

说明书

发明名称: 一种基于强化学习的NPC对战决策方法及相关产品

[0001] 本申请要求于2023年09月26日提交中国专利局、申请号为2023112509261、申请名称为“一种基于强化学习的NPC对战决策方法及相关产品”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

[0002] 本申请涉及人工智能技术领域，尤其涉及一种基于强化学习的NPC对战决策方法及相关产品。

背景技术

[0003] 在科技不断发展的今天，人们的消遣方式也逐渐增多，游戏是大多数人会选择消遣方式之一。NPC是non-player character的缩写，是游戏中一种角色类型，意思是非玩家角色，指的是游戏中不受真人玩家操纵的游戏角色。NPC可以提供游戏中的各种服务和体验，增强游戏的逼真度和互动性，NPC在游戏中是十分重要的。NPC可以推进游戏的剧情走向，可以为玩家角色发布任务，可以与玩家角色进行交易，还可以帮助玩家角色进行战斗。

[0004] 对于战斗NPC来说，需要面临在游戏世界中各式各样的场景和敌人的，在各种各样的场景下与各种各样的敌人进行战斗。战斗NPC需要在面对不同场景下的不同敌人做出不同的对战决策。例如，在平原的游戏场景中面临十只低级敌人，战斗NPC会做出对战决策A；在火山场景中面临一只精英级别的敌人，战斗NPC会做出对战决策B。相关技术对于战斗NPC的对战决策往往使用行为树的方式确定，行为树就是一棵节点层次分明的树，控制着战斗NPC的一系列对战策略。

[0005] 但是，使用行为树决定在开放式虚拟场景中战斗NPC的对战决策往往需要耗费大量的人力成本，因此，如何降低确定NPC的对战决策使用的人力成本成为了本领域亟需解决的技术问题。

[0006] 发明内容

- [0007] 本申请实施例提供了一种基于强化学习的NPC对战决策方法及相关产品，旨在解决相关技术中耗费大量的人力成本确定NPC对战决策的问题。
- [0008] 本申请第一方面提供了一种基于强化学习的NPC对战决策方法，包括：
- [0009] 获取虚拟场景中目标NPC的当前状态信息，所述目标NPC的对战对象的当前状态信息，所述目标NPC与所述对战对象的当前相对位姿信息，以及所述目标NPC与所述对战对象的历史对战状态信息；所述目标NPC的当前状态信息包括所述目标NPC释放技能的上下文信息和所述目标NPC的技能集合；
- [0010] 基于所述目标NPC的当前状态信息、所述对战对象的当前状态信息以及所述当前相对位姿信息进行特征提取及拼接，得到拼接后的特征；
- [0011] 通过长短期记忆网络对所述拼接后的特征及所述历史对战状态信息进行处理，得到所述长短期记忆网络输出的第一处理结果；
- [0012] 基于所述第一处理结果、所述目标NPC释放技能的上下文信息、所述目标NPC的技能集合和强化学习的奖励项的奖励值，通过强化学习的方式对所述目标NPC下一时刻应采取的对战策略进行决策；其中，所述奖励项的奖励值为基于削弱所述对战对象的战斗能力确定的。
- [0013] 本申请第二方面提供了一种基于强化学习的NPC对战决策装置，包括：
- [0014] 信息获取模块，用于获取虚拟场景中目标NPC的当前状态信息，所述目标NPC的对战对象的当前状态信息，所述目标NPC与所述对战对象的当前相对位姿信息，以及所述目标NPC与所述对战对象的历史对战状态信息；所述目标NPC的当前状态信息包括所述目标NPC释放技能的上下文信息和所述目标NPC的技能集合；
- [0015] 拼接模块，用于基于所述目标NPC的当前状态信息、所述对战对象的当前状态信息以及所述当前相对位姿信息进行特征提取及拼接，得到拼接后的特征；
- [0016] 长短期记忆网络模块，用于通过长短期记忆网络对所述拼接后的特征及所述历史对战状态信息进行处理，得到所述长短期记忆网络输出的第一处理结果；
- [0017] 强化学习模块，用于基于所述第一处理结果、所述目标NPC释放技能的上下文信息、所述目标NPC的技能集合和强化学习的奖励项，通过强化学习的方式对

所述目标NPC下一时刻应采取的对战策略进行决策；其中，所述奖励项的奖励值为基于削弱所述对战对象的战斗能力确定的。

[0018] 本申请第三方面提供了一种基于强化学习的NPC对战决策设备，所述设备包括处理器以及存储器：

[0019] 所述存储器用于存储计算机程序，并将所述计算机程序传输给所述处理器；

[0020] 所述处理器用于根据所述计算机程序中的指令执行第一方面提供的基于强化学习的NPC对战决策方法的步骤。

[0021] 本申请第四方面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用于存储计算机程序，所述计算机程序被基于强化学习的NPC对战决策设备执行时实现第一方面提供的基于强化学习的NPC对战决策方法的步骤。

[0022] 本申请第五方面提供了一种计算机程序产品，包括计算机程序，该计算机程序被基于强化学习的NPC对战决策设备执行时实现第一方面提供的基于强化学习的NPC对战决策方法的步骤。

[0023] 从以上技术方案可以看出，本申请实施例具有以下优点：

[0024] 本申请技术方案中利用获取到的虚拟场景中目标NPC的当前状态信息、对战对象的当前状态信息以及当前相对位姿信息进行特征提取及拼接，得到拼接后的特征。通过长短期记忆网络对拼接后的特征及历史对战状态信息进行处理，得到长短期记忆网络输出的第一处理结果。之后基于第一处理结果、目标NPC释放技能的上下文信息、目标NPC的技能集合和强化学习的奖励项，通过强化学习的方式对目标NPC下一时刻应采取的对战策略进行决策。由于使用了某个时刻的信息得到拼接后的特征，拼接后的特征可以很好的表示某个时刻下的对战状态信息，在长短期记忆网络中使用了这个时刻下的对战状态信息和历史对战状态信息进行处理，得到第一处理结果，第一处理结果是综合这个时刻下的对战状态信息和这个历史某个时刻的对战状态信息得到的结果，第一处理结果将现在的对战状态信息和历史对战状态信息综合起来，有助于更精准的确定目标NPC下一时刻的对战策略。通过基于第一处理结果与相关信息通过强化学习的方式对目标NPC下一时刻应采取的对战策略进行决策，只需要获取相关的信息并进行一定的处理和设置奖励项之后就可以十分简单高效的确定目标NPC在不

同场景中下一时刻的对战决策，与相关技术相比无需建立数量众多、结构复杂的行为树，节约了大量的人力成本，并提高了处理效率和决策精度。

附图说明

- [0025] 图1为一种战斗NPC的行为树示意图；
- [0026] 图2为本申请实施例提供的一种基于强化学习的NPC对战决策方法的场景架构图；
- [0027] 图3为本申请实施例提供的一种基于强化学习的NPC对战决策方法的流程图；
- [0028] 图4为本申请实施例提供的一种基于强化学习的NPC对战决策方法的应用场景示意图；
- [0029] 图5为本申请实施例提供的一种基于强化学习的NPC对战决策方法的结构示意图；
- [0030] 图6为本申请实施例提供的一种基于强化学习的NPC对战决策方法的网络结构图；
- [0031] 图7为本申请实施例提供的一种基于强化学习的NPC对战决策装置的结构示意图；
- [0032] 图8为本申请实施例中服务器的一个结构示意图；
- [0033] 图9为本申请实施例中终端设备的一个结构示意图。

具体实施方式

- [0034] 正如上文提到的，行为树是一棵节点层次分明的树，可以使用行为树控制战斗NPC的一系列对战策略。
- [0035] 图1为一种战斗NPC的行为树示意图，其中局面信息中可以包括战斗的场景和敌人，通过风格选择节点选择战斗NPC的对战风格，例如，风格1可以是偏激进的以伤换伤的对战风格，风格2可以是偏均衡的对战风格，风格3可以是偏保守的优先保证战斗NPC状态的对战风格。选定了风格后，经过自身状态判断节点、对手状态判断节点和距离判定节点后来到策略选择节点，通过策略选择节点选择策略1节点或者策略2节点，策略1节点下又存在等待节点、移动节点和出

招节点、出招节点下可以是战斗NPC可以释放的技能节点，也就是出招选择节点，当出招选择节点选择的技能被执行后，向局面信息节点返回执行结果。

[0036] 图1所示的战斗NPC的行为树示意图中的局面信息指的是特定的某种局面信息。在一种可能的实现方式中，局面信息包括在森林的游戏场景下对战三只低级敌人，那么这个行为树所作出的所有对战决策都是针对于在森林的游戏场景下对战三只低级敌人这一局面下的。如果存在一个在深海的游戏场景下对战五只鱼类敌人，图1所示的战斗NPC的行为树则无法使用，需要建立一个对应深海的游戏场景下对战五只鱼类敌人的行为树来确定在这一场景下战斗NPC的对战策略。

[0037] 在一款游戏中可能会存在数量十分庞大的战斗局面，相关技术需要针对每一种战斗局面使用局面信息建立对应局面信息的行为树，之后利用行为树确定NPC的对战策略。不同NPC对应同一局面信息的行为树也有所不同，在一种可能的实现方式中，在相同的局面信息下，NPC1可能只存在两个风格，但是NPC2可能存在五个风格，这就代表着对应于相同的局面信息，NPC1和NPC2的行为树也是完全不同的。这就导致了利用行为树确定NPC的对战策略需要建立数量十分庞大的行为树，进一步造成了人力成本高的问题。

[0038] 鉴于以上问题，在本申请中提供了一种基于强化学习的NPC对战决策方法及相关产品，目的是使用较低的人工成本确定NPC的对战决策。

[0039] 本申请实施例提供的基于强化学习的NPC对战决策方法的执行主体可以为计算机设备，计算机设备可以包括终端设备或服务器。例如在终端设备上获取虚拟场景中的信息。作为示例，终端设备具体可以包括但不限于手机、台式电脑、平板电脑、笔记本电能、掌上电脑、车载终端和飞行器等。本申请实施例提供的基于强化学习的NPC对战决策方法的执行主体也可以是服务器，即可以在服务器上通过长短期记忆网络对拼接后的特征及历史对战状态信息进行处理，得到所述长短期记忆网络输出的第一处理结果。本申请实施例提供的基于强化学习的NPC对战决策方法也可以由终端设备和服务器协同执行。故本申请实施例中对于执行本申请技术方案的实现主体不做限定。

- [0040] 图2示例性地展示了一种基于强化学习的NPC对战决策方法的场景架构图。图中包括服务器以及多种形式的终端设备。图2所示的服务器可以是独立的物理服务器，也可以是多个物理服务器构成的服务器集群或者分布式系统。另外，服务器还可以是提供云服务、云数据库、云计算、云函数、云存储、网络服务、云通信、中间件服务、域名服务、安全服务、CDN、以及大数据和人工智能平台等基础云计算服务的云服务器。
- [0041] 图3为本申请实施例提供的一种基于强化学习的NPC对战决策方法的流程图。该方法可以由计算机设备执行，在本实施例中，以该计算机设备是终端设备为例进行说明。如图3所示的基于强化学习的NPC对战决策方法中，包括：
- [0042] S301：获取虚拟场景中目标NPC的当前状态信息，目标NPC的对战对象的当前状态信息，目标NPC与对战对象的当前相对位姿信息，以及目标NPC与对战对象的历史对战状态信息。
- [0043] 虚拟场景可以是游戏场景。在一种可能的实现方式中，虚拟场景可以是开放式的虚拟场景，开放式的虚拟场景指的是自由度极高的虚拟场景，例如开放式的游戏世界。
- [0044] 终端设备获取目标NPC的当前状态信息，目标NPC的对战对象的当前状态信息，目标NPC与对战对象的当前相对位姿信息，以及目标NPC与对战对象的历史对战状态信息。目标NPC是虚拟场景中的NPC，在一种可能的实现方式中，目标NPC可以是开放式的虚拟场景中的战斗NPC。
- [0045] 终端设备获取的目标NPC的当前状态信息包括目标NPC释放技能的上下文信息和目标NPC的技能集合。在一种可能的实现方式中，目标NPC正在释放一个需要持续释放的技能A，技能A需要20秒的释放时间，在终端设备获取的目标NPC的当前状态信息的这个时刻，目标NPC已经释放了5秒技能A，目标NPC的当前状态信息包括了已经释放5秒技能A的信息。目标NPC可以使用技能A、技能B、……和技能J这十个技能，终端设备获取这十个技能形成的目标NPC的技能集合。释放技能的上下文信息可以是目标NPC已经释放的技能中还没有完成的技能的遗留效果。在一种可能的实现方式中，目标NPC在某一时刻释放了技能A，技能A的效果是在十秒内持续对目标NPC的对战对象造成伤害。在这一时刻之后

五秒的另一时刻，释放技能的上下文信息可以是技能A已经持续的对目标NPC的对战对象造成了五秒的伤害，技能A还会持续的对目标NPC的对战对象再造成五秒的伤害等信息。

[0046] 在一种可能的实现方式中，目标NPC在当前时刻在虚拟场景中的坐标为(100,300,200)，目标NPC的正面朝向x轴的正方向，上述信息可以作为目标NPC的当前状态信息。目标NPC的血量为300，攻击力为50，防御力为30，暴击率为50%，上述信息同样可以作为目标NPC的当前状态信息。目标NPC的当前状态信息还可以包括与目标NPC所属同一阵营的队友当前的战斗能力属性信息、移动信息、队友释放技能的上下文信息和队友的技能集合。

[0047] 在某一时刻与目标NPC进行对战的对象信息为目标NPC的对战对象的当前状态信息。在一种可能的实现方式中，与目标NPC进行对战的对象的坐标为(105,302,201)，与目标NPC进行对战的对象的正面朝向y轴的负半轴方向，上述信息可以作为目标NPC的对战对象的当前状态信息。目标NPC的对战对象的血量为5000，攻击力为100，防御力为10，暴击率为10%，上述信息同样可以作为目标NPC的对战对象的当前状态信息。

[0048] 目标NPC与对战对象的当前相对位姿信息用于表示目标NPC和对战对象的相对姿态，目标NPC与对战对象的相对位置。在一种可能的实现方式中，目标NPC的坐标为(100,300,200)，目标NPC进行对战的对象的坐标为(105,302,201)，利用目标NPC的坐标和目标NPC进行对战的对象的坐标可以确定目标NPC与对战对象的当前相对位置。目标NPC的正面朝向对战的对象，对战对象的正面朝向y轴的负半轴方向，目标NPC与对战对象正面的相对朝向信息可以作为相对姿态。

[0049] 目标NPC与对战对象的历史对战状态信息指的是在当前时刻之前的某个时刻，目标NPC与对战对象的历史对战状态。

[0050] S302：基于目标NPC的当前状态信息、对战对象的当前状态信息以及当前相对位姿信息进行特征提取及拼接，得到拼接后的特征。

[0051] 终端设备对目标NPC的当前状态信息、对战对象的当前状态信息以及当前相对位姿信息进行特征提取，特征提取后得到可以被神经网络识别的特征。在特征提取时不同的特征类型会有不同的处理方法。在一种可能的实现方式

中，目标NPC的当前状态信息为离散类别的数据，可以使用如独热编码(One Hot Encoding)等编码方式进行处理；当前相对位姿信息为连续的特征，可以使用如全连接网络编码等编码进行处理。终端设备经过特征提取得到多个特征，将多个特征拼接后得到拼接后的特征，在一种可能的实现方式中，拼接后的特征可以是长向量。

[0052] S303：通过长短期记忆网络对拼接后的特征及历史对战状态信息进行处理，得到长短期记忆网络输出的第一处理结果。

[0053] 长短期记忆(long short-term memory, LSTM)网络通过门控制将短期记忆与长期记忆结合起来，长短期记忆网络可以较好的处理长期信息保存和短期输入跳跃的问题。

[0054] 拼接后的特征用于代表某一时刻的特征，历史对战状态信息是在这一时刻之前的对战状态信息，长短期记忆网络对拼接后的特征及历史对战状态信息进行处理得到长短期记忆网络输出的第一处理结果。通过LSTM网络，可以从时序的角度更好的融合拼接后的特征及历史对战状态信息，并建立起拼接后的特征及历史对战状态信息之间的时序关系，这些信息均会体现在第一处理结果中，以便为后续的策略决策提供依据。

[0055] S304：基于第一处理结果、目标NPC释放技能的上下文信息、目标NPC的技能集合和强化学习奖励项的奖励值，通过强化学习的方式对目标NPC下一时刻应采取的对战策略进行决策。

[0056] 强化学习可以得到目标NPC在虚拟环境中执行的动作，目标NPC执行的每一个动作都会有一个对应的奖励项，奖励项是期望目标NPC能够达到的结果。一般对于目标NPC来说，削弱对战对象的战斗能力为期望目标NPC能够达到的结果，所以一般将削弱对战对象的战斗能力作为奖励项。基于削弱所述对战对象的战斗能力，可以确定奖励项的奖励值，一般来说，对该对战对象的战斗能力削弱的越大，对应的奖励项的奖励值就越大。

[0057] 在一种可能的实现方式中，奖励项可以是对战对象的血量减少百分比，例如，目标NPC执行动作1可以让对战对象的血量减少2%，目标NPC执行动作2可以让

对战对象的血量减少5%，动作1和动作2都会有对应的奖励值，因为动作2让对战对象的血量减少了更多的百分比，所以动作2的奖励值会高于动作1的奖励值。

[0058] 在另一种可能的实现方式中，奖励项可以同时考虑对战对象的血量减少百分比和目标NPC血量减少的百分比，例如，目标NPC执行动作3可以让对战对象的血量减少2%，自身血量减少5%；目标NPC执行动作4可以让对战对象的血量减少5%，自身血量减少20%。动作3和动作4都有对应的奖励值，虽然动作4让对战对象的血量减少的百分比更多，但是动作4让目标NPC血量减少的百分比同样更多，综合考虑动作3和动作4，动作3的奖励值一般会高于动作4的奖励值。

[0059] 在再一种可能的实现方式中，奖励项可以是同时考虑对战对象的血量减少百分比和目标NPC血量减少的百分比，并且为对象的血量减少百分比和目标NPC血量减少的百分比设置一定的权重，综合的得到最终的奖励值。

[0060] 削弱对战对象的战斗能力除了直接削减对战对象的血量减少百分比外，还可以是降低对战对象的当前状态信息。在一种可能的实现方式中，目标NPC执行动作5可以让对战对象的防御力降低50%，目标NPC执行动作6可以让对战对象的血量减少3%，动作5和动作6都会有对应的奖励值，综合考虑动作5和动作6，动作5的奖励值一般会高于动作6的奖励值。

[0061] 在一种可能的实现方式中，奖励项还可以考虑击败对战对象的时间。目标NPC执行动作7击败对战对象需要三十秒，目标NPC执行动作8击败对战对象只需要二十秒，动作7和动作8都会有对应的奖励值，综合考虑动作7和动作8，动作8的奖励值会高于动作7的奖励值。

[0062] 在另一种可能的实现方式中，奖励项还可以考虑目标NPC和对战对象之间的距离。目标NPC执行动作9时，目标NPC和对战对象之间一直保持较远的距离；目标NPC执行动作10时，目标NPC和对战对象之间的距离会很近。动作9和动作10都会有对应的奖励值，综合考虑动作9和动作10，动作10的奖励值会高于动作9的奖励值。

[0063] 强化学习的奖励项会按照实际需求确定，一般来说强化学习的奖励项会有多个指标，每个指标有对应的权重，在一种可能的实现方式中，可以对多个奖励项

各自的奖励值进行加权求和得到奖励值。终端设备通过强化学习的方式对目标NPC下一时刻应采取的对战策略进行决策。

- [0064] 在一种可能的实现方式中，终端设备对目标NPC释放技能的上下文信息和目标NPC的技能集合进行特征提取，得到目标NPC的多个技能特征；技能集合包括目标NPC的每一种技能的名称、效果标签、冷却时间、释放距离和当前的可用性。例如，目标NPC有两个技能，技能A和技能B；技能A的技能名称为回复、效果标签为回复自身10%的血量百分比、冷却时间为30秒，释放距离为1，当前的可用性为可用；技能B的技能名称为火球、效果标签为减少对战对象20点的血量值、冷却时间为2秒，释放距离为20，当前的可用性为不可用，技能集合中包括技能A对应的上述多个特征和技能B对应的上述多个特征。对技能A和技能B的特征进行拼接，得到技能特征拼接结果；基于第一处理结果、技能特征拼接结果和强化学习的奖励项，通过强化学习的方式对所述目标NPC下一时刻应采取的对战策略进行决策。
- [0065] 可见，考虑到第一处理结果体现了目标NPC和对战对象在全局时序上的整体特征，而目标NPC技能是对目标NPC进行策略决策的关键因素之一，为了提升与目标NPC技能相关的特征在后续决策中的比重，先将目标NPC释放技能的上下文信息和目标NPC的技能集合进行特征拼接，然后将得到的技能特征拼接结果和第一处理结果用于强化学习，使得强化学习中，能够获取更为清晰且占比更重的目标NPC技能信息，最终提高通过强化学习对目标NPC下一时刻应采取的对战策略的准确性。
- [0066] 基于第一处理结果、技能特征拼接结果和前一时刻采取的对战策略获得的奖励项的奖励值，预测目标NPC下一时刻的第一概率分布、第二概率分布和第三概率分布。第一概率分布为关于释放的技能的概率分布。在一种可能的实现方式中，目标NPC有两个技能，技能A和技能B，本段对于技能A和技能B的描述与上一段的描述一致，由于技能A当前的可用性为可用，技能B当前的可用性为不可用，所以第一概率分布中对于释放技能A的概率为100%，对于释放技能B的概率为0%。在另一种可能的实现方式中，除了技能A和技能B外，目标NPC掌握技能C，技能C的技能名称为冰冻、效果标签为使对战对象的移动速度降低为0、冷却

时间为20秒，释放距离为20，当前的可用性为可用，终端设备可以预测基于第一处理结果和技能特征拼接结果预测目标NPC下一时刻关于释放三个技能的概率分布。

[0067] 第二概率分布为关于移动的方向的概率分布，第三概率分布为关于移动的距离的概率分布。在一种可能的实现方式中，目标NPC掌握技能D，技能D的释放距离为5，第一处理结果中显示的目标NPC与对战对象之间的距离为10，为了释放技能D，目标NPC需要通过移动减少与对战对象之间的距离，目标NPC可以沿着直线向对战对象移动，目标NPC也可以通过走弧线的方式向对战对象移动，第二概率分布为关于移动的方向的概率分布。目标NPC可以沿着直线向对战对象移动，目标NPC可以沿着直线向对战对象移动8，目标NPC也可以沿着直线向对战对象移动5，第三概率分布为关于移动的距离的概率分布。在另一种可能的实现方式中，目标NPC掌握技能E，技能E的效果为回复自己50点的生命值，目标NPC释放技能E时无需进行任何的移动或转向，这时可以直接控制目标NPC释放技能E，无需进行任何的移动或转向。在再一种可能的实现方式中，目标NPC掌握技能F，技能F的效果为攻击以目标NPC为中心，范围10内的敌对角色，目标NPC释放技能F时无需进行任何的移动或转向，这时可以直接控制目标NPC释放技能F，无需进行任何的移动或转向。

[0068] 在一种可能的实现方式中，第二概率分布包括宏观移动方向概率分布和微观移动方向概率分布。微观移动指的在虚拟环境中进行的微小的移动，例如，目标NPC向左移动了三个像素点。宏观移动指的是目标NPC在虚拟环境中进行了较大范围的移动，例如，目标NPC在虚拟环境中由坐标(0,0,0)移动至坐标(0,0,1000)。

[0069] 第三概率分布包括宏观移动距离概率分布和微观移动距离概率分布，基于宏观移动距离概率分布中的最高值确定目标宏观移动距离，并基于微观移动距离概率分布中的最高值确定目标微观移动距离。在一种可能的实现方式中，宏观移动距离可以从一个开放式虚拟场景中的一个坐标移动到另一个坐标中的距离，微观移动距离可以是抬腿的高度，摆手的幅度等。

- [0070] 基于第一概率分布中的最高值确定目标技能，基于第二概率分布中的最高值确定目标方向，并基于第三概率分布中的最高值确定目标距离。在一种可能的实现方式中，第一概率分布中技能C为最高值，第二概率分布中沿着直线向对战对象移动为最高值，第三概率分布中移动5为最高值。基于上述信息可以模拟目标NPC下一时刻释放目标技能、向目标方向移动并移动所述目标距离，目标NPC下一时刻，第二时刻释放技能C，沿着直线向对战对象移动5。当目标NPC在下一时刻释放技能C，沿着直线向对战对象移动5之后可以得到奖励值。基于第二时刻的动作，可以得到目标NPC在第二时刻之后的第三时刻执行的动作，并得到第三时刻执行的动作对应的奖励值。以获得的奖励项的持续累计值最高为目标，根据奖励项的持续累计值对目标NPC下一时刻应采取的对战策略进行决策。
- [0071] 本申请提供的方法通过长短期记忆网络对拼接后的特征及历史对战状态信息进行处理，得到长短期记忆网络输出的第一处理结果。本申请提供的方法不仅考虑到当前时刻的特征，还考虑到历史对战状态信息，在使用第一处理结果进行强化学习时可以得到更合理的目标NPC动作。本申请通过强化学习的方式对目标NPC下一时刻应采取的对战策略进行决策。只要将某一时刻对应的信息作为强化学习的输入，即可得到目标NPC下一时刻应采取的对战策略，与相关技术相比，无需建立数量众多，结构复杂的行为树，在节约了大量的人力成本的同时，还减少了确定目标NPC对战策略所需的时间。
- [0072] 在一种可能的实现方式中，目标NPC的当前状态信息还包括目标NPC当前的战斗能力属性信息和移动信息，战斗能力属性信息可以是目标NPC当前的血量、攻击力、防御力、暴击率和移动速度等属性信息，也可以是对目标NPC的战斗能力属性进行评估的分值。例如，目标NPC当前的战斗能力属性信息可以是血量200、攻击力10、防御力30、暴击率10%和移动速度50。目标NPC当前的战斗能力属性信息还可以是将上述信息综合评价得到的战斗能力评分，例如，一点血量的评分为5、一点攻击力的评分为10、一点防御力的评分为3、一点暴击率的评分为15，一点移动速度的评分为5；将血量200、攻击力10、防御力30、暴击率10%和移动速度50进行评分后，得到的战斗能力评分为1590分，1590分可以作为

目标NPC当前的战斗能力属性信息。移动信息是目标NPC在虚拟场景在进行移动的相关信息。

- [0073] 对战对象的当前状态信息包括对战对象当前的战斗能力属性信息和移动信息与目标NPC的当前状态信息包括对战对象当前的战斗能力属性信息和移动信息类似，这里不做过多的赘述。对战对象的当前状态信息中还包括了作战对象目标NPC释放技能的上下文信息。例如，对战对象拥有技能X，技能X需要十秒的释放时间，在某个时刻对战对象已经进行了三秒的技能X释放，这个信息会作为对战对象目标NPC释放技能的上下文信息。
- [0074] 当前相对位姿信息包括所述对战对象与所述目标NPC的相对位置和相对朝向。
- [0075] 终端设备对目标NPC的当前状态信息、对战对象的当前状态信息以及当前相对位姿信息分别进行特征提取，得到目标NPC的当前状态特征、作战对象的当前状态特征以及目标NPC与对战对象的当前相对位姿特征。终端设备对目标NPC的当前状态特征中包含的多个特征进行拼接，得到第一拼接结果；对作战对象的当前状态特征中包含的多个特征进行拼接，得到第二拼接结果；对当前相对位姿特征中包含的多个特征进行拼接，得到第三拼接结果。拼接时可以按照预设的特征顺序进行拼接，例如，终端设备对目标NPC的当前状态特征中包含的多个特征进行拼接时，预设的特征顺序规定按照血量特征、攻击力特征、防御力特征、暴击率特征、移动速度特征和移动信息特征的顺序进行拼接，最终得到的第一拼接结果中特征的顺序和预设的特征顺序保持一致。
- [0076] 终端设备采用第一深度神经网络、第二深度神经网络和第三深度神经网络分别提取第一拼接结果、第二拼接结果和第三拼接结果中的特征，得到第一提取结果、第二提取结果和第三提取结果。在一种可能的实现方式中，第一深度神经网络、第二深度神经网络和第三深度神经网络可以使用DNN(Deep-Learning Neural Network)神经网络。
- [0077] 终端设备对第一提取结果、第二提取结果和第三提取结果进行拼接，得到拼接后的特征。
- [0078] 本申请提供的方法中，对目标NPC的当前状态信息、对战对象的当前状态信息和当前相对位姿信息分别进行特征提取和拼接。本申请考虑到目标NPC和对战

对象的区别，并未将目标NPC的当前状态信息和对战对象的当前状态信息提取后拼接到一起，而是分别进行处理。通过分别处理目标NPC的当前状态信息和对战对象的当前状态信息可以较大幅度的考虑目标NPC和对战对象的区别，最终提高通过强化学习对目标NPC下一时刻应采取的对战策略的准确性。

- [0079] 在一种可能的实现方式中，终端设备还可以获取虚拟场景中目标NPC的全局环境信息，全局环境信息包括时间信息、对战对象的数目和环境预警感知信息。其中，环境预警感知信息包括对战对象释放技能前的预警范围和对战对象释放技能后的持续效果。例如，对战对象在某个时刻释放技能喷火，喷火的影响范围是对战对象正面朝向的一个扇形范围，这个扇形的范围可以作为对战对象释放技能前的预警范围。对战对象释放喷火后，会在扇形范围内形成持续五秒的火焰，扇形范围内形成的持续五秒的火焰可以作为对战对象释放技能后的持续效果。
- [0080] 终端设备可以将全局环境信息与第一特征提取结果、第二特征提取结果和第三特征提取结果进行拼接，得到拼接后的特征。
- [0081] 本申请提供的方法考虑到了全局环境信息对目标NPC对战决策的影响，将获取到的全局环境信息与目标NPC当前状态信息对应的第一特征提取结果、对战对象的当前状态信息对应的第二特征提取结果以及当前相对位姿信息对应的第三特征提取结果进行拼接。在强化学习的过程中可以更准确的得到目标NPC的对战决策。
- [0082] 在一种可能的实现方式中，终端设备可以获取目标NPC下一时刻释放的目标技能、向目标方向移动并移动所述目标距离后得到的奖励项的最新奖励值，目标技能可以是攻击技能，目标NPC下一时刻向东移动20的距离后释放技能A攻击对战对象，目标NPC向东移动20的距离且释放技能A攻击对战对象后得到最新奖励值。在另一种可能的实现方式中，终端设备可以获取控制目标NPC下一时刻释放目标技能后得到的奖励项的最新奖励值，目标技能可以是治愈性技能，目标NPC下一时刻为自己释放目标技能，释放后得到最新奖励值。
- [0083] 终端设备可以根据最新奖励值、虚拟场景中目标NPC的最新状态信息，目标NPC的对战对象的最新状态信息，目标NPC与对战对象的最新相对位姿信息，

以及目标NPC与对战对象的最新历史对战状态信息，通过强化学习的方式对目标NPC在目标时刻应采取的对战策略进行决策，目标时刻为下一时刻之后的时刻。

- [0084] 在一种可能的实现方式中，若目标NPC下一时刻向东移动20的距离后释放技能A攻击对战对象，则可以获得20的最新奖励值。在目标NPC向东移动20的距离后释放技能A攻击对战对象后，目标NPC与对战对象的相对位姿信息，目标NPC的对战对象的状态信息都会发生变化，终端设备可以通过强化学习的方式对目标NPC在下一时刻之后的目标时刻应采取的对战策略进行决策。
- [0085] 在一种可能的实现方式中，目标NPC下一时刻，第二时刻释放技能A，沿着直线向对战对象移动5。目标NPC在执行上述动作之后获得奖励值30，但是由于执行了上述动作，目标NPC站在了带有火焰的扇形范围中，目标NPC的奖励值会减10。在第二时刻，利用奖励值作为评价目标NPC动作的标准。基于第二时刻的动作，可以得到目标NPC在第二时刻之后的第三时刻执行的动作，并得到第三时刻执行的动作对应的奖励值。根据第二时刻和第三时刻获得的累计奖励值对所述目标NPC下一时刻应采取的对战策略进行决策。
- [0086] 当确定若目标NPC下一时刻释放所述目标技能、向目标方向移动并移动目标距离，使累计的奖励项的奖励值最大，则决策出目标NPC下一时刻应释放所述目标技能、应向所述目标方向移动并应移动所述目标距离。在一种可能的实现方式中，目标NPC下一时刻，第二时刻释放技能A，沿着直线向对战对象移动5这个动作使目标NPC奖励项的持续累计值最大，这时，将释放技能A，沿着直线向对战对象移动5这个动作作为目标NPC的对战策略。
- [0087] 当确定若目标NPC下一时刻释放所述目标技能、向目标方向移动并移动目标距离，无法使奖励项的持续累计值最大，则通过模拟确定出使奖励项最大化的技能、移动方向和移动距离，作为所述目标NPC下一时刻应释放的技能、应移动的方向和应移动的距离。在另一种可能的实现方式中，沿着直线向对战对象移动5这个动作使目标NPC奖励项的持续累计不是最大的，则通过模拟确定出动作二，释放技能B，原地不动；动作三，释放技能C，沿着直线向对战对象的反方

向移动6。动作三可以使奖励项最大，终端设备将释放技能C，沿着直线向对战对象的反方向移动6作为目标NPC的对战策略。

[0088] 通过将每一个控制完目标NPC释放技能的时刻的奖励值作为最新奖励值，以此指示强化学习基于此继续进行下一个时刻的对战策略决策，从而可以持续的通过强化学习控制目标NPC与对战对象进行交互，保证了控制的持续性。

[0089] 图4为本申请实施例提供的一种基于强化学习的NPC对战决策方法的应用场景示意图，这里的游戏客户端示例性的选择了智能手机，当然除了智能手机外还可以是台式电脑、平板电脑、笔记本电脑、掌上电脑和车载终端等。游戏客户端向服务器发送信息，游戏客户端向服务器发送目标NPC的当前状态信息，目标NPC的对战对象的当前状态信息，目标NPC与所述对战对象的当前相对位姿信息，以及目标NPC与对战对象的历史对战状态信息；目标NPC的当前状态信息包括目标NPC释放技能的上下文信息和所述目标NPC的技能集合等信息。多个服务器通过强化学习得到目标NPC的对战决策，将对战决策返回游戏客户端。

[0090] 图5为本申请实施例提供的一种基于强化学习的NPC对战决策方法的结构示意图。图5中的Agent为目标NPC，State表示目标NPC的当前状态信息，将目标NPC的当前状态信息经过DNN神经网络得到policy $\pi_{\theta}(s, a)$ ，parameter θ 为DNN的参数，policy $\pi_{\theta}(s, a)$ 表示Agent的动作。通过Take action将动作输出至环境Environment，根据动作对环境的影响确定奖励值Reward，将环境Environment中的部分状态Observe state返回Agent。Agent基于返回的奖励值Reward和部分状态Observe state进行下一次的动作输出。

[0091] 图6为本申请提供的一种基于强化学习的NPC对战决策方法的网络结构图。图6中，最下面一行的env_info表示全局环境信息，relative_info表示目标NPC与所述对战对象的当前相对位姿信息，enemy_info表示目标NPC的对战对象的当前状态信息，self_info表示目标NPC的当前状态信息，lstm_info表示目标NPC与所述对战对象的历史对战状态信息。

[0092] 将目标NPC的当前状态信息self_info中的信息经过特征提取层self_layer得到目标NPC的当前状态信息对应的特征，通过拼接层self_concat对目标NPC的当前状态

信息对应的特征进行拼接得到目标NPC的当前状态信息对应的长向量，将目标NPC的当前状态信息对应的长向量输入第一深度神经网络DNN得到第一提取结果。将目标NPC的当前状态信息中目标NPC释放技能的上下文信息和所述目标NPC的技能集合输入特征提取层self_layer得到技能相关的特征，将技能相关的特征输入拼接层skill_concat，通过拼接层skill_concat拼接为长向量。

- [0093] 如果目标NPC的当前状态信息中包括作为队友的NPC的当前状态信息时，可以增加一个池化层Pooling，通过池化层Pooling保证输出特征的维度一致。
- [0094] 将对战对象的当前状态信息enemy_info中的信息经过特征提取层enemy_layer得到对战对象的当前状态信息对应的特征，将对战对象的当前状态信息对应的特征输入拼接层enemy_concat得到对战对象的当前状态信息对应的长向量，将对战对象的当前状态信息对应的长向量输入第二深度神经网络DNN得到第二提取结果。
- [0095] 将目标NPC与所述对战对象的当前相对位姿信息relative_info中的信息经过特征提取层relative_layer得到当前相对位姿信息对应的特征，将当前相对位姿信息对应的特征输入破拼接层relative_concat得到相对位姿信息对应的长向量，将相对位姿信息对应的长向量输入第三深度神经网络DNN得到第三提取结果。
- [0096] 在连接层concat中将全局环境信息、第一提取结果、第二提取结果和第三提取结果进行拼接，得到拼接后的特征。
- [0097] 通过长短期记忆网络LSTM对拼接后的特征及历史对战状态信息lstm_info进行处理，得到长短期记忆网络输出的第一处理结果。基于第一处理结果、目标NPC释放技能的上下文信息、目标NPC的技能集合对应的特征和强化学习的奖励项通过强化学习的方式得到多个输出头Head，多个输出头Head中可以包括释放技能的输出头，确定方向的输出头和确定距离的输出头。对于每一个输出头都有一个对应的 $\pi(\cdot | s)$ 来表示输出头对应的概率分布。例如，释放技能的输出头对应了目标NPC关于释放的技能的第一概率分布，确定方向的输出头对应了目标NPC关于移动的方向的第二概率分布，确定距离的输出头对应了目标NPC关于移动的距离的第三概率分布。每个输出头还对应了一个函数值 $v(s)$ ， $v(s)$ 用于评估输出头输出动作的好坏，函数值 $v(s)$ 与奖励项有关。释放技能的输出头会用

到拼接层skill_concat输出的技能相关的长向量，通过拼接层skill_concat保证释放技能的输出头与目标NPC的技能相互对应。

- [0098] 本申请使用的网络结构通过状态的输入得到动作的输出，可以表示从状态到动作的映射关系。根据输入的状态不同，本申请也可以使用不同的网络结构来处理。在一种可能的实现方式中，输入的状态特征为连续的特征或离散的特征，可以使用DNN神经网络。在另一种可能的实现方式中，输入的特征为二维平面网络结构，例如用于显示对战对象释放技能前的预警范围的单通道图片，这时可以把单通道图片离散化为网格，通过CNN神经网络处理。
- [0099] 基于前文实施例提供的一种基于强化学习的NPC对战决策方法，本申请中还相应提供了一种基于强化学习的NPC对战决策装置700。以下结合图7进行说明。图7为本申请实施例提供的一种基于强化学习的NPC对战决策装置的结构示意图。如图7所示的一种基于强化学习的NPC对战决策装置包括：
- [0100] 信息获取模块701，用于获取虚拟场景中目标NPC的当前状态信息，所述目标NPC的对战对象的当前状态信息，所述目标NPC与所述对战对象的当前相对位姿信息，以及所述目标NPC与所述对战对象的历史对战状态信息；所述目标NPC的当前状态信息包括所述目标NPC释放技能的上下文信息和所述目标NPC的技能集合；
- [0101] 拼接模块702，用于基于所述目标NPC的当前状态信息、所述对战对象的当前状态信息以及所述当前相对位姿信息进行特征提取及拼接，得到拼接后的特征；
- [0102] 长短期记忆网络模块703，用于通过长短期记忆网络对所述拼接后的特征及所述历史对战状态信息进行处理，得到所述长短期记忆网络输出的第一处理结果；
- [0103] 强化学习模块704，用于基于所述第一处理结果、所述目标NPC释放技能的上下文信息、所述目标NPC的技能集合和强化学习的奖励项，通过强化学习的方式对所述目标NPC下一时刻应采取的对战策略进行决策；其中，所述奖励项的奖励值为基于削弱所述对战对象的战斗能力确定的。

- [0104] 在一种可能的实现方式中，所述目标NPC的当前状态信息还包括所述目标NPC当前的战斗能力属性信息和移动信息；所述对战对象的当前状态信息包括所述对战对象当前的战斗能力属性信息、移动信息和所述作战对象目标NPC释放技能的上下文信息；所述当前相对位姿信息包括所述对战对象与所述目标NPC的相对位置和相对朝向；
- [0105] 所述拼接模块具体用于：
- [0106] 对所述目标NPC的当前状态信息、所述对战对象的当前状态信息以及所述当前相对位姿信息分别进行特征提取，得到所述目标NPC的当前状态特征、所述作战对象的当前状态特征以及所述目标NPC与所述对战对象的当前相对位姿特征；
- [0107] 对所述目标NPC的当前状态特征中包含的多个特征进行拼接，得到第一拼接结果；
- [0108] 对所述作战对象的当前状态特征中包含的多个特征进行拼接，得到第二拼接结果；
- [0109] 对所述当前相对位姿特征中包含的多个特征进行拼接，得到第三拼接结果；
- [0110] 采用第一深度神经网络、第二深度神经网络和第三深度神经网络分别提取所述第一拼接结果、所述第二拼接结果和所述第三拼接结果中的特征，得到第一提取结果、第二提取结果和第三提取结果；
- [0111] 对所述第一提取结果、所述第二提取结果和所述第三提取结果进行拼接，得到拼接后的特征。
- [0112] 在一种可能的实现方式中，所述基于强化学习的NPC对战决策装置还包括：
- [0113] 全局环境信息获取模块，用于获取所述虚拟场景中所述目标NPC的全局环境信息，所述全局环境信息包括时间信息、对战对象的数目和环境预警感知信息；其中，所述环境预警感知信息包括所述对战对象释放技能前的预警范围和所述对战对象释放技能后的持续效果；
- [0114] 所述拼接模块具体用于：
- [0115] 对所述全局环境信息、所述第一特征提取结果、所述第二特征提取结果和所述第三特征提取结果进行拼接，得到拼接后的特征。

- [0116] 在一种可能的实现方式中，强化学习模块具体用于：
- [0117] 对所述目标NPC释放技能的上下文信息和所述目标NPC的技能集合进行特征提取，得到所述目标NPC的多个技能特征；所述技能集合包括所述目标NPC的每一种技能的名称、效果标签、冷却时间、释放距离和当前的可用性；
- [0118] 对所述多个技能特征进行拼接，得到技能特征拼接结果；
- [0119] 基于所述第一处理结果、所述技能特征拼接结果和强化学习的奖励项的奖励值，通过强化学习的方式对所述目标NPC下一时刻应采取的对战策略进行决策。
- [0120] 在一种可能的实现方式中，所述强化学习模块具体用于：
- [0121] 基于所述第一处理结果、所述技能特征拼接结果和前一时刻采取的对战策略获得的奖励项的奖励值，预测所述目标NPC下一时刻的第一概率分布、第二概率分布和第三概率分布，其中第一概率分布为关于释放的技能的概率分布，所述第二概率分布为关于移动的方向的概率分布，所述第三概率分布为关于移动的距离的概率分布；
- [0122] 基于所述第一概率分布中的最高值确定目标技能，基于所述第二概率分布中的最高值确定目标方向，并基于所述第三概率分布中的最高值确定目标距离；
- [0123] 若所述目标技能为具有移动效果的攻击技能，控制所述目标NPC下一时刻释放所述目标技能、向所述目标方向移动并移动所述目标距离；
- [0124] 若所述目标技能为不具备移动效果的攻击技能或者所述目标技能为治愈性的技能，控制所述目标NPC下一时刻释放所述目标技能。
- [0125] 所述装置进一步包括目标时刻对战策略确定模块：
- [0126] 获取控制所述目标NPC下一时刻释放所述目标技能、向所述目标方向移动并移动所述目标距离后得到的奖励项的最新奖励值，或者获取控制所述目标NPC下一时刻释放所述目标技能后得到的奖励项的最新奖励值；
- [0127] 根据最新奖励值、所述虚拟场景中目标NPC的最新状态信息，所述目标NPC的对战对象的最新状态信息，所述目标NPC与所述对战对象的最新相对位姿信息，以及所述目标NPC与所述对战对象的最新历史对战状态信息，通过强化学习的方式对所述目标NPC在目标时刻应采取的对战策略进行决策；所述目标时刻为所述下一时刻之后的时刻。

- [0128] 在一种可能的实现方式中，所述强化学习包括多个奖励项，强化学习模块，还用于基于所述第一处理结果、所述目标NPC释放技能的上下文信息、所述目标NPC的技能集合和所述多个奖励项各自奖励值的加权求和，通过强化学习的方式对所述目标NPC下一时刻应采取的对战策略进行决策基于所述第一处理结果、所述目标NPC释放技能的上下文信息、所述目标NPC的技能集合和强化学习的奖励项的奖励值，通过强化学习的方式对所述目标NPC下一时刻应采取的对战策略进行决策。
- [0129] 在一种可能的实现方式中，所述目标NPC的当前状态信息还包括与所述目标NPC所属同一阵营的队友当前的战斗能力属性信息、移动信息、所述队友释放技能的上下文信息和所述队友的技能集合。
- [0130] 本申请实施例提供了一种基于强化学习的NPC对战决策设备，该基于强化学习的NPC对战决策设备可以为服务器。图8是本申请实施例提供的一种服务器结构示意图，该服务器900可因配置或性能不同而产生比较大的差异，可以包括一个或一个以上中央处理器(central processing units, CPU)922(例如，一个或一个以上处理器)和存储器932，一个或一个以上存储应用程序942或数据944的存储介质930(例如一个或一个以上海量存储设备)。其中，存储器932和存储介质930可以是短暂存储或持久存储。存储在存储介质930的程序可以包括一个或一个以上模块(图示没标出)，每个模块可以包括对服务器中的一系列指令操作。更进一步地，中央处理器922可以设置为与存储介质930通信，在服务器900上执行存储介质930中的一系列指令操作。
- [0131] 服务器900还可以包括一个或一个以上电源926，一个或一个以上有线或无线网络接口950，一个或一个以上输入输出接口958，和/或，一个或一个以上操作系统941。
- [0132] 其中，CPU 922用于执行如下步骤：
- [0133] 获取虚拟场景中目标NPC的当前状态信息，所述目标NPC的对战对象的当前状态信息，所述目标NPC与所述对战对象的当前相对位姿信息，以及所述目标NPC与所述对战对象的历史对战状态信息；所述目标NPC的当前状态信息包括所述目标NPC释放技能的上下文信息和所述目标NPC的技能集合；

- [0134] 基于所述目标NPC的当前状态信息、所述对战对象的当前状态信息以及所述当前相对位姿信息进行特征提取及拼接，得到拼接后的特征；
- [0135] 通过长短期记忆网络对所述拼接后的特征及所述历史对战状态信息进行处理，得到所述长短期记忆网络输出的第一处理结果；
- [0136] 基于所述第一处理结果、所述目标NPC释放技能的上下文信息、所述目标NPC的技能集合和强化学习的奖励项，通过强化学习的方式对所述目标NPC下一时刻应采取的对战策略进行决策；其中，所述奖励项的奖励值为基于削弱所述对战对象的战斗能力确定的。
- [0137] 本申请实施例还提供了另一种基于强化学习的NPC对战决策设备，该基于强化学习的NPC对战决策设备可以为终端设备。如图9所示，为了便于说明，仅示出了与本申请实施例相关的部分，具体技术细节未揭示的，请参照本申请实施例方法部分。以该终端设备为手机为例：
- [0138] 图9示出的是与本申请实施例提供的手机的部分结构的框图。参考图9，手机包括：射频(英文全称：Radio Frequency，英文缩写：RF)电路1010、存储器1020、输入单元1030、显示单元1040、传感器1050、音频电路1060、无线保真(英文全称：wireless fidelity，英文缩写：WiFi)模块1070、处理器1080、以及电源1090等部件。本领域技术人员可以理解，图9中示出的手机结构并不构成对手机的限定，可以包括比图示更多或更少的部件，或者组合某些部件，或者不同的部件布置。
- [0139] 下面结合图9对手机的各个构成部件进行具体的介绍：
- [0140] RF电路1010可用于收发信息或通话过程中，信号的接收和发送，特别地，将基站的下行信息接收后，给处理器1080处理；另外，将设计上行的数据发送给基站。通常，RF电路1010包括但不限于天线、至少一个放大器、收发信机、耦合器、低噪声放大器(英文全称：Low Noise Amplifier，英文缩写：LNA)、双工器等。此外，RF电路1010还可以通过无线通信与网络和其他设备通信。上述无线通信可以使用任一通信标准或协议，包括但不限于全球移动通讯系统(英文全称：Global System of Mobile communication，英文缩写：GSM)、通用分组无线服务(英文全称：General Packet Radio Service，GPRS)、码分多址(英文全称：

Code Division Multiple Access, 英文缩写: CDMA)、宽带码分多址(英文全称: Wideband Code Division Multiple Access, 英文缩写: WCDMA)、长期演进(英文全称: Long Term Evolution, 英文缩写: LTE)、电子邮件、短消息服务(英文全称: Short Messaging Service, SMS)等。

- [0141] 存储器1020可用于存储软件程序以及模块, 处理器1080通过运行存储在存储器1020的软件程序以及模块, 从而执行手机的各种功能应用以及数据处理。存储器1020可主要包括存储程序区和存储数据区, 其中, 存储程序区可存储操作系统、至少一个功能所需的应用程序(比如声音播放功能、图像播放功能等)等; 存储数据区可存储根据手机的使用所创建的数据(比如音频数据、电话本等)等。此外, 存储器1020可以包括高速随机存取存储器, 还可以包括非易失性存储器, 例如至少一个磁盘存储器件、闪存器件、或其他易失性固态存储器件。
- [0142] 输入单元1030可用于接收输入的数字或字符信息, 以及产生与手机的用户设置以及功能控制有关的键信号输入。具体地, 输入单元1030可包括触控面板1031以及其他输入设备1032。触控面板1031, 也称为触摸屏, 可收集用户在其上或附近的触摸操作(比如用户使用手指、触笔等任何适合的物体或附件在触控面板1031上或在触控面板1031附近的操作), 并根据预先设定的程式驱动相应的连接装置。可选的, 触控面板1031可包括触摸检测装置和触摸控制器两个部分。其中, 触摸检测装置检测用户的触摸方位, 并检测触摸操作带来的信号, 将信号传送给触摸控制器; 触摸控制器从触摸检测装置上接收触摸信息, 并将它转换成触点坐标, 再送给处理器1080, 并能接收处理器1080发来的命令并加以执行。此外, 可以采用电阻式、电容式、红外线以及表面声波等多种类型实现触控面板1031。除了触控面板1031, 输入单元1030还可以包括其他输入设备1032。具体地, 其他输入设备1032可以包括但不限于物理键盘、功能键(比如音量控制按键、开关按键等)、轨迹球、鼠标、操作杆等中的一种或多种。
- [0143] 显示单元1040可用于显示由用户输入的信息或提供给用户的信息以及手机的各种菜单。显示单元1040可包括显示面板1041, 可选的, 可以采用液晶显示器(英文全称: Liquid Crystal Display, 英文缩写: LCD)、有机发光二极管(英文全称: Organic Light-Emitting Diode, 英文缩写: OLED)等形式来配置显示面板1041。

进一步的，触控面板1031可覆盖显示面板1041，当触控面板1031检测到在其上或附近的触摸操作后，传送给处理器1080以确定触摸事件的类型，随后处理器1080根据触摸事件的类型在显示面板1041上提供相应的视觉输出。虽然在图9中，触控面板1031与显示面板1041是作为两个独立的部件来实现手机的输入和输入功能，但是在某些实施例中，可以将触控面板1031与显示面板1041集成而实现手机的输入和输出功能。

[0144] 手机还可包括至少一种传感器1050，比如光传感器、运动传感器以及其他传感器。具体地，光传感器可包括环境光传感器及接近传感器，其中，环境光传感器可根据环境光线的明暗来调节显示面板1041的亮度，接近传感器可在手机移动到耳边时，关闭显示面板1041和/或背光。作为运动传感器的一种，加速计传感器可检测各个方向上(一般为三轴)加速度的大小，静止时可检测出重力的大小及方向，可用于识别手机姿态的应用(比如横竖屏切换、相关游戏、磁力计姿态校准)、振动识别相关功能(比如计步器、敲击)等；至于手机还可配置的陀螺仪、气压计、湿度计、温度计、红外线传感器等其他传感器，在此不再赘述。

[0145] 音频电路1060、扬声器1061，传声器1062可提供用户与手机之间的音频接口。音频电路1060可将接收到的音频数据转换后的电信号，传输到扬声器1061，由扬声器1061转换为声音信号输出；另一方面，传声器1062将收集的声音信号转换为电信号，由音频电路1060接收后转换为音频数据，再将音频数据输出处理器1080处理后，经RF电路1010以发送给比如另一手机，或者将音频数据输出至存储器1020以便进一步处理。

[0146] WiFi属于短距离无线传输技术，手机通过WiFi模块1070可以帮助用户收发电子邮件、浏览网页和访问流式媒体等，它为用户提供了无线的宽带互联网访问。虽然图9示出了WiFi模块1070，但是可以理解的是，其并不属于手机的必须构成，完全可以根据需要在不改变发明的本质的范围内而省略。

[0147] 处理器1080是手机的控制中心，利用各种接口和线路连接整个手机的各个部分，通过运行或执行存储在存储器1020内的软件程序和/或模块，以及调用存储在存储器1020内的数据，执行手机的各种功能和处理数据，从而对手机进行整体数据及信息收集。可选的，处理器1080可包括一个或多个处理单元；优选的，处

理器1080可集成应用处理器和调制解调处理器，其中，应用处理器主要处理操作系统、用户界面和应用程序等，调制解调处理器主要处理无线通信。可以理解的是，上述调制解调处理器也可以不集成到处理器1080中。

- [0148] 手机还包括给各个部件供电的电源1090(比如电池)，优选的，电源可以通过电源管理系统与处理器1080逻辑相连，从而通过电源管理系统实现管理充电、放电、以及功耗管理等功能。
- [0149] 尽管未示出，手机还可以包括摄像头、蓝牙模块等，在此不再赘述。
- [0150] 在本申请实施例中，该手机所包括的处理器1080还具有以下功能：
- [0151] 获取虚拟场景中目标NPC的当前状态信息，所述目标NPC的对战对象的当前状态信息，所述目标NPC与所述对战对象的当前相对位姿信息，以及所述目标NPC与所述对战对象的历史对战状态信息；所述目标NPC的当前状态信息包括所述目标NPC释放技能的上下文信息和所述目标NPC的技能集合；
- [0152] 基于所述目标NPC的当前状态信息、所述对战对象的当前状态信息以及所述当前相对位姿信息进行特征提取及拼接，得到拼接后的特征；
- [0153] 通过长短期记忆网络对所述拼接后的特征及所述历史对战状态信息进行处理，得到所述长短期记忆网络输出的第一处理结果；
- [0154] 基于所述第一处理结果、所述目标NPC释放技能的上下文信息、所述目标NPC的技能集合和强化学习的奖励项，通过强化学习的方式对所述目标NPC下一时刻应采取的对战策略进行决策；其中，所述奖励项的奖励值为基于削弱所述对战对象的战斗能力确定的。
- [0155] 本申请实施例还提供一种计算机可读存储介质，用于存储计算机程序，该计算机程序在基于强化学习的NPC对战决策设备上运行时，使得该基于强化学习的NPC对战决策设备用于执行前述各个实施例所述的一种基于强化学习的NPC对战决策方法中的任意一种实施方式。
- [0156] 本申请实施例还提供一种包括计算机程序的计算机程序产品，当其在基于强化学习的NPC对战决策设备上运行时，使得基于强化学习的NPC对战决策设备执行前述各个实施例所述的一种基于强化学习的NPC对战决策方法中的任意一种实施方式。

- [0157] 所属领域的技术人员可以清楚地了解到，为描述的方便和简洁，上述描述的系统、设备的具体工作过程，可以参考前述方法实施例中的对应过程，在此不再赘述。
- [0158] 在本申请所提供的几个实施例中，应该理解到，所揭露的系统和方法，可以通过其它的方式实现。例如，以上所描述的系统实施例仅仅是示意性的，例如，所述系统的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个系统可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，装置或单元的间接耦合或通信连接，可以是电性，机械或其它的形式。
- [0159] 所述作为分离部件说明的系统可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。
- [0160] 另外，在本申请各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。上述集成的单元既可以采用硬件的形式实现，也可以采用软件功能单元的形式实现。
- [0161] 所述集成的单元如果以软件功能单元的形式实现并作为独立的产品销售或使用，可以存储在一个计算机可读取存储介质中。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分或者该技术方案的全部或部分可以以软件产品的形式体现出来，该计算机软件产品存储在一个存储介质中，包括若干指令用以使得一台计算机设备(可以是个人计算机，服务器，或者网络设备)执行本申请各个实施例所述方法的全部或部分步骤。而前述的存储介质包括：U盘、移动硬盘、只读存储器(英文全称：Read-Only Memory，英文缩写：ROM)、随机存取存储器(英文全称：Random Access Memory，英文缩写：RAM)、磁碟或者光盘等各种可以存储计算机程序的介质。

[0162] 以上实施例仅用以说明本申请的技术方案，而非对其限制；尽管参照前述实施例对本申请进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本申请各实施例技术方案的精神和范围。

权利要求书

- [权利要求 1] 一种基于强化学习的NPC对战决策方法，所述方法由计算机设备执行，所述方法包括：
- 获取虚拟场景中目标NPC的当前状态信息，所述目标NPC的对战对象的当前状态信息，所述目标NPC与所述对战对象的当前相对位姿信息，以及所述目标NPC与所述对战对象的历史对战状态信息；所述目标NPC的当前状态信息包括所述目标NPC释放技能的上下文信息和所述目标NPC的技能集合；
- 基于所述目标NPC的当前状态信息、所述对战对象的当前状态信息以及所述当前相对位姿信息进行特征提取及拼接，得到拼接后的特征；
- 通过长短期记忆网络对所述拼接后的特征及所述历史对战状态信息进行处理，得到所述长短期记忆网络输出的第一处理结果；
- 基于所述第一处理结果、所述目标NPC释放技能的上下文信息、所述目标NPC的技能集合和强化学习的奖励项的奖励值，通过强化学习的方式对所述目标NPC下一时刻应采取的对战策略进行决策；其中，所述奖励项的奖励值为基于削弱所述对战对象的战斗能力确定的。
- [权利要求 2] 根据权利要求1所述的方法，所述基于所述目标NPC的当前状态信息、所述对战对象的当前状态信息以及所述当前相对位姿信息进行特征提取及拼接，得到拼接后的特征，包括：
- 对所述目标NPC的当前状态信息、所述对战对象的当前状态信息以及所述当前相对位姿信息分别进行特征提取，得到所述目标NPC的当前状态特征、所述作战对象的当前状态特征以及所述目标NPC与所述对战对象的当前相对位姿特征；
- 对所述目标NPC的当前状态特征中包含的多个特征进行拼接，得到第一拼接结果；

对所述作战对象的当前状态特征中包含的多个特征进行拼接，得到第二拼接结果；

对所述当前相对位姿特征中包含的多个特征进行拼接，得到第三拼接结果；

采用第一深度神经网络、第二深度神经网络和第三深度神经网络分别提取所述第一拼接结果、所述第二拼接结果和所述第三拼接结果中的特征，得到第一提取结果、第二提取结果和第三提取结果；

对所述第一提取结果、所述第二提取结果和所述第三提取结果进行拼接，得到拼接后的特征。

[权利要求 3]

根据权利要求2所述的方法，所述方法还包括：

获取所述虚拟场景中所述目标NPC的全局环境信息，所述全局环境信息包括时间信息、对战对象的数目和环境预警感知信息；其中，所述环境预警感知信息包括所述对战对象释放技能前的预警范围和所述对战对象释放技能后的持续效果；

对所述第一提取结果、所述第二提取结果和所述第三提取结果进行拼接，得到拼接后的特征，包括：

对所述全局环境信息、所述第一特征提取结果、所述第二特征提取结果和所述第三特征提取结果进行拼接，得到拼接后的特征。

[权利要求 4]

根据权利要求1所述的方法，所述基于所述第一处理结果、所述目标NPC释放技能的上下文信息、所述目标NPC的技能集合和强化学习的奖励项的奖励值，通过强化学习的方式对所述目标NPC下一时刻应采取的对战策略进行决策，包括：

对所述目标NPC释放技能的上下文信息和所述目标NPC的技能集合进行特征提取，得到所述目标NPC的多个技能特征；所述技能集合包括所述目标NPC的每一种技能的名称、效果标签、冷却时间、释放距离和当前的可用性；

对所述多个技能特征进行拼接，得到技能特征拼接结果；

基于所述第一处理结果、所述技能特征拼接结果和强化学习的奖励项的奖励值，通过强化学习的方式对所述目标NPC下一时刻应采取的对战策略进行决策。

[权利要求 5]

根据权利要求4所述的方法，所述基于所述第一处理结果、所述技能特征拼接结果和强化学习的奖励项的奖励值，通过强化学习的方式对所述目标NPC下一时刻应采取的对战策略进行决策，包括：

基于所述第一处理结果、所述技能特征拼接结果和前一时刻采取的对战策略获得的奖励项的奖励值，预测所述目标NPC下一时刻的第一概率分布、第二概率分布和第三概率分布，其中第一概率分布为关于释放的技能的概率分布，所述第二概率分布为关于移动的方向的概率分布，所述第三概率分布为关于移动的距离的概率分布；基于所述第一概率分布中的最高值确定目标技能，基于所述第二概率分布中的最高值确定目标方向，并基于所述第三概率分布中的最高值确定目标距离；

若所述目标技能为具有移动效果的攻击技能，控制所述目标NPC下一时刻释放所述目标技能、向所述目标方向移动并移动所述目标距离；

若所述目标技能为不具备移动效果的攻击技能或者所述目标技能为治愈性的技能，控制所述目标NPC下一时刻释放所述目标技能。

[权利要求 6]

根据权利要求5所述的方法，还包括：

获取控制所述目标NPC下一时刻释放所述目标技能、向所述目标方向移动并移动所述目标距离后得到的奖励项的最新奖励值，或者获取控制所述目标NPC下一时刻释放所述目标技能后得到的奖励项的最新奖励值；

根据最新奖励值、所述虚拟场景中目标NPC的最新状态信息，所述目标NPC的对战对象的最新状态信息，所述目标NPC与所述对战对象的最新相对位姿信息，以及所述目标NPC与所述对战对象的最新历史对战状态信息，通过强化学习的方式对所述目标NPC

在目标时刻应采取的对战策略进行决策；所述目标时刻为所述下一时刻之后的时刻。

[权利要求 7]

根据权利要求1-6任一项所述的方法，所述强化学习包括多个奖励项，所述基于所述第一处理结果、所述目标NPC释放技能的上下文信息、所述目标NPC的技能集合和强化学习的奖励项的奖励值，通过强化学习的方式对所述目标NPC下一时刻应采取的对战策略进行决策包括：

基于所述第一处理结果、所述目标NPC释放技能的上下文信息、所述目标NPC的技能集合和所述多个奖励项各自奖励值的加权求和，通过强化学习的方式对所述目标NPC下一时刻应采取的对战策略进行决策基于所述第一处理结果、所述目标NPC释放技能的上下文信息、所述目标NPC的技能集合和强化学习的奖励项的奖励值，通过强化学习的方式对所述目标NPC下一时刻应采取的对战策略进行决策。

[权利要求 8]

根据权利要求1所述的方法，所述目标NPC的当前状态信息还包括与所述目标NPC所属同一阵营的队友当前的战斗能力属性信息、移动信息、所述队友释放技能的上下文信息和所述队友的技能集合。

[权利要求 9]

一种基于强化学习的NPC对战决策装置，包括：
信息获取模块，用于获取虚拟场景中目标NPC的当前状态信息，所述目标NPC的对战对象的当前状态信息，所述目标NPC与所述对战对象的当前相对位姿信息，以及所述目标NPC与所述对战对象的历史对战状态信息；所述目标NPC的当前状态信息包括所述目标NPC释放技能的上下文信息和所述目标NPC的技能集合；
拼接模块，用于基于所述目标NPC的当前状态信息、所述对战对象的当前状态信息以及所述当前相对位姿信息进行特征提取及拼接，得到拼接后的特征；

长短期记忆网络模块，用于通过长短期记忆网络对所述拼接后的特征及所述历史对战状态信息进行处理，得到所述长短期记忆网络输出的第一处理结果；

强化学习模块，用于基于所述第一处理结果、所述目标NPC释放技能的上下文信息、所述目标NPC的技能集合和强化学习的奖励项的奖励值，通过强化学习的方式对所述目标NPC下一时刻应采取的对战策略进行决策；其中，所述奖励项的奖励值为基于削弱所述对战对象的战斗能力确定的。

[权利要求 10]

一种基于强化学习的NPC对战决策，所述设备包括处理器以及存储器：

所述存储器用于存储计算机程序，并将所述计算机程序传输给所述处理器；

所述处理器用于根据所述计算机程序中的指令执行权利要求1至8中任一项所述的基于强化学习的NPC对战决策方法的步骤。

[权利要求 11]

一种计算机可读存储介质，所述计算机可读存储介质用于存储计算机程序，所述计算机程序被基于强化学习的NPC对战决策设备执行时实现权利要求1至8任一项所述的基于强化学习的NPC对战决策方法的步骤。

[权利要求 12]

一种计算机程序产品，包括计算机程序，该计算机程序被基于强化学习的NPC对战决策设备执行时实现权利要求1至8任一项所述的基于强化学习的NPC对战决策方法的步骤。