说　明　书

应用程序巡检方法、装置、存储介质和计算机系统

技术领域

本说明书一个或多个实施例涉及计算机领域，尤其涉及一种基于强化学习的应用程序巡检方法、装置、存储介质和计算机系统。

背景技术

对能够运行于终端的应用程序进行巡检时，通常需要进行大量的人工干预。例如，需要由测试人员来识别终端对应用程序显示的页面中能够用于进行点击交互的触发区域，对触发区域执行点击动作，以便根据点击动作的执行结果来判别当前页面是否存在漏洞。

希望有一种新的技术方案，以期能够更加高效且全面的完成对应用程序进行巡检。

发明内容

本说明书一个或多个实施例中提供了一种基于强化学习的应用程序巡检方法、装置、存储介质和计算机系统。

本说明书提供了一种基于强化学习的应用程序巡检方法，所述方法包括：从终端对目标应用程序显示的第一页面中，确定用于进行点击交互的N个触发区域；根据所述N个触发区域，确定所述第一页面的页面特征；将所述页面特征输入强化学习模型，使其输出与动作空间相对应的动作决策信息，所述动作空间包括若干预设动作以及分别用于触发所述N个触发区域的N个触发动作；根据所述动作决策信息，对所述第一页面执行目标动作，使所述终端显示第二页面；获取在所述第一页面中执行所述目标动作的环境奖励，其中所述环境奖励在所述第二页面是所述终端中新出现页面时的取值，大于所述第二页面不是所述终端中新出现页面时的取值，所述环境奖励用于更新所述强化学习模型。

本说明书提供了一种计算机系统，包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序，所述处理器执行所述计算机程序时实现如下步骤：从终端对目标应用程序显示的第一页面中，确定用于进行点击交互的N个触发区域；根据所述N个触发区域，确定所述第一页面的页面特征；将所述页面特征输入强化学习模型，使其输出与动作空间相对应的动作决策信息，所述动作空间包括若干预设动作以及分别用于触发所述N个触发区域的N个触发动作；根据所述动作决策信息，对所述第一页面执行目标

动作，使所述终端显示第二页面；获取在所述第一页面中执行所述目标动作的环境奖励，其中所述环境奖励在所述第二页面是所述终端中新出现页面时的取值，大于所述第二页面不是所述终端中新出现页面时的取值，所述环境奖励用于更新所述强化学习模型。

本说明书提供了一种基于强化学习的应用程序巡检装置，包括：目标检测单元，配置为从终端对目标应用程序显示的第一页面中，确定用于进行点击交互的N个触发区域；特征处理单元，配置为根据所述N个触发区域，确定所述第一页面的页面特征；模型调用单元，配置为将所述页面特征输入强化学习模型，使其输出与动作空间相对应的动作决策信息，所述动作空间包括若干预设动作以及分别用于触发所述N个触发区域的N个触发动作；动作执行单元，配置为根据所述动作决策信息，对所述第一页面执行目标动作，使所述终端显示第二页面；奖励计算单元，配置为获取在所述第一页面中执行所述目标动作的环境奖励，其中所述环境奖励在所述第二页面是所述终端中新出现页面时的取值，大于所述第二页面不是所述终端中新出现页面时的取值，所述环境奖励用于更新所述强化学习模型。

本说明书提供了一种计算机可读非易失性存储介质，所述存储介质存储有计算机程序，所述计算机程序被处理器执行时实现上述的基于强化学习的应用程序巡检方法。

本说明书实施例中采用的技术方案能够达到以下有益效果中的至少一项：可以自动识别终端在当前时刻下对目标应用程序显示的页面中能够用于进行点击交互的N个触发区域，利用强化学习模型实现对终端在当前时刻下对目标应用程序显示的页面提供更好的动作决策信息，使得后续过程中根据该动作决策信息在该页面中执行目标动作后，能够尽可能的促使终端对目标应用程序显示未曾显示过的页面，从而能够促使终端尽可能多的探索并显示由目标应用程序提供的全部页面，同时尽可能的降低对目标应用程序提供的相同页面重复进行巡检的频次，有利于更加快速、高效且全面的完成对目标应用程序进行巡检。

附图说明

为了更清楚地说明本说明书实施例的技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其它的附图。

图1为本说明书实施例中示例性提供的强化学习系统的示意图；

图2为本说明书实施例中所提供技术方案的系统架构图；

图3为本说明书实施例中提供的一种基于强化学习的应用程序巡检方法的流程图；

图4为本说明书实施例中提供的一种基于强化学习的应用程序巡检装置的示意图；

图 5 为本说明书实施例中提供的一种计算机系统的结构示意图。

具体实施方式

为使本说明书的目的、技术方案和优点更加清楚，下面将结合本说明书具体实施例及相应的附图对本说明书技术方案进行清楚、完整地描述。显然，所描述的实施例仅是本说明书一部分实施例，而不是全部的实施例。基于本说明书中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本说明书保护的范围。

单个应用程序通常会提供大量能够显示于终端的页面，前述应用程序可以是能够直接运行于终端的移动应用程序，或者也可以是依赖其它移动应用程序运行的小程序。应用程序提供的单个页面中通常可以包括多个元素，前述的元素通常可以包括字符、图像或者其结合。当某个页面被显示于终端时，用户可以对所显示页面中的元素执行触发操作（例如点击操作），使终端显示被触发的元素所关联的其它页面，进而完成用户期望完成的事务。

当被触发的元素并未正确关联期望关联的其它页面时，终端则可能继续显示当前页面或者直接终止运行相关应用程序，从而无法完成用户期望完成的事务。因此，需要通过对应用程序进行巡检，以期发现其提供的大量页面内容中未正确关联其它页面的元素。

如前所述，通过人工干预的方式完成对应用程序进行巡检时，需要由测试人员来识别并点击应用程序所提供页面中能够用于进行点击交互的触发区域（即应用程序所提供页面中能够用于进行点击交互的元素），存在如下示例的诸多问题：识别并点击触发区域的速度相对较慢，难以快速完成对应用程序提供的大量页面进行巡检；受测试人员的主观因素影响，应用程序所提供页面中的部分触发区域可能被遗漏而未被正确点击，导致未能对应用程序提供的部分页面进行巡检；极易发生对应用程序提供的相同页面重复进行巡检。

鉴于此，本说明书实施例中至少提供了一种基于强化学习的应用程序巡检方法、装置、存储介质和计算机系统，可以自动识别终端在当前时刻下对目标应用程序显示的页面中能够用于进行点击交互的 N 个触发区域，利用强化学习模型实现对该页面提供更好的动作决策信息，使得后续过程中根据该动作决策信息在该页面中执行目标动作后，能够尽可能的促使终端对目标应用程序显示未曾显示过的页面，从而能够促使终端尽可能多的探索并显示由目标应用程序提供的全部页面，同时尽可能的降低对目标应用程序提供的相同页面重复进行巡检的频次，有利于更加快速、高效且全面的完成对目标应用程序进行巡检。

强化学习是机器学习中的一个重要领域，强调如何基于环境而行动，以取得最大化的预期利益。强化学习是除了监督学习和非监督学习之外的第三种基本的机器学习方法。与监督学习不同的是，强化学习不需要带标签的输入输出对。其关注点在于对未知领域的寻

找探索与对已有知识的利用之间的平衡。

图 1 为本说明书实施例中示例性提供的强化学习系统的示意图。如图 1 所示，强化学习系统可以包括智能体和执行环境，智能体通过与执行环境的交互及反馈，不断进行学习，优化其策略。具体而言，智能体可以观察并获得执行环境的状态(state)，根据一定的策略对执行环境的状态确定需要提供的动作(action)决策信息。基于动作决策信息选择的动作被作用于执行环境，会改变执行环境的状态，同时产生一个反馈给智能体，该反馈又称为环境奖励 (reward)。接着，智能体可以根据获得的环境奖励来判断之前提供的动作决策信息是否正确，策略是否需要调整，进而更新其策略。总而言之，智能体通过反复不断地观察状态、确定所需提供的动作决策信息、收到反馈，使得智能体可以不断更新其策略，最终目标是能够学习到一个策略，使得获得的环境奖励累积最大化。智能体的策略得到充分的更新后，智能体则能够在不同的状态下，提供更好的动作决策信息。

前述的策略是从执行环境的状态到动作决策信息的映射。其中，动作决策信息可以包括确定性的动作，或者可以包括动作空间中所包括的多个动作各自对应的执行概率。

图 2 为本说明书实施例中提供的一种基于强化学习的应用程序巡检方法的流程图。该方法示例性描述了对目标应用程序进行巡检的过程。其中如前所述，该目标应用程序可以是能够独立运行于终端中的移动应用程序，或者该目标应用程序也可以是依赖于某个移动应用程序运行的小程序。前述的终端可以包括但不限于个人计算机、手机以及平板电脑等。

该方法可以由任何具有计算/处理能力的装置、设备、平台或设备集群执行。运行目标应用程序的终端对应强化学习系统中的执行环境。对用于执行该方法的实体而言，可以集成在执行环境中，或者作为智能体的组成部分而集成在包括强化学习模型的智能体中，或者被划分为多个模块，部分模块集成在执行环境中，而其余模块则集成在智能体中。

参见图 2 所示，该方法可以包括但不限于如下步骤 S201~步骤 S215 中的部分或全部。

步骤 S201，从终端对目标应用程序显示的第一页面中，确定用于进行点击交互的 N 个触发区域。

当需要对某个目标应用程序进行巡检时，测试人员可以向终端提供该目标应用程序对应的 ULR scheme，使终端运行该目标应用程序并显示由该目标应用程序提供的页面。参照图 3 所示，第一页面是指利用终端对目标应用程序进行巡检的过程中，终端在当前时刻下向用户真实呈现的当前页面。当终端对由目标应用程序提供的任意页面例如页面 P 进行显示的过程中，终端通过相应的显示组件向用户真实呈现的当前页面（即第一页面），通常并不仅仅包括页面 P，还可能包括由终端提供的框架性元素，例如终端内置电池的电量

指示图标、时间信息、终端所接入网络对应的状态指示图标等等，其中框架性元素通常并不支持点击交互。因此而不难理解，前述步骤 S201 中所述的用于进行点击交互的 N 个触发区域，实质上对应由目标应用程序提供的某个页面中所包括的 N 个元素。

在一种可能的实施方式中，可以通过预先训练的目标检测模型来处理第一页面，实现从该第一页面中标记出能够用于进行点击交互的 N 个触发区域。通常的，可以采用对预训练模型进行微调的方式来获得用于处理第一页面的目标检测模型；在一个较为典型的示例中，可以将 SOTA（state-of-the-art）模型 DINO 作为预训练模型，选择对应的训练数据集来对 DINO 进行微调，目标是通过进行微调后的 DINO 处理终端对任意应用程序显示的页面时，均能够较好的识别并标记出该页面中能够用于进行点击交互的全部触发区域。

在另一种可能的实施方式中，可通过相应的页面分析工具获取第一页面的页面转储信息，例如从终端的内存中将第一页面的相关信息转储到 dump.xml 文件，dump.xml 文件则包括第一页面的页面转储信息。页面转储信息中包括由目标应用程序提供的某个页面所包括的 N 个元素各自对应在第一页面中的位置信息，进而可以根据 N 个元素各自对应在第一页面中的位置信息，从第一页面中标记出用于进行点击交互的 N 个触发区域。

需要说明的是，N 的取值是从第一页面中确定的触发区域的具体数量而非预设数值。

步骤 S203，根据 N 个触发区域，确定第一页面的页面特征。

可以对 N 个触发区域各自对应在第一页面中的图像区域进行编码，获得 N 个触发区域各自对应的图像特征；示例性的，可以通过 CPID（Contrastive Language-Image Pre-training）的编码器（encoder）处理该 N 个触发区域各自对应在第一页面中的图像区域，获得该 N 个触发区域各自对应的图像特征（记为 Image embedding）。类似的，可以对该 N 个触发区域各自对应在第一页面中的位置信息进行编码，获得该 N 个触发区域各自对应的位置特征（记为 Position embedding）。进而，可以通过合并该 N 个触发区域各自对应的 Image embedding 和 Position embedding，获得该第一页面的页面特征（记为 Page embedding）。

单个触发区域对应在第一页面中的图像区域通常呈矩形。与之相应的是，单个触发区域对应在第一页面中的位置信息，可以利用该触发区域对应的图像区域在图像坐标系的横坐标极小值（x-min）、纵坐标极小值（y-min）、横坐标极大值（x-max）以及纵坐标极大值（y-max）表达。通常的，x-min、y-min、x-max、y-max 的数值可能相对较大，因此可以首先对其进行标准化处理，将 x-min、y-min、x-max、y-max 的值映射到 0~1 之间，然后再对标准化处理后的 x-min、y-min、x-max、y-max 进行编码，获得相应的 Position embedding。

前文虽然示例性描述了获得 N 个触发区域各自对应的 Image embedding 和 Position

embedding 的过程，然而可以理解的前述过程仅仅是示例性，例如还可以通过其它编码器来处理触发区域对应在第一页面中的图像区域，仅需要确保所获得的 Image embedding 以及 Position embedding，能够较为准确的表达对应的图像区域和位置信息即可。

前述 Image embedding 和 Position embedding 通常可以是维数相同的特征向量。在此基础上，可以首先对对应相同触发区域的 Image embedding 和 Position embedding 进行合并（例如相加或者依次拼接），获得 N 个触发区域各自对应的嵌入特征；进而可以利用该 N 个触发区域各自对应的嵌入特征组成第一页面的 Page embedding。

步骤 S205，将第一页面的页面特征输入强化学习模型，使其输出与动作空间对应的动作决策信息，动作空间包括若干预设动作以及分别用于触发 N 个触发区域的 N 个触发动作。

参照图 3 所示，用于实现该方法的实体，例如基于强化学习的应用巡检装置，可以通过调用强化学习模型来处理终端所显示的当前页面，获得当前页面的动作决策信息。

强化学习模型属于强化学习系统中的智能体，第一页面的页面特征可以理解为智能体观测到的执行环境的状态。强化学习模型所采用的策略可以包括典型的深度确定性策略梯度（Deep Deterministic Policy Gradient，DDPG）算法、SAC（Soft Actor-Critic）算法、Q-learning 算法或者 DQN（Deep Q-Network）等等，其中 SAC 算法在 DDPG 算法的基础上引入了信息熵作为奖励的一部分，可以提高了探索效率，而且支持在较为复杂的执行环境中高效稳定地学习，因此本说明书实施例中，强化学习模型所采用的策略优选 SAC 算法。

因可能对第一页面执行的各个触发动作是离散动作，强化学习模型的动作空间可以设计为分类分布。除此之外，第一页面所支持的部分触发动作，例如向上滑动屏幕、向下滑动屏幕以及向左滑动屏幕等等，可能并不依赖第一页面所包括的 N 个触发区域；其中向上滑动屏幕和向下滑动屏幕通常会导致终端显示不完全相同于第一页面的其它页面，而向左滑动屏幕通常可以使终端由显示第一页面返回到已显示过的前一页面。在此基础上，假设终端对目标应用程序显示的任意页面中至多包括 M 个触发区域，以及假设终端对目标应用程序显示的任意页面还支持 K 个并不依赖于触发区域的触发动作，动作空间中则可以包括 M 个用于触发 M 个触发区域的触发动作以及 K 个预设动作（不依赖触发区域的触发动作）。

在一种可能的实施方式中，动作空间中的前 M 个触发动作对应可能存在的 M 个触发区域，动作空间中的后 K 个触发动作则对应前述的 K 个预设动作。相应的，可以根据 N 个触发区域各自对应在第一页面中的位置信息对该 N 个触发区域进行排序，例如结合前文示例的触发区域的位置信息所包括的 x-min，按照 N 个触发区域各自对应的 x-min 由小到大的顺序对 N 个触发区域进行排序；如果存在至少两个触发区域各自对应的 x-min 大小相

同，则按照 y-min 由小到大的顺序对前述至少两个触发区域进行排序。进而，对于进行排序处理后的 N 个触发区域中任意的第 i 个触发区域，其对应动作空间下的第 i 个触发动作。

强化学习模型采用相应的策略来处理第一页面的页面 Page embedding 时，可以计算出动作空间下的 M+K 个触发动作各自对应的执行概率。与之相应的，强化学习模型输出的动作决策信息，可以用于指示该动作空间中的 M+K 个触发动作分别对应的执行概率；或者，强化学习模型可以从 M+K 个触发动作中，进一步的选择一个执行概率相对较大的触发动作作为期望执行的目标动作，其输出的动作决策信息则可以用于指示该目标动作。

在一个更为具体的示例中，动作空间下例如具体包括 1~M+K 等 M+K 个动作序号，其中前 N 个动作序号与依次排列的 N 个触发区域一一对应，而后 K 个动作序号与依次排列的 K 个预设动作一一对应。强化学习模型可以针对第一页面的 Page embedding，输出 M+K 个动作序号各自对应的执行概率，或者从 M+K 个序号中选择一个目标动作序号并输出。

步骤 S207，根据动作决策信息，对第一页面执行目标动作，使终端显示第二页面。

当动作决策信息指示动作空间下的 M+K 个触发动作各自对应的执行概率时，可以从该 M+K 个动作中选择一个执行概率相对较大的触发动作作为目标动作。

当目标动作属于并不依赖触发区域的预设动作的情况下，直接执行目标动作即可。

当目标动作属于与 N 个触发区域相对应的 N 个触发动作的情况下，可以根据目标动作所对应的目标触发区域在第一页面中的位置信息执行目标动作，即根据目标动作所对应的目标触发区域在第一页面中的位置信息，触发目标触发区域对应在第一页面中的图像区域。

对目标应用程序进行巡检的过程中，需要使终端能够尽可能全面的访问由目标应用程序提供的全部页面，同时能够尽可能的避免重复访问相同页面，以利更加快速、全面且高效的完成对应用程序进行巡检。为了达到此目的，我们期望完成在第一页面中执行目标动作后，终端所显示的第二页面是新出现页面而并非已出现页面。

为了实现判断第二页面是否为新出现页面，在一种可能的实施方式中，在步骤 S209，获取用于表征第二页面的页面编码，并根据页面编码确定第二页面是否为新出现页面。

在一种可能的实施方式中，开始对目标应用程序进行巡检后，可以维护一个编码集合。该编码集合中包括预设参考页面所对应的页面编码，以及终端对目标应用程序显示的全部页面各自对应的页面编码；其中该参考页面是指开始对目标应用程序进行巡检前，具体如测试人员触发终端开始运行目标应用程序之前，终端所显示的最后一个页面。

与之相应的，可以通过相应的页面分析工具获取第二页面的页面转储信息，其中包括第二页面的框架信息，框架信息中包括第二页面中的多个元素各自对应在第二页面中的位

置信息。该多个元素不仅包括由目标应用程序提供的页面所包括元素，还可能包括由终端提供的框架性元素，框架性元素的位置信息可能对页面相似性判断造成负面影响，因此可以从框架信息中剔除由终端提供的框架性元素的位置信息。剔除框架性元素的位置信息之后，框架信息中可能仍然包括由目标应用程序提供的大量元素的位置信息，为了更加有利于高效实施相似性判断，可以继续剔除未携带字符的元素所对应的位置信息，仅保留第一页面中由目标应用程序提供并且携带字符的至少一个目标元素各自对应的位置信息。进而可以根据前述至少一个目标元素各自对应的位置信息确定第二页面所对应的页面编码。示例性的，可以按照某种顺序对至少一个目标元素各自对应的位置信息进行拼接，通过对过长的拼接结果进行裁剪，对过短的拼接结果进行填充，可以获得具有预定长度的页面编码。

获得第二页面的页面编码后，则可以继续查询编码集合中是否包括第二页面的页面编码，如果是确定第二页面并非新出现页面。

前述判断第二页面是否为新出现页面的过程是示例性的。例如，也可以直接计算第二页面分别与预设参考页面以及终端对目标应用程序已显示的各个页面间的相似度，当存在至少一个相似度大于相似度阈值的情况下，则确定第二页面并非新出现页面。

步骤 S211，获取在第一页面中执行目标动作的环境奖励，其中环境奖励在第二页面是终端中新出现页面时的取值，大于第二页面不是终端中新出现页面时的取值，环境奖励用于更新强化学习模型。

结合强化学习模型的训练及工作原理，此处设置环境奖励在第二页面是终端中新出现页面时的取值，大于第二页面不是终端中新出现页面时的取值，目的在于后续过程中利用该环境奖励更新强化学习模型后，更新后的强化学习模型能够对终端所显示的页面提供更好的动作决策信息，尽可能的促使后续完成根据相应的动作决策信息在终端所显示页面中执行目标动作后，终端能够对由目标应用程序提供且未曾对其进行显示过的页面进行显示。

参照前文所述，目标应用程序提供的页面中，可能存在部分元素未能正确关联期望关联的其它。当目标触发区域对应的元素并未正确关联期望关联的其它页面时，在第一页面中对目标触发区域执行触发动作后，终端可能继续显示第一页面或者终止运行目标应用程序；其中如果终止运行目标应用程序的情况下，终端则可能显示前述的参考页面。

当目标触发区域对应的元素并未正确关联期望关联的其它页面时，意味着无法通过目标触发区域到达新的页面并继续对目标应用程序进行巡检，因此希望后续过程中利用该环境奖励更新强化学习模型后，更新后的强化学习模型能够对终端所显示的页面提供更好的动作决策信息，尽可能避免继续对该目标触发区域执行触发动作。基于此，当第二页面相

似于终端中已出现的第一页面或者预设参考页面时，环境奖励的取值为第一预设数值；当第二页面相似于终端中除第一页面和预设参考页面外的其它已出现页面时，环境奖励的取值为第二预设数值；其中第一预设数值小于第二预设数值。

在一个更为具体的示例中，当第二页面是新出现页面时，环境奖励的取值例如可以为预设数值 1。当第二页面并非新出现页面时，如果第二页面相似于第一页面或者预设参考页面，那么环境奖励的取值例如可以为预设数值-1，否则环境奖励的取值例如可以为 0。

为了实现在对目标应用程序进行巡检的过程中，基于强化学习模型输出的动作决策信息尽可能的促使终端对目标应用程序显示未曾显示过的页面，可选地，在步骤 S213，获取在第一页面中执行目标动作的探索奖励，探索奖励用于更新强化学习模型。

探索奖励作为强化学习模型的内在奖励，可通过包含随机网络蒸馏（Random Network Distillation，RND）算法或内在好奇模型（Intrinsic curiosity module，ICM）算法处理第二页面的页面特征得到。以 RND 算法为例，参照图 3 所示，RND 算法包括一个随机初始化之后即固定不变的目标网络（target network）以及一个可训练的预测网络（predictor network），用于实现该方法的实体例如基于强化学习的应用巡检装置，可以通过调用该 RND 算法来处理终端所显示的当前页面（即第二页面），获得 target network 和 predictor network 对第二页面分别输出的处理结果，进而可以将两个处理结果间的余弦距离作为探索奖励。

参照前文所述，当第二页面相似于终端中已出现的第一页面或者预设参考页面时，说明第一页面中所包括的由目标应用程序提供的目标页面存在漏洞，例如目标页面中与目标触发区域对应的元素可能并未正确关联期望关联的其它页面。基于此，为了提醒测试人员目标页面中存在漏洞，还可以接着执行如下步骤 S215，当第二页面相似于终端中已出现的第一页面或者预设参考页面时，生成与第一页面对应的错误提示信息。

示例性的，错误提示信息中可以包括第一页面中由目标应用程序提供的目标页面的页面标识。当目标触发动作与 N 个目标触发区域中的某个目标触发区域对应时，错误提示信息中还可以包括用于指示目标触发区域所对应元素的指示信息，例如还可以包括该目标触发区域对应在第一页面中的位置信息或者目标触发区域所对应元素的元素标识。

前述步骤 S201~步骤 S215 描述了对目标应用程序提供的一个页面进行一次巡检的过程，具体技术场景中可以通过多次执行前述过程，自动完成对目标应用程序进行全面巡检。

参照前文所述，智能体通过反复不断地观察状态、确定所需提供的动作决策信息、收到反馈，使得智能体可以不断更新其策略，最终目标是能够学习到一个策略，使得获得的环境奖励累积最大化。智能体的策略得到充分的更新后，智能体能够在不同的状态下，提

供更好的动作决策信息。因此，在前述步骤 S201~步骤 S215 的基础上，还可以通过如下步骤 S217 和步骤 S219 来更新强化学习模型。

步骤 S217，向预先设置的回放缓冲区加入训练样本，训练样本包括第一页面的页面特征、目标动作、奖励信息和第二页面的页面特征，其中奖励信息中至少包括环境奖励。

在一些实施例中，奖励信息中还可以包括在第一页面中执行目标动作的探索奖励。

步骤 S219，当回放缓存区已加入的训练样本达到预设阈值时，利用存储于回放缓冲区的训练样本训练强化学习模型。其中，当完成利用存储于回放缓冲区的训练样本训练强化学习模型后，可重新初始化回放缓存区，以利将新的训练样本加入回放缓冲区。

在一些实施例中，当奖励信息中还包括在第一页面中执行目标动作的探索奖励的情况下，还可以利用回放缓存区中的各个训练样本所分别包括的第二页面，对应的更新用于输出前述探索奖励的算法，例如对应更新 RND 算法中的预测网络。

通过对强化学习模型进行多轮更新后，强化学习模型则可以根据终端在当前时刻下对目标应用程序显示的当前页面所对应的页面特征，提供优秀的动作决策信息来促使终端对目标应用程序提供的大量页面进行显示，从而完成对目标应用程序进行巡检。

与前述方法实施例基于相同的构思，本说明书实施例中提供了一种基于强化学习模型的应用程序巡检装置 400。参见图 4 所示，所述巡检装置包括：目标检测单元 401，配置为从终端对目标应用程序显示的第一页面中，确定用于进行点击交互的 N 个触发区域；特征处理单元 403，配置为根据所述 N 个触发区域，确定所述第一页面的页面特征；模型调用单元 405，配置为将所述页面特征输入强化学习模型，使其输出与动作空间相对应的动作决策信息，所述动作空间包括若干预设动作以及分别用于触发所述 N 个触发区域的 N 个触发动作；动作执行单元 407，配置为根据所述动作决策信息，对所述第一页面执行目标动作，使所述终端显示第二页面；奖励计算单元 409，配置为获取在所述第一页面中执行所述目标动作的环境奖励，其中所述环境奖励在所述第二页面是所述终端中新出现页面时的取值，大于所述第二页面不是所述终端中新出现页面时的取值，所述环境奖励用于更新所述强化学习模型。

为了描述的方便，描述以上装置时以功能分为各种单元/模块分别描述。当然，在实施本说明书时可以把各模块/单元的功能在同一个或多个软件和/或硬件中实现。

The specification also provides a computer-readable non-transitory storage medium, wherein the storage medium stores a computer program, which, when executed by a processor, can be used to perform one or more steps of one or more methods described or illustrated herein

or provides functionality described or illustrated herein. Herein, the computer-readable non-transitory storage medium or media may include one or more semiconductor-based or other integrated circuits (ICs) (such, as for example, field-programmable gate arrays (FPGAs) or application-specific ICs (ASICs)), hard disk drives (HDDs), hybrid hard drives (HHDs), optical discs, optical disc drives (ODDs), magneto-optical discs, magneto-optical drives, floppy diskettes, floppy disk drives (FDDs), magnetic tapes, solid-state drives (SSDs), RAM-drives, SECURE DIGITAL cards or drives, any other suitable computer-readable non-transitory storage media, or any suitable combination of two or more of these, where appropriate. A computer-readable non-transitory storage medium may be volatile, non-volatile, or a combination of volatile and non-volatile, where appropriate.

The specification also provides a computer system. FIG. 5 illustrates an example computer system 1000. In particular embodiments, one or more computer systems 1000 perform one or more steps of one or more methods described or illustrated herein. In particular embodiments, one or more computer systems 1000 provide functionality described or illustrated herein. In particular embodiments, software running on one or more computer systems 1000 performs one or more steps of one or more methods described or illustrated herein or provides functionality described or illustrated herein. Particular embodiments include one or more portions of one or more computer systems 1000. Herein, reference to a computer system may encompass a computing device, and vice versa, where appropriate. Moreover, reference to a computer system may encompass one or more computer systems, where appropriate.

This disclosure contemplates any suitable number of computer systems 1000. This disclosure contemplates computer system 1000 taking any suitable physical form. As example and not by way of limitation, computer system 1000 may be an embedded computer system, a system-on-chip (SOC), a single-board computer system (SBC) (such as, for example, a computer-on-module (COM) or system-on-module (SOM)), a desktop computer system, a laptop or notebook computer system, an interactive kiosk, a mainframe, a mesh of computer systems, a mobile telephone, a personal digital assistant (PDA), a server, a tablet computer system, or a combination of two or more of these. Where appropriate, computer system 1000 may include one or more computer systems 1000; be unitary or distributed; span multiple locations; span multiple machines; span multiple data centers; or reside in a cloud, which may include one or more cloud components in one or more networks. Where appropriate, one or more computer systems 1000 may perform without substantial spatial or temporal limitation one or more steps of one or more methods described or illustrated herein. As an example and not by way of limitation, one or more computer systems 1000 may perform in real time or in batch mode one or

more steps of one or more methods described or illustrated herein. One or more computer systems 1000 may perform at different times or at different locations one or more steps of one or more methods described or illustrated herein, where appropriate.

In particular embodiments, computer system 1000 includes a processor 1002, memory 1004, storage 1006, an input/output (I/O) interface 1008, a communication interface 1010, and a bus 1012. Although this disclosure describes and illustrates a particular computer system having a particular number of particular components in a particular arrangement, this disclosure contemplates any suitable computer system having any suitable number of any suitable components in any suitable arrangement.

In particular embodiments, processor 1002 includes hardware for executing instructions, such as those making up a computer program. As an example and not by way of limitation, to execute instructions, processor 1002 may retrieve (or fetch) the instructions from an internal register, an internal cache, memory 1004, or storage 1006; decode and execute them; and then write one or more results to an internal register, an internal cache, memory 1004, or storage 1006. In particular embodiments, processor 1002 may include one or more internal caches for data, instructions, or addresses. This disclosure contemplates processor 1002 including any suitable number of any suitable internal caches, where appropriate. As an example and not by way of limitation, processor 1002 may include one or more instruction caches, one or more data caches, and one or more translation lookaside buffers (TLB s). Instructions in the instruction caches may be copies of instructions in memory 1004 or storage 1006, and the instruction caches may speed up retrieval of those instructions by processor 1002. Data in the data caches may be copies of data of 80 in memory 1004 or storage 1006 for instructions executing at processor 1002 to operate on; the results of previous instructions executed at processor 1002 for access by subsequent instructions executing at processor 1002 or for writing to memory 1004 or storage 1006; or other suitable data. The data caches may speed up read or write operations by processor 1002. The TLBs may speed up virtual-address translation for processor 1002. In particular embodiments, processor 1002 may include one or more internal registers for data, instructions, or addresses. This disclosure contemplates processor 1002 including any suitable number of any suitable internal registers, where appropriate. Where appropriate, processor 1002 may include one or more arithmetic logic units (ALUs); be a multi-core processor; or include one or more processors 1002. Although this disclosure describes and illustrates a particular processor, this disclosure contemplates any suitable processor.

In particular embodiments, memory 1004 includes main memory for storing instructions for processor 1002 to execute or data for processor 1002 to operate on. As an example and not by

way of limitation, computer system 1000 may load instructions from storage 1006 or another source (such as, for example, another computer system 1000) to memory 1004. Processor 1002 may then load the instructions from memory 1004 to an internal register or internal cache. To execute the instructions, processor 1002 may retrieve the instructions from the internal register or internal cache and decode them. During or after execution of the instructions, processor 1002 may write one or more results (which may be intermediate or final results) to the internal register or internal cache. Processor 1002 may then write one or more of those results to memory 1004. In particular embodiments, processor 1002 executes only instructions in one or more internal registers or internal caches or in memory 1004 (as opposed to storage 1006 or elsewhere) and operates only on data in one or more internal registers or internal caches or in memory 1004 (as opposed to storage 1006 or elsewhere). One or more memory buses (which may each include an address bus and a data bus) may couple processor 1002 to memory 1004. Bus 1012 may include one or more memory buses, as described below. In particular embodiments, one or more memory management units (MMUs) reside between processor 1002 and memory 1004 and facilitate accesses to memory 1004 requested by processor 1002. In particular embodiments, memory 1004 includes random access memory (RAM). This RAM may be volatile memory, where appropriate. Where appropriate, this RAM may be dynamic RAM (DRAM) or static RAM (SRAM). Moreover, where appropriate, this RAM may be single-ported or multi-ported RAM. This disclosure contemplates any suitable RAM. Memory 1004 may include one or more memories 1004, where appropriate. Although this disclosure describes and illustrates particular memory, this disclosure contemplates any suitable memory.

In particular embodiments, storage 1006 includes mass storage for data or instructions. As an example and not by way of limitation, storage 1006 may include a hard disk drive (HDD), a floppy disk drive, flash memory, an optical disc, a magneto-optical disc, magnetic tape, or a Universal Serial Bus (USB) drive or a combination of two or more of these. Storage 1006 may include removable or non-removable (or fixed) media, where appropriate. Storage 1006 may be internal or external to computer system 1000, where appropriate. In particular embodiments, storage 1006 is non-volatile, solid-state memory. In particular embodiments, storage 1006 includes read-only memory (ROM). Where appropriate, this ROM may be mask-programmed ROM, programmable ROM (PROM), erasable PROM (EPROM), electrically erasable PROM (EEPROM), electrically alterable ROM (EAROM), or flash memory or a combination of two or more of these. This disclosure contemplates mass storage 1006 taking any suitable physical form. Storage 1006 may include one or more storage control units facilitating communication between processor 1002 and storage 1006, where appropriate. Where appropriate, storage 1006 may

include one or more storages 1006. Although this disclosure describes and illustrates particular storage, this disclosure contemplates any suitable storage.

In particular embodiments, I/O interface 1008 includes hardware, software, or both, providing one or more interfaces for communication between computer system 1000 and one or more I/O devices. Computer system 1000 may include one or more of these I/O devices, where appropriate. One or more of these I/O devices may enable communication between a person and computer system 1000. As an example and not by way of limitation, an I/O device may include a keyboard, keypad, microphone, monitor, mouse, printer, scanner, speaker, still camera, stylus, tablet, touch screen, trackball, video camera, another suitable I/O device or a combination of two or more of these. An I/O device may include one or more sensors. This disclosure contemplates any suitable I/O devices and any suitable I/O interfaces 1008 for them. Where appropriate, I/O interface 1008 may include one or more device or software drivers enabling processor 1002 to drive one or more of these I/O devices. I/O interface 1008 may include one or more I/O interfaces 1008, where appropriate. Although this disclosure describes and illustrates a particular I/O interface, this disclosure contemplates any suitable I/O interface.

In particular embodiments, communication interface 1010 includes hardware, software, or both providing one or more interfaces for communication (such as, for example, packet-based communication) between computer system 1000 and one or more other computer systems 1000 or one or more networks. As an example and not by way of limitation, communication interface 1010 may include a network interface controller (NIC) or network adapter for communicating with an Ethernet or other wire-based network or a wireless NIC (WNIC) or wireless adapter for communicating with a wireless network, such as a WI-FI network. This disclosure contemplates any suitable network and any suitable communication interface 1010 for it. As an example and not by way of limitation, computer system 1000 may communicate with an ad hoc network, a personal area network (PAN), a local area network (LAN), a wide area network (WAN), a metropolitan area network (MAN), or one or more portions of the Internet or a combination of two or more of these. One or more portions of one or more of these networks may be wired or wireless. As an example, computer system 1000 may communicate with a wireless PAN (WPAN) (such as, for example, a BLUETOOTH WPAN), a WI-FI network, a WI-MAX network, a cellular telephone network (such as, for example, a Global System for Mobile Communications (GSM) network), or other suitable wireless network or a combination of two or more of these. Computer system 1000 may include any suitable communication interface 1010 for any of these networks, where appropriate. Communication interface 1010 may include one or more communication interfaces 1010, where appropriate. Although this disclosure describes and

illustrates a particular communication interface, this disclosure contemplates any suitable communication interface.

In particular embodiments, bus 1012 includes hardware, software, or both coupling components of computer system 1000 to each other. As an example and not by way of limitation, bus 1012 may include an Accelerated Graphics Port (AGP) or other graphics bus, an Enhanced Industry Standard Architecture (EISA) bus, a front-side bus (FSB), a HYPERTRANSPORT (HT) interconnect, an Industry Standard Architecture (ISA) bus, an INFINIBAND interconnect, a low-pin-count (LPC) bus, a memory bus, a Micro Channel Architecture (MCA) bus, a Peripheral Component Interconnect (PCI) bus, a PCI-Express (PCIe) bus, a serial advanced technology attachment (SATA) bus, a Video Electronics Standards Association local (VLB) bus, or another suitable bus or a combination of two or more of these. Bus 1012 may include one or more buses 1012, where appropriate. Although this disclosure describes and illustrates a particular bus, this disclosure contemplates any suitable bus or interconnect.

还需要说明的是，术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含，从而使得包括一系列要素的过程、方法、商品或者设备不仅包括那些要素，而且还包括没有明确列出的其他要素，或者是还包括为这种过程、方法、商品或者设备所固有的要素。在没有更多限制的情况下，由语句“包括一个……”限定的要素，并不排除在包括所述要素的过程、方法、商品或者设备中还存在另外的相同要素。

本说明书中的各个实施例均采用递进的方式描述，各个实施例之间相同相似的部分互相参见即可，每个实施例重点说明的都是与其他实施例的不同之处。尤其，对于系统实施例而言，由于其基本相似于方法实施例，所以描述的比较简单，相关之处参见方法实施例的部分说明即可。

以上所述仅为本说明书的实施例而已，并不用于限制本说明书。对于本领域技术人员来说，本说明书可以有各种更改和变化。凡在本说明书的精神和原理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权利要求范围之内。

权　利　要　求　书

1．一种基于强化学习的应用程序巡检方法，所述方法包括：

从终端对目标应用程序显示的第一页面中，确定用于进行点击交互的 N 个触发区域；

根据所述 N 个触发区域，确定所述第一页面的页面特征；

将所述页面特征输入强化学习模型，使其输出与动作空间相对应的动作决策信息，所述动作空间包括若干预设动作以及分别用于触发所述 N 个触发区域的 N 个触发动作；

根据所述动作决策信息，对所述第一页面执行目标动作，使所述终端显示第二页面；

获取在所述第一页面中执行所述目标动作的环境奖励，其中所述环境奖励在所述第二页面是所述终端中新出现页面时的取值，大于所述第二页面不是所述终端中新出现页面时的取值，所述环境奖励用于更新所述强化学习模型。

2．根据权利要求 1 所述的方法，所述从终端对目标应用程序显示的第一页面中，确定用于进行点击交互的 N 个触发区域，包括：将所述第一页面输入预先训练的深度学习模型，使所述深度学习模型在所述第一页面中检测出用于进行点击交互的 N 个触发区域。

3．根据权利要求 1 所述的方法，所述根据所述 N 个触发区域，确定所述第一页面的页面特征，包括：

对所述 N 个触发区域各自对应在所述第一页面中的图像区域进行编码，获得所述 N 个触发区域各自对应的图像特征；

对所述 N 个触发区域各自对应在所述第一页面中的位置信息进行编码，获得所述 N 个触发区域各自对应的位置特征；

合并所述 N 个触发区域各自对应的图像特征和位置特征，获得所述页面特征。

4．根据权利要求 1 所述的方法，所述动作决策信息用于指示所述动作空间中的各个动作分别对应的执行概率，或者，所述动作决策信息用于指示期望执行的所述目标动作。

5．根据权利要求 1 所述的方法，所述根据所述动作决策信息，对所述第一页面执行目标动作，包括：当所述目标动作不属于所述若干预设动作的情况下，根据所述目标动作所对应的目标触发区域在所述第一页面中的位置信息，执行所述目标动作。

6．根据权利要求 1 所述的方法，当所述第二页面不是新出现页面并且相似于所述第一页面或者预设参考页面时，所述环境奖励的取值为第一预设数值；当所述第二页面不是新出现页面并且并不相似于所述第一页面和所述预设参考页面时，所述环境奖励的取值为第二预设数值；其中所述第一预设数值小于所述第二预设数值。

7．根据权利要求 1 所述的方法，所述方法还包括：当所述第二页面相似于所述终端

中已出现的所述第一页面或者预设参考页面时，生成与所述第一页面对应的错误提示信息。

8．根据权利要求 1 所述的方法，所述方法还包括：获取用于表征所述第二页面的页面编码，根据所述页面编码确定所述第二页面是否为所述终端中新出现页面。

9．根据权利要求 8 所述的方法，所述获取用于表征所述第二页面的页面编码，包括：

获取所述第二页面在所述终端中所对应的页面转储信息；

从所述页面转储信息中提取由所述目标应用程序提供的内容信息；

从所述内容信息中确定出携带文字的若干目标元素各自对应的坐标信息；

根据所述若干目标元素各自对应的坐标信息，确定所述第二页面的页面编码。

10．根据权利要求 1 所述的方法，所述方法还包括：获取在所述第一页面中执行所述目标动作的探索奖励，所述探索奖励用于更新所述强化学习模型。

11．根据权利要求 1-10 中任一项所述的方法，所述方法还包括：

向回放缓冲区加入训练样本，所述训练样本包括所述第一页面的页面特征、所述目标动作、奖励信息和所述第二页面的页面特征，其中所述奖励信息中至少包括所述环境奖励；

当所述回放缓存区已加入的训练样本达到预设阈值时，利用存储于所述回放缓冲区的训练样本训练所述强化学习模型。

12．一种计算机系统，包括存储器、处理器及存储在所述存储器上并可在所述处理器上运行的计算机程序，所述处理器执行所述计算机程序时实现如下步骤：

从终端对目标应用程序显示的第一页面中，确定用于进行点击交互的 N 个触发区域；

根据所述 N 个触发区域，确定所述第一页面的页面特征；

将所述页面特征输入强化学习模型，使其输出与动作空间相对应的动作决策信息，所述动作空间包括若干预设动作以及分别用于触发所述 N 个触发区域的 N 个触发动作；

根据所述动作决策信息，对所述第一页面执行目标动作，使所述终端显示第二页面；

获取在所述第一页面中执行所述目标动作的环境奖励，其中所述环境奖励在所述第二页面是所述终端中新出现页面时的取值，大于所述第二页面不是所述终端中新出现页面时的取值，所述环境奖励用于更新所述强化学习模型。

13．根据权利要求 12 所述的计算机系统，所述从终端对目标应用程序显示的第一页面中，确定用于进行点击交互的 N 个触发区域，包括：将所述第一页面输入预先训练的深度学习模型，使所述深度学习模型在所述第一页面中检测出用于进行点击交互的 N 个触发区域。

14．根据权利要求 12 所述的计算机系统，所述根据所述 N 个触发区域，确定所述第

一页面的页面特征，包括：

对所述 N 个触发区域各自对应在所述第一页面中的图像区域进行编码，获得所述 N 个触发区域各自对应的图像特征；

对所述 N 个触发区域各自对应在所述第一页面中的位置信息进行编码，获得所述 N 个触发区域各自对应的位置特征；

合并所述 N 个触发区域各自对应的图像特征和位置特征，获得所述页面特征。

15．根据权利要求 12 所述的计算机系统，所述动作决策信息用于指示所述动作空间中的各个动作分别对应的执行概率，或者，所述动作决策信息用于指示期望执行的所述目标动作。

16．根据权利要求 12 所述的计算机系统，所述根据所述动作决策信息，对所述第一页面执行目标动作，包括：当所述目标动作不属于所述若干预设动作的情况下，根据所述目标动作所对应的目标触发区域在所述第一页面中的位置信息，执行所述目标动作。

17．根据权利要求 12 所述的计算机系统，当所述第二页面不是新出现页面并且相似于所述第一页面或者预设参考页面时，所述环境奖励的取值为第一预设数值；当所述第二页面不是新出现页面并且并不相似于所述第一页面和所述预设参考页面时，所述环境奖励的取值为第二预设数值；其中所述第一预设数值小于所述第二预设数值。

18．根据权利要求 12 所述的计算机系统，所述处理器执行所述计算机程序时还实现如下步骤：当所述第二页面相似于所述终端中已出现的所述第一页面或者预设参考页面时，生成与所述第一页面对应的错误提示信息。

19．一种基于强化学习的应用程序巡检装置，所述装置包括：

目标检测单元，配置为从终端对目标应用程序显示的第一页面中，确定用于进行点击交互的 N 个触发区域；

特征处理单元，配置为根据所述 N 个触发区域，确定所述第一页面的页面特征；

模型调用单元，配置为将所述页面特征输入强化学习模型，使其输出与动作空间相对应的动作决策信息，所述动作空间包括若干预设动作以及分别用于触发所述 N 个触发区域的 N 个触发动作；

动作执行单元，配置为根据所述动作决策信息，对所述第一页面执行目标动作，使所述终端显示第二页面；

奖励计算单元，配置为获取在所述第一页面中执行所述目标动作的环境奖励，其中所述环境奖励在所述第二页面是所述终端中新出现页面时的取值，大于所述第二页面不是所

述终端中新出现页面时的取值，所述环境奖励用于更新所述强化学习模型。

20．一种计算机可读非易失性存储介质，所述存储介质存储有计算机程序，所述计算机程序被处理器执行时实现上述权利要求 1-11 中任一项所述的方法。