚拟角色执行动作。第一客户端基于第一虚拟角色执行动作后,反馈收集到的第i+1对战状态数据,从而学习建立了从一个状态映射到一个动作的映射关系,进而通过对战过程中的对战状态数据训练第n决策模型,得到第n+1决策模型。通过对战过程中的对战状态数据训练决策模型,提高决策模型的决策能力,使其更加接近真实玩家的表现,进而增强决策模型的对战拟人性。

[0199] 进一步的,服务器基于第一奖励值和第二奖励值训练决策模型,其中第一奖励值通过相邻对战状态数据中的第一虚拟角色和第二虚拟角色的角色属性值变化确定,第二奖励值通过对战胜负结果确定,对战状态数据中的角色属性值变化在一定程度上也可以反映对战胜负结果,因此在本申请实施例中服务器通过第一奖励值和第二奖励值训练决策模型以达到以胜利为导向的目的,从而进一步增强了决策模型的对战拟人性和强度。

[0200] 进一步的,引入第三奖励值训练决策模型,第三奖励值通过合理性参数确定,合理性参数用于衡量在对战状态下决策模型所指示的虚拟对象执行动作指令所表征动作的合理性,合理性参数不同,第三奖励值也不同,进而引导决策模型学习拟人化的基础规则和打法,进一步增强决策模型的对战拟人性。另外,在不同的训练阶段,通过不同的奖励值训练决策模型,以达到不同的训练目的,即在迭代轮数未达到轮数阈值的情况下,服务器基于第一奖励值以及第二奖励值确定的基础奖励值以及第三奖励值确定的辅助奖励值训练决策模型,使得在低轮次的迭代过程中侧重训练决策模型的对战拟人性,而在迭代轮数达到轮数阈值的情况下,服务器基于第一奖励值以及第二奖励值确定的基础奖励值训练决策模型,使得在高轮次的迭代过程中侧重训练决策模型的强度,即以胜利为目标。

[0201] 进一步的,服务器将训练得到的决策模型添加至第二模型池,并基于决策模型的历史胜率从第二模型中采样得到对手决策模型,第二模型池中的决策模型的历史胜率越高,被服务器采样作为对手决策模型的概率越大,通过高胜率的对手决策模型进行训练,有利于提高决策模型的强度,进而使其更接近真实玩家的表现。另外,当第二模型池中的对手决策模型与第一模型池中的决策模型对战次数达到次数阈值时,服务器将重新从第二模型池中采样对手决策模型,以保证决策模型训练的全面性,同时,基于对战结果更新第二模型池中的决策模型的历史胜率,以保证决策模型训练的准确性。

[0202] 进一步的,服务器将对战状态数据中包含的基础数据、角色增益数据、元件数据输入决策模型,输出动作采样概率,为了避免客户端返回无效动作,从而造成计算资源的浪费,服务器通过动作屏蔽信息对输出的动作采样概率进行概率屏蔽处理,进而得到决策模型最终输出的动作指令。

[0203] 上述实施例中以决策模型的训练方法应用于游戏进行说明,在另一种可能的场景下,本申请实施例提供的决策模型的训练方法可应用于其他工业领域中,如智能安防机器人。

[0204] 当应用于智能安防机器人时,训练得到机器人对应的决策模型,从而使得机器人在基于对应的决策模型指示的决策行动时,可基于自身特征进行攻击或防御。

[0205] 在一种可能的实施方式中,对智能安防机器人对应的决策模型进行训练时,构建第一模型池和第二模型池,第一模型池中包含至少一个决策模型,第二模型池中包含一个历史迭代过程中训练得到的决策模型,从第二模型池中选择对手决策模型和第一模型池中的决策模型进行训练。决策模型和对手决策模型输出的动作指令用于指示其对应的机器人执行动作并进行攻击或者防御,基于攻击或者防御过程中的状态数据对决策模型进行多轮训练,提高决策模型的决策能力,使得机器人在不同的情况下基于自身的攻击特性或者防御特性进行安防,从而提高智能安防机器人的安防效果。

[0206] 上述仅以应用于智能安防机器人进行示意性说明,但并不对此构成限定,本申请实施例提供的决策模型的训练方法可应用于任意需进行自动格斗的对象上。

[0207] 下述为本申请装置实施例,可以用于执行本申请方法实施例。对于本申请装置实施例中未披露的细节,请参照本申请方法实施例。

[0208] 请参考图11,其示出了本申请一个示例性实施例提供的决策模型的训练装置的结构框图,该装置包括:

[0209] 获取模块1101,用于获取第一模型池,所述第一模型池中包含至少一个决策模型,所述决策模型用于指示虚拟角色在对战时采用的对战策略;

[0210] 第一输出模块1102,用于在第n轮迭代过程中,将第i对战状态数据输入第n决策模型,得到所述第n决策模型输出的动作指令,所述第n决策模型为上一轮迭代过程训练得到的模型,所述第i对战状态数据用于表征对战过程中第i对战时刻时的对战状态,所述动作指令用于指示第一客户端控制第一虚拟角色执行动作,并反馈所述第一虚拟角色执行动作后的收集到的第i+1对战状态数据;

[0211] 训练模块1103,用于基于对战过程中的对战状态数据训练所述第n决策模型,得到第n+1决策模型;

[0212] 第一添加模块1104,用于将所述第n+1决策模型添加至所述第一模型池。

[0213] 可选地,所述训练模块1103,包括:

[0214] 第一确定单元,用于基于相邻对战时刻对应的所述对战状态数据,确定第一奖励值;

[0215] 第二确定单元,用于基于对战结果确定第二奖励值,所述对战结果用于指示对战的胜负结果;

[0216] 训练单元,用于基于所述第一奖励值和所述第二奖励值训练所述第n决策模型,得到所述第n+1决策模型。

[0217] 可选地,所述第一确定单元,还用于:

[0218] 获取所述第i对战状态数据中的第i角色属性值,以及所述第i+1对战状态数据中的第i+1角色属性值,其中,角色属性值包括所述第一虚拟角色以及与第二虚拟角色的属性值,所述第二虚拟角色是与所述第一虚拟角色对战的虚拟角色;

[0219] 基于所述第i角色属性值与所述第i+1角色属性值的属性值变化情况,确定所述第一奖励值。

[0220] 可选地,所述训练单元,还用于:

[0221] 在迭代轮数达到轮数阈值的情况下,基于所述第一奖励值和所述第二奖励值训练所述第n决策模型,得到所述第n+1决策模型。

[0222] 可选地,所述训练模块1103,还包括:

[0223] 第三确定单元,用于在所述迭代轮数未达到所述轮数阈值的情况下,基于所述对战状态数据以及所述动作指令,确定第三奖励值;

[0224] 所述训练单元,还用于基于所述第一奖励值、所述第二奖励值以及所述第三奖励值训练所述第n决策模型,得到所述第n+1决策模型。

[0225] 可选地,所述第三确定单元,用于:

[0226] 在所述迭代轮数未达到所述轮数阈值的情况下,基于所述对战状态数据以及动作指令确定合理性参数,所述合理性参数用于表征在所述对战状态数据所表征对战状态下执行所述动作指令所表征动作的合理性;

[0227] 基于所述合理性参数确定所述第三奖励值,所述第三奖励值与所述合理性参数呈正相关关系。

[0228] 可选地,所述训练单元,还用于:

[0229] 基于所述第一奖励值、所述第二奖励值以及基础奖励权重,确定基础奖励值;

[0230] 基于所述第三奖励值和辅助奖励权重,确定辅助奖励值,所述辅助奖励权重小于所述基础奖励权重,且所述辅助奖励权重与所述迭代轮数呈负相关关系;

[0231] 基于基础奖励值和所述辅助奖励值训练所述第n决策模型,得到所述第n+1决策模型。

[0232] 可选地,所述装置还包括:

[0233] 采样模块,用于从第二模型池中采样得到对手决策模型,所述第二模型池中包含历史迭代过程中训练得到的至少一个所述决策模型;

[0234] 第二输出模块,用于在第n轮迭代过程中,将第i对手对战状态数据输入所述对手决策模型,得到所述对手决策模型输出的对手动作指令,所述对手动作指令用于指示第二客户端控制第二虚拟角色执行动作,并反馈所述第二虚拟角色执行动作后的收集到的第i+

1对手对战状态数据,所述第二虚拟角色是与所述第一虚拟角色对战的虚拟角色。

[0235] 可选地,所述采样模块,用于:

[0236] 基于所述第二模型池中各个决策模型对应的历史胜率,从所述第二模型池中采样得到所述对手决策模型,其中,所述决策模型的采样率与所述历史胜率呈正相关关系;

[0237] 所述装置还包括:

[0238] 更新模块,用于基于所述第一虚拟角色与所述第二虚拟角色的对战结果更新所述对手决策模型的所述历史胜率。

[0239] 可选地,所述采样模块,还用于:

[0240] 在采用所述对手决策模型进行对战的对战次数达到次数阈值的情况下,重新从所述第二模型池中采样得到所述对手决策模型。

[0241] 可选地,所述装置还包括:

[0242] 第二添加模块,用于将所述第n+1决策模型添加至所述第二模型池。

[0243] 可选地,所述第i对战状态数据包含基础数据、角色增益数据以及元件数据;

[0244] 所述第一输出模块1102,包括:

[0245] 第一输入单元,用于将所述基础数据输入所述第n决策模型的第一全连接网络;

[0246] 第二输入单元,用于将所述角色增益数据输入所述第n决策模型的第二全连接网络;

[0247] 第三输入单元,用于将所述元件数据输入所述第n决策模型的第三全连接网络;

[0248] 拼接单元,用于拼接所述第一全连接网络、所述第二全连接网络以及所述第三全连接网络的全连接处理结果,并输入所述第n决策模型的第四全连接网络,得到所述第四全连接网络输出的动作采样概率;

[0249] 采样单元,用于基于所述动作采样概率进行动作采样,得到所述动作指令。

[0250] 可选地,所述装置还包括:

[0251] 第一确定模块,用于基于所述第i对战状态数据确定动作屏蔽信息,所述动作屏蔽信息用于指示在当前对战状态下所述第一虚拟角色无法执行的动作;

[0252] 屏蔽模块,用于基于所述动作屏蔽信息对所述动作采样概率进行概率屏蔽处理,所述概率屏蔽处理用于屏蔽所述第一虚拟角色无法执行的动作的采样概率;

[0253] 所述采样单元,还用于:

[0254] 基于概率屏蔽处理后的所述动作采样概率进行动作采样,得到所述动作指令。

[0255] 可选地,所述装置还包括:

[0256] 第二确定模块,用于在满足迭代训练结束条件的情况下,将所述第一模型池中最后一轮训练得到的决策模型确定为目标决策模型。

[0257] 综上所述,在本申请实施例中,每轮迭代中,将第i对战状态数据输入第n决策模型,得到第n决策模型输出的动作指令,该动作指令用于指示第一客户端控制第一虚拟角色执行动作。第一客户端基于第一虚拟角色执行动作后,反馈收集到的第i+1对战状态数据,从而学习建立了从一个状态映射到一个动作的映射关系,进而通过对战过程中的对战状态数据训练第n决策模型,得到第n+1决策模型。通过对战过程中的对战状态数据训练决策模型,提高决策模型的决策能力,使其更加接近真实玩家的表现,进而增强决策模型的对战拟人性。

[0258] 需要说明的是:上述实施例提供的装置,仅以上述各功能模块的划分进行举例说明,实际应用中,可以根据需要而将上述功能分配由不同的功能模块完成,即将装置的内部结构划分成不同的功能模块,以完成以上描述的全部或者部分功能。另外,上述实施例提供的装置与方法实施例属于同一构思,其具体实现过程详见方法实施例,这里不再赘述。

[0259] 请参考图12,其示出了本申请一个示例性实施例提供的计算机设备的结构框图,该计算机设备可用于训练上述决策模型。

[0260] 所述计算机设备1200包括中央处理单元(Central Processing Unit,CPU)1201、包括随机存取存储器(Random Access Memory,RAM)1202和只读存储器(Read-Only Memory,ROM)1203的系统存储器1204,以及连接系统存储器1204和中央处理单元1201的系统总线1205。所述计算机设备1200还包括帮助服务器内的各个器件之间传输信息的基本输入/输出系统(Input/Output系统,I/O系统)1206,和用于存储操作系统1213、应用程序1214和其他程序模块1215的大容量存储设备1207。

[0261] 所述基本输入/输出系统1206包括有用于显示信息的显示器1208和用于用户输入信息的诸如鼠标、键盘之类的输入设备1209。其中所述显示器1208和输入设备1209都通过连接到系统总线1205的输入输出控制器1210连接到中央处理单元1201。所述基本输入/输出系统1206还可以包括输入输出控制器1210以用于接收和处理来自键盘、鼠标、或电子触控笔等多个其他设备的输入。类似地,输入输出控制器1210还提供输出到显示屏、打印机或其他类型的输出设备。

[0262] 所述大容量存储设备1207通过连接到系统总线1205的大容量存储控制器(未示出)连接到中央处理单元1201。所述大容量存储设备1207及其相关联的计算机可读存储介质为计算机设备1200提供非易失性存储。也就是说,所述大容量存储设备1207可以包括诸如硬盘或者只读光盘(Compact Disc Read-Only Memory,CD-ROM)驱动器之类的计算机可读存储介质(未示出)。

[0263] 不失一般性,所述计算机可读存储介质可以包括计算机存储介质和通信介质。计算机存储介质包括以用于存储诸如计算机可读存储指令、数据结构、程序模块或其他数据等信息的任何方法或技术实现的易失性和非易失性、可移动和不可移动介质。计算机存储介质包括RAM、ROM、可擦除可编程只读寄存器(Erasable Programmable Read Only Memory,EPROM)、电子抹除式可复写只读存储器(Electrically-Erasable Programmable Read-Only Memory,EEPROM)、闪存或其他固态存储其技术,CD-ROM、数字多功能光盘(Digital Versatile Disc,DVD)或其他光学存储、磁带盒、磁带、磁盘存储或其他磁性存储设备。当然,本领域技术人员可知所述计算机存储介质不局限于上述几种。上述的系统存储器1204和大容量存储设备1207可以统称为存储器。

[0264] 存储器存储有一个或多个程序,一个或多个程序被配置成由一个或多个中央处理单元1201执行,一个或多个程序包含用于实现上述方法实施例的指令,中央处理单元1201执行该一个或多个程序实现上述各个方法实施例提供的方法。

[0265] 根据本申请的各种实施例,所述计算机设备1200还可以通过诸如因特网等网络连接到网络上的远程服务器运行。也即计算机设备1200可以通过连接在所述系统总线1205上的网络接口单元1211连接到网络1212,或者说,也可以使用网络接口单元1211来连接到其他类型的网络或远程服务器系统(未示出)。

[0266] 所述存储器还包括一个或者一个以上的程序,所述一个或者一个以上程序存储于存储器中,所述一个或者一个以上程序包含用于进行本申请实施例提供的方法中由计算机设备所执行的步骤。

[0267] 本申请实施例还提供了一种计算机可读存储介质,该计算机可读存储介质存储有至少一条指令,所述至少一条指令由所述处理器加载并执行以实现如上各个实施例所述的决策模型的训练方法。

[0268] 根据本申请的一个方面,提供了一种计算机程序产品或计算机程序,该计算机程序产品或计算机程序包括计算机指令,该计算机指令存储在计算机可读存储介质中。计算机设备的处理器从计算机可读存储介质读取该计算机指令,处理器执行该计算机指令,使得该计算机设备执行上述方面的各种可选实现方式中提供的决策模型的训练方法。

[0269] 本领域技术人员应该可以意识到,在上述一个或多个示例中,本申请实施例所描述的功能可以用硬件、软件、固件或它们的任意组合来实现。当使用软件实现时,可以将这些功能存储在计算机可读存储介质中或者作为计算机可读存储介质上的一个或多个指令或代码进行传输。计算机可读存储介质包括计算机存储介质和通信介质,其中通信介质包括便于从一个地方向另一个地方传送计算机程序的任何介质。存储介质可以是通用或专用计算机能够存取的任何可用介质。

[0270] 以上所述仅为本申请的可选实施例,并不用以限制本申请,凡在本申请的精神和原则之内,所作的任何修改、等同替换、改进等,均应包含在本申请的保护范围之内。

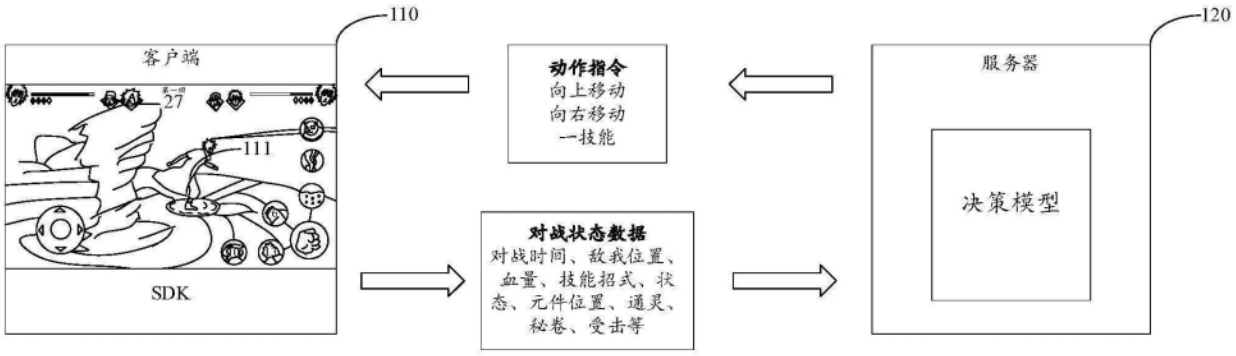

110
客户端
27
111
SDK
动作指令
向上移动
向右移动
一技能
对战状态数据
对战时间、敌我位置、血量、技能招式、状态、元件位置、通灵、秘卷、受击等
120
服务器
决策模型

图1

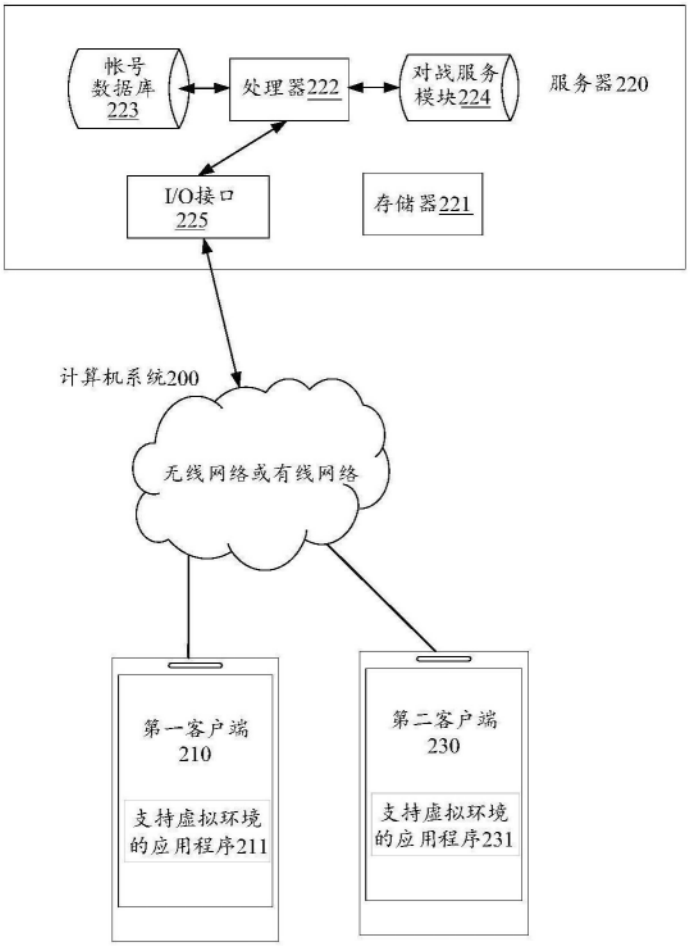

帐号数据库223
处理器222
对战服务模块224
服务器220
I/O接口225
存储器221
计算机系统200
无线网络或有线网络
第一客户端210
支持虚拟环境的应用程序211
第二客户端230
支持虚拟环境的应用程序231

图2

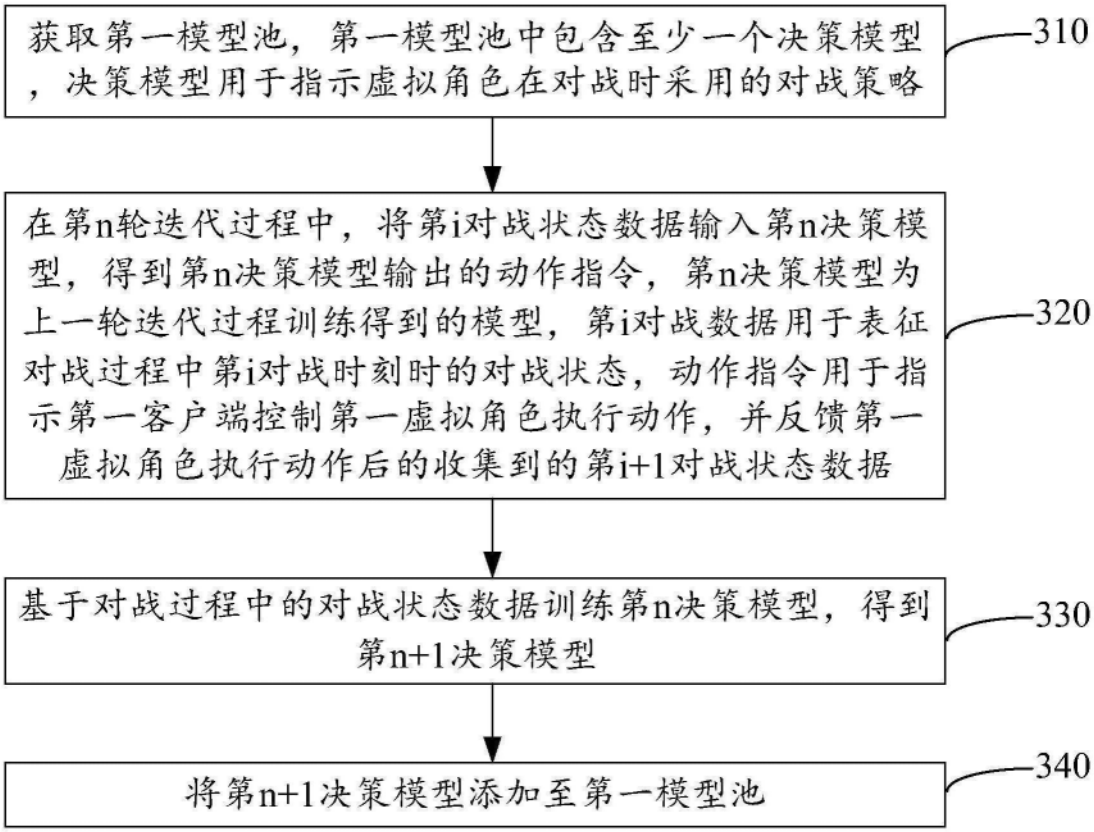

获取第一模型池，第一模型池中包含至少一个决策模型，决策模型用于指示虚拟角色在对战时采用的对战策略
310
在第n轮迭代过程中，将第i对战状态数据输入第n决策模型，得到第n决策模型输出的动作指令，第n决策模型为上一轮迭代过程训练得到的模型，第i对战数据用于表征对战过程中第i对战时刻时的对战状态，动作指令用于指示第一客户端控制第一虚拟角色执行动作，并反馈第一虚拟角色执行动作后的收集到的第i+1对战状态数据
320
基于对战过程中的对战状态数据训练第n决策模型，得到第n+1决策模型
330
将第n+1决策模型添加至第一模型池
340

图3

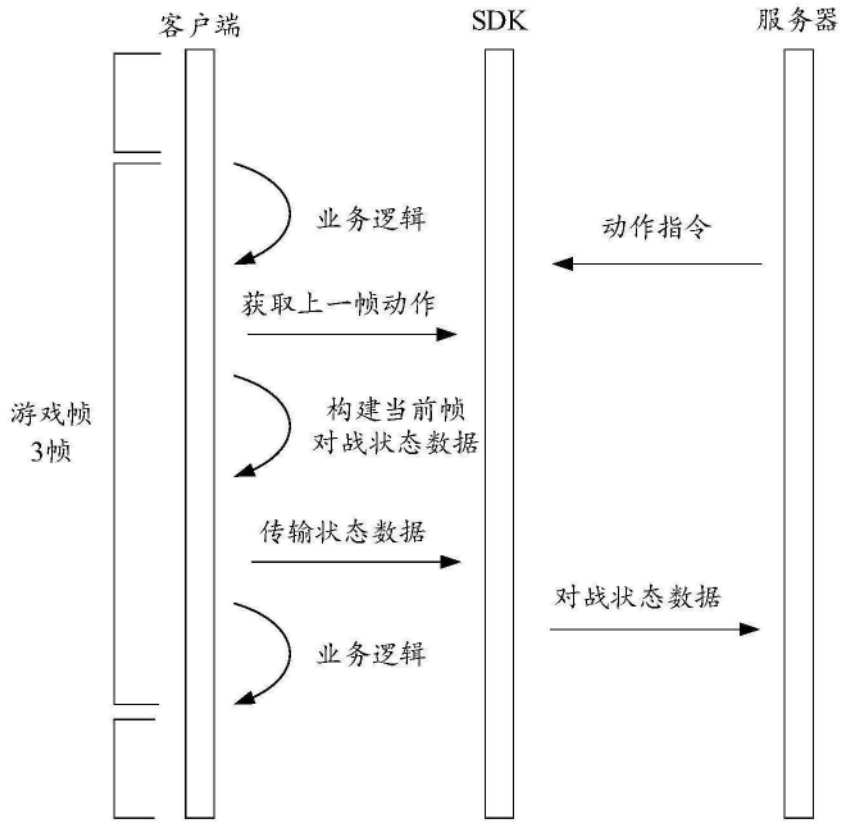

客户端
SDK
服务器
业务逻辑
动作指令
获取上一帧动作
游戏帧
3帧
构建当前帧
对战状态数据
传输状态数据
对战状态数据
业务逻辑

图4

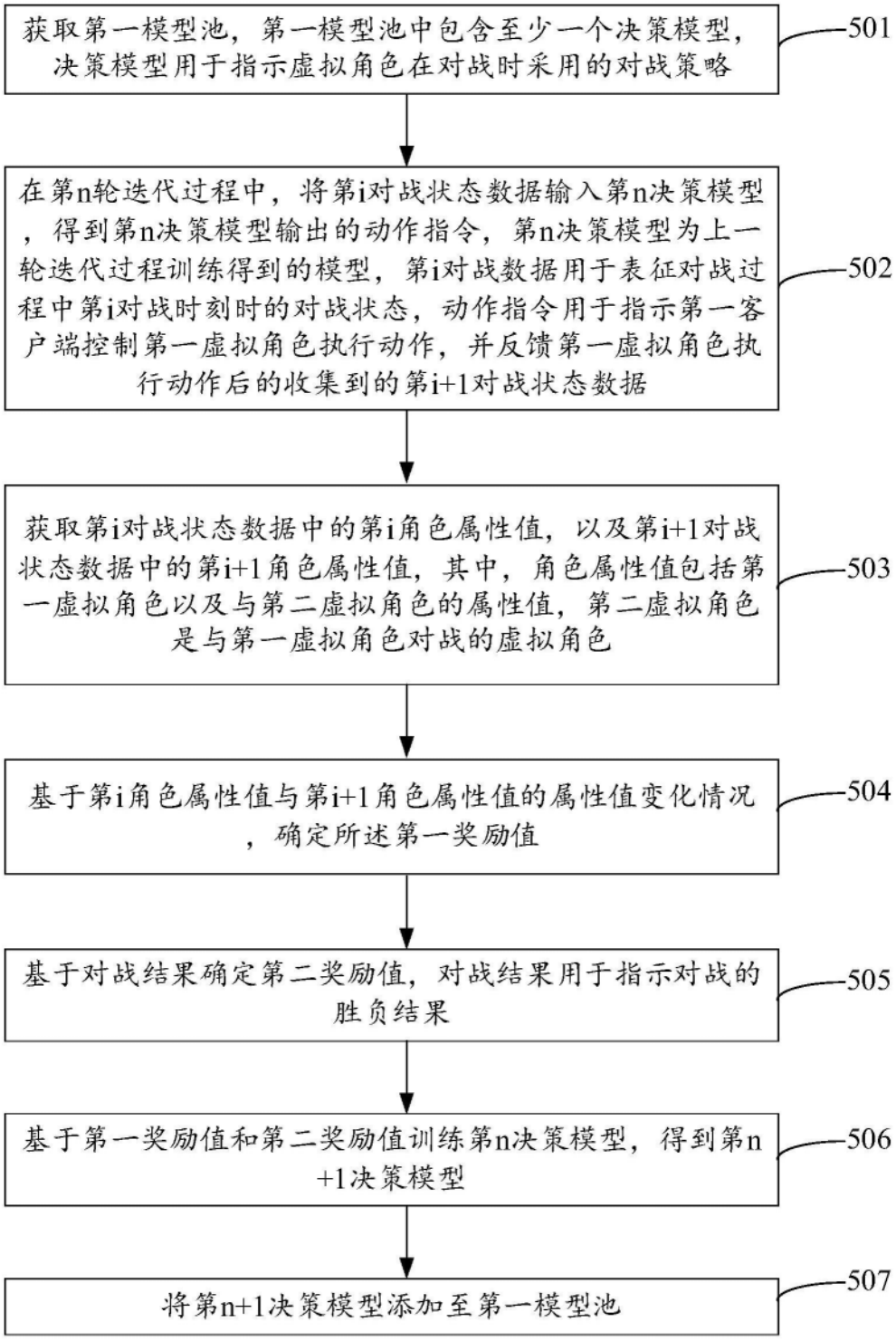

获取第一模型池，第一模型池中包含至少一个决策模型，决策模型用于指示虚拟角色在对战时采用的对战策略
501
在第n轮迭代过程中，将第i对战状态数据输入第n决策模型，得到第n决策模型输出的动作指令，第n决策模型为上一轮迭代过程训练得到的模型，第i对战数据用于表征对战过程中第i对战时刻时的对战状态，动作指令用于指示第一客户端控制第一虚拟角色执行动作，并反馈第一虚拟角色执行动作后的收集到的第i+1对战状态数据
502
获取第i对战状态数据中的第i角色属性值，以及第i+1对战状态数据中的第i+1角色属性值，其中，角色属性值包括第一虚拟角色以及与第二虚拟角色的属性值，第二虚拟角色是与第一虚拟角色对战的虚拟角色
503
基于第i角色属性值与第i+1角色属性值的属性值变化情况，确定所述第一奖励值
504
基于对战结果确定第二奖励值，对战结果用于指示对战的胜负结果
505
基于第一奖励值和第二奖励值训练第n决策模型，得到第n+1决策模型
506
将第n+1决策模型添加至第一模型池
507

图5

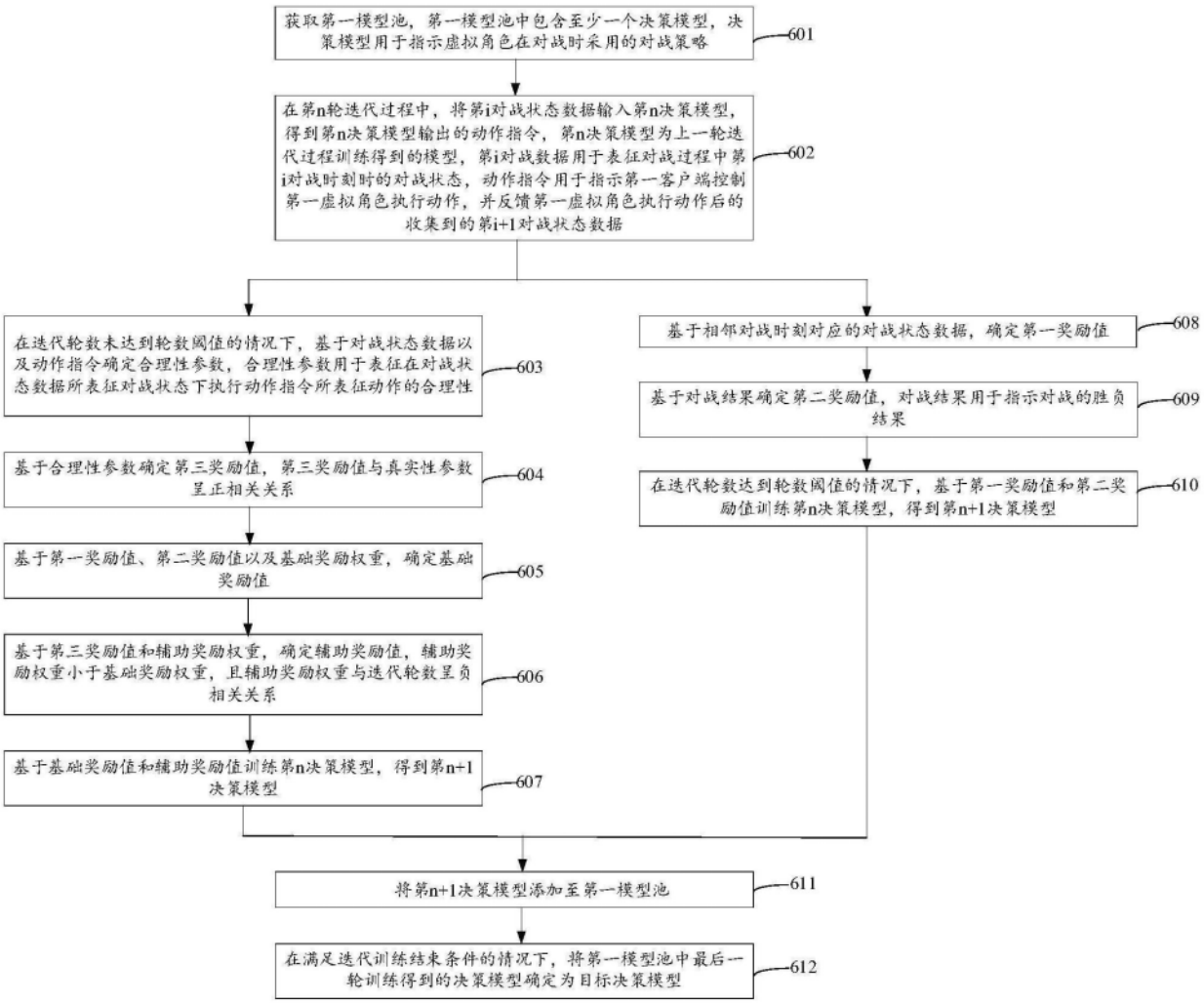

获取第一模型池，第一模型池中包含至少一个决策模型，决策模型用于指示虚拟角色在对战时采用的对战策略
601
在第n轮迭代过程中，将第i对战状态数据输入第n决策模型，得到第n决策模型输出的动作指令，第n决策模型为上一轮迭代过程训练得到的模型，第i对战数据用于表征对战过程中第i对战时刻时的对战状态，动作指令用于指示第一客户端控制第一虚拟角色执行动作，并反馈第一虚拟角色执行动作后的收集到的第i+1对战状态数据
602
在迭代轮数未达到轮数阈值的情况下，基于对战状态数据以及动作指令确定合理性参数，合理性参数用于表征在对战状态数据所表征对战状态下执行动作指令所表征动作的合理性
603
基于合理性参数确定第三奖励值，第三奖励值与真实性参数呈正相关关系
604
基于第一奖励值、第二奖励值以及基础奖励权重，确定基础奖励值
605
基于第三奖励值和辅助奖励权重，确定辅助奖励值，辅助奖励权重小于基础奖励权重，且辅助奖励权重与迭代轮数呈负相关关系
606
基于基础奖励值和辅助奖励值训练第n决策模型，得到第n+1决策模型
607
基于相邻对战时刻对应的对战状态数据，确定第一奖励值
608
基于对战结果确定第二奖励值，对战结果用于指示对战的胜负结果
609
在迭代轮数达到轮数阈值的情况下，基于第一奖励值和第二奖励值训练第n决策模型，得到第n+1决策模型
610
将第n+1决策模型添加至第一模型池
611
在满足迭代训练结束条件的情况下，将第一模型池中最后一轮训练得到的决策模型确定为目标决策模型
612

图6

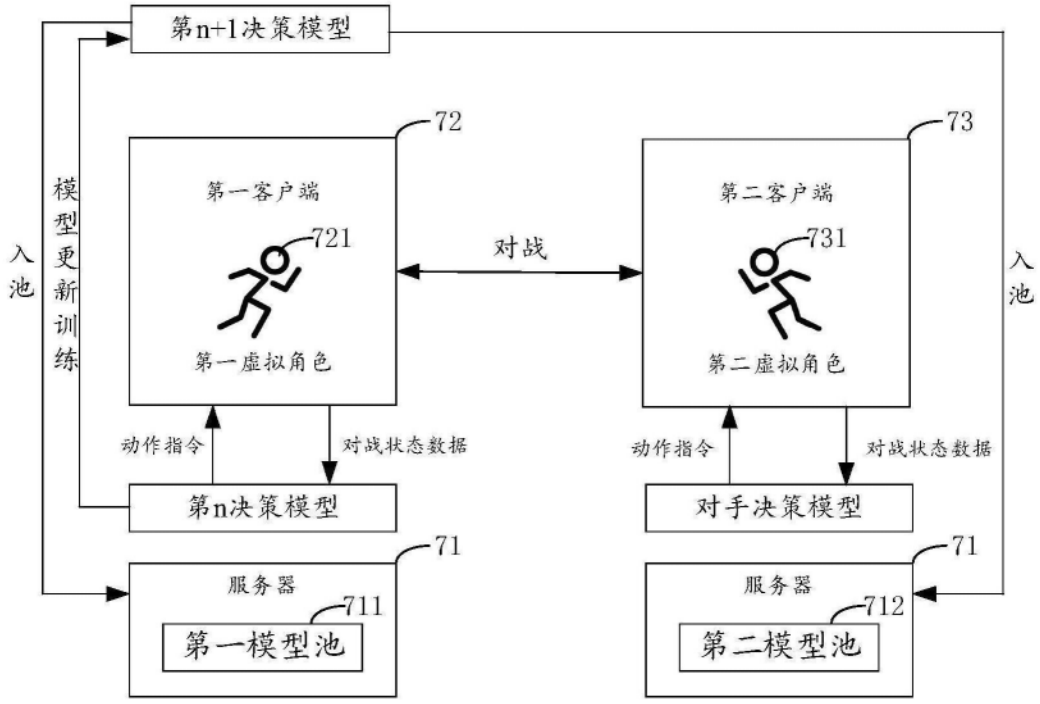

第n+1决策模型
72
第一客户端
721
第一虚拟角色
对战
73
第二客户端
731
第二虚拟角色
入池
模型更新训练
入池
动作指令
对战状态数据
第n决策模型
动作指令
对战状态数据
对手决策模型
71
服务器
711
第一模型池
71
服务器
712
第二模型池

图7

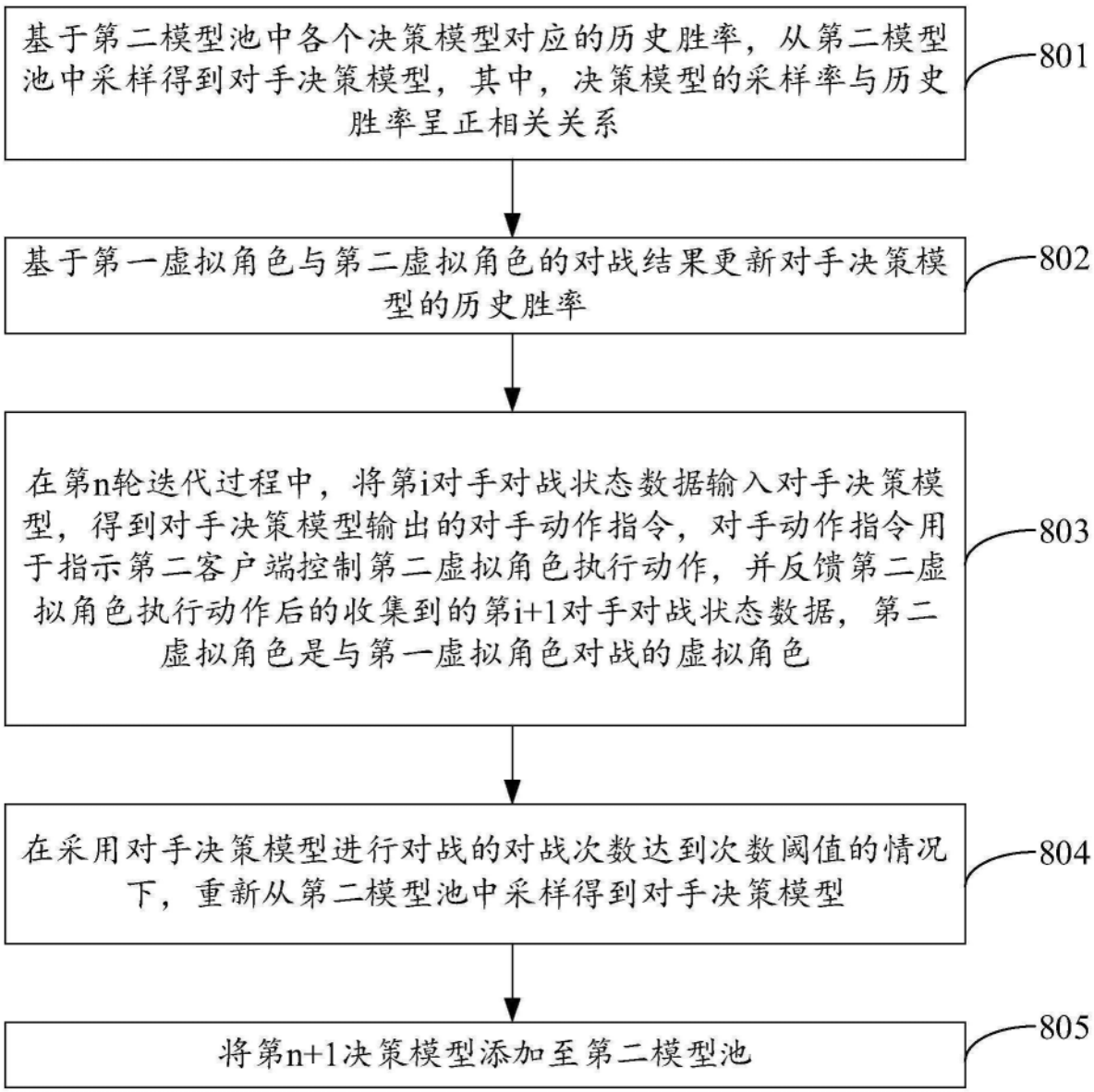

基于第二模型池中各个决策模型对应的历史胜率，从第二模型池中采样得到对手决策模型，其中，决策模型的采样率与历史胜率呈正相关关系
801
基于第一虚拟角色与第二虚拟角色的对战结果更新对手决策模型的历史胜率
802
在第n轮迭代过程中，将第i对手对战状态数据输入对手决策模型，得到对手决策模型输出的对手动作指令，对手动作指令用于指示第二客户端控制第二虚拟角色执行动作，并反馈第二虚拟角色执行动作后的收集到的第i+1对手对战状态数据，第二虚拟角色是与第一虚拟角色对战的虚拟角色
803
在采用对手决策模型进行对战的对战次数达到次数阈值的情况下，重新从第二模型池中采样得到对手决策模型
804
将第n+1决策模型添加至第二模型池
805

图8

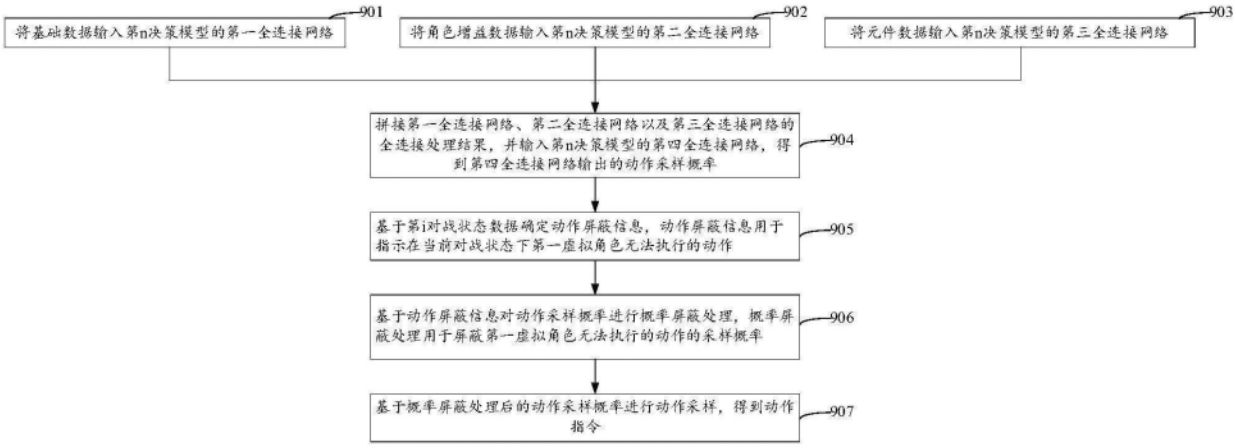

901
将基础数据输入第n决策模型的第一全连接网络
902
将角色增益数据输入第n决策模型的第二全连接网络
903
将元件数据输入第n决策模型的第三全连接网络
拼接第一全连接网络、第二全连接网络以及第三全连接网络的全连接处理结果，并输入第n决策模型的第四全连接网络，得到第四全连接网络输出的动作采样概率
904
基于第j对战状态数据确定动作屏蔽信息，动作屏蔽信息用于指示在当前对战状态下第一虚拟角色无法执行的动作
905
基于动作屏蔽信息对动作采样概率进行概率屏蔽处理，概率屏蔽处理用于屏蔽第一虚拟角色无法执行的动作的采样概率
906
基于概率屏蔽处理后的动作采样概率进行动作采样，得到动作指令
907

图9

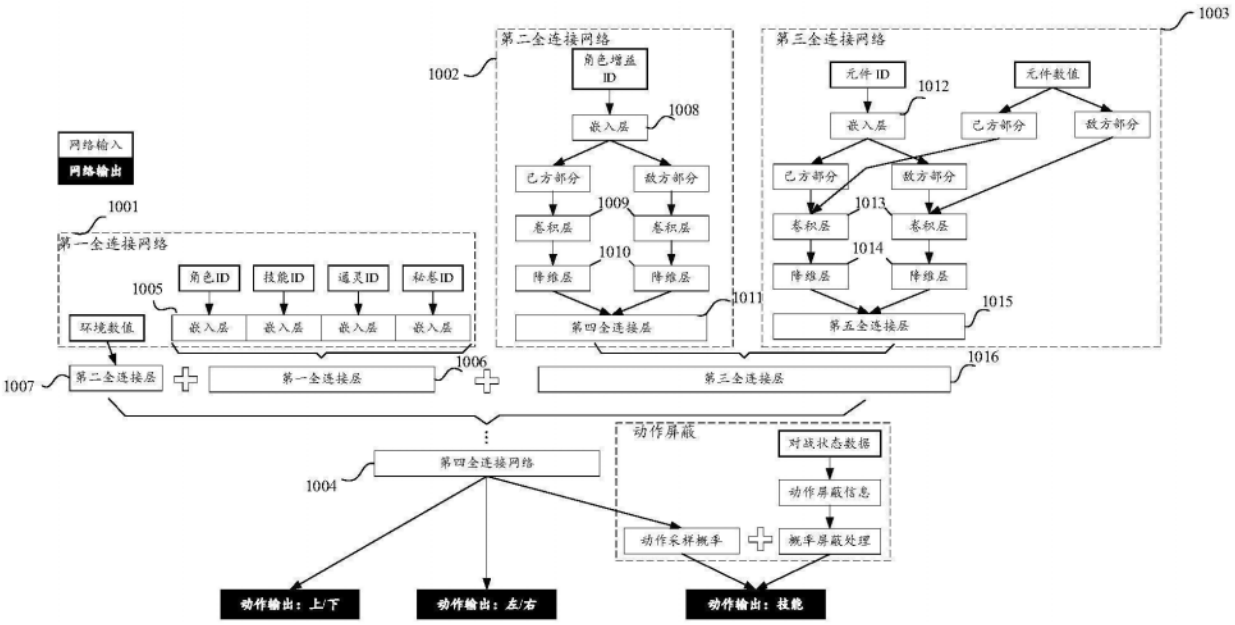

网络输入
网络输出
1001
第一全连接网络
角色ID
技能ID
通灵ID
秘卷ID
1005
环境数值
嵌入层
嵌入层
嵌入层
嵌入层
1007
第二全连接层
1006
第一全连接层
1002
第二全连接网络
角色增益ID
1008
嵌入层
己方部分
敌方部分
卷积层
卷积层
1009
1010
降维层
降维层
第四全连接层
1011
1003
第三全连接网络
元件ID
元件数值
1012
嵌入层
己方部分
敌方部分
己方部分
敌方部分
卷积层
卷积层
1013
1014
降维层
降维层
第五全连接层
1015
1016
第三全连接层
第四全连接网络
1004
动作屏蔽
对战状态数据
动作屏蔽信息
动作采样概率
概率屏蔽处理
动作输出：上/下
动作输出：左/右
动作输出：技能

图10

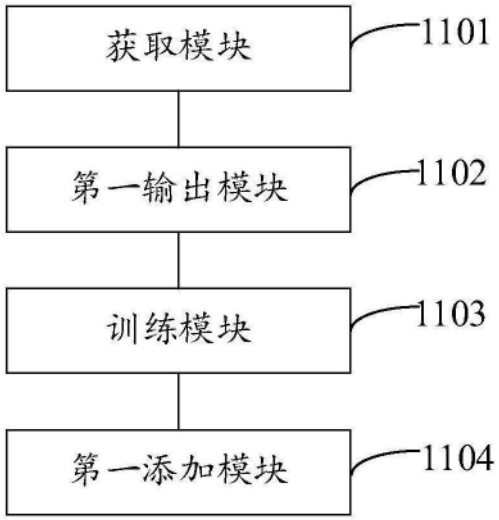

获取模块
1101
第一输出模块
1102
训练模块
1103
第一添加模块
1104

图11

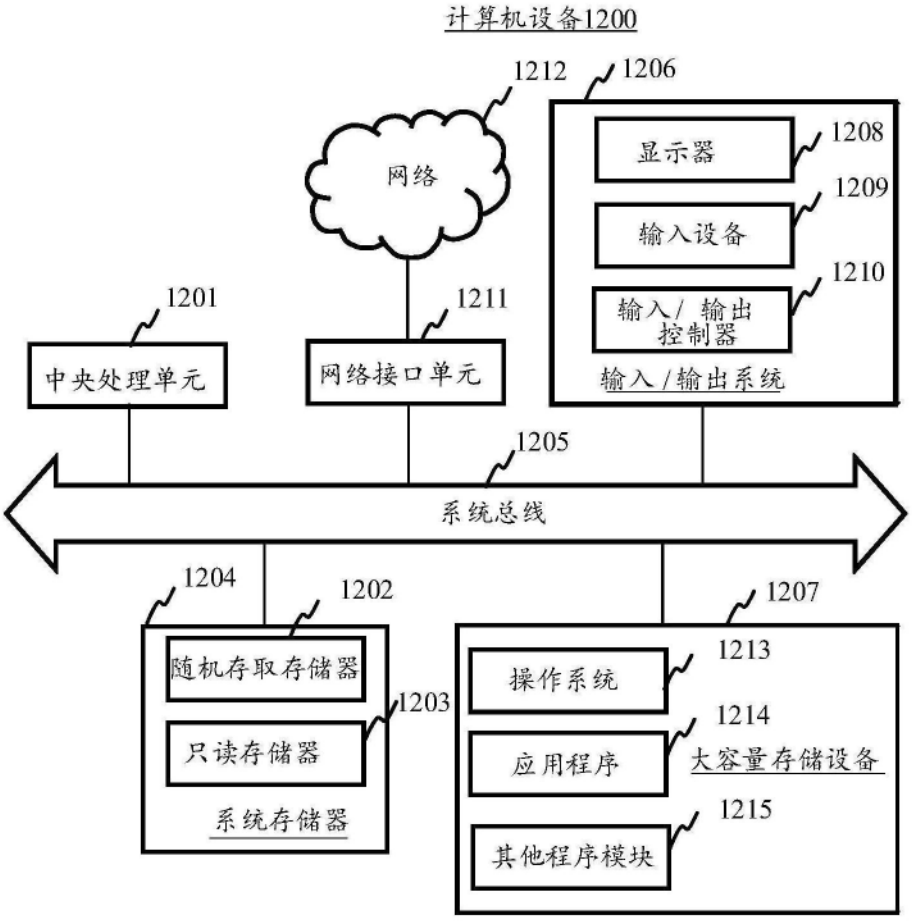

计算机设备1200
1212
1206
网络
显示器
1208
1209
输入设备
1210
输入/ 输出
控制器
输入/输出系统
1201
1211
中央处理单元
网络接口单元
1205
系统总线
1204
1202
1207
随机存取存储器
1213
操作系统
1203
只读存储器
1214
应用程序
大容量存储设备
系统存储器
1215
其他程序模块

图12