



(12) 发明专利申请

(10) 申请公布号 CN 118689681 A

(43) 申请公布日 2024. 09. 24

(21) 申请号 202410213906.5

(22) 申请日 2024.02.27

(30) 优先权数据

10-2023-0038039 2023.03.23 KR

10-2023-0074808 2023.06.12 KR

(71) 申请人 三星电子株式会社

地址 韩国京畿道水原市

申请人 高丽大学校产学协力团

(72) 发明人 宋宜锡 金性范 金载训 金桢寅

方炳祐 李晟旻 李俊渊 李知润

郑在润

(74) 专利代理机构 北京铭硕知识产权代理有限公司

公司 11286

专利代理师 张川绪 史泉

(51) Int.Cl.

G06F 11/07 (2006.01)

G06F 16/18 (2019.01)

G06N 3/08 (2023.01)

G06N 3/045 (2023.01)

G06N 3/0475 (2023.01)

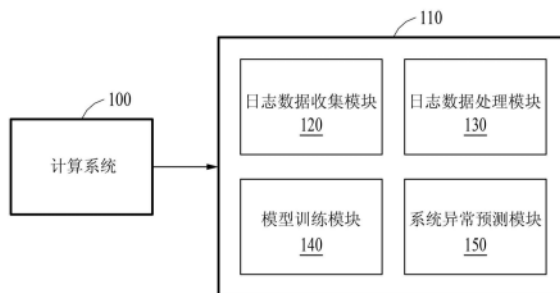
权利要求书2页 说明书17页 附图11页

(54) 发明名称

预测计算系统中的错误的方法和装置以及训练方法

(57) 摘要

公开了预测计算系统中的错误的方法和装置以及训练方法。错误预测方法包括：接收在计算系统的操作期间由计算系统生成的日志数据；将日志数据令牌化为令牌；将令牌输入到生成各个令牌的得分的鉴别器模型，每个得分对应于相应令牌是异常令牌的概率；基于所述得分确定异常得分；以及基于异常得分确定计算系统中未来发生错误的可能性。



1. 一种预测计算系统中的错误的方法,所述方法包括:
接收在计算系统的操作期间由计算系统生成的日志数据;
将日志数据令牌化为令牌;
将令牌输入到生成各个令牌的得分的鉴别器模型,每个得分对应于相应令牌是异常令牌的概率;
基于所述得分确定异常得分;以及
基于异常得分确定计算系统中未来发生错误的可能性。
2. 根据权利要求1所述的方法,还包括:对所述得分执行增强处理,其中,增强处理调整所述得分,使得所述得分的标准差增大。
3. 根据权利要求1所述的方法,其中,鉴别器模型是神经网络模型,神经网络模型被配置为基于相应令牌是异常令牌的可能性来确定相应令牌的得分。
4. 根据权利要求1所述的方法,其中,确定异常得分的步骤包括:从所述得分之中选择一组最高得分。
5. 根据权利要求4所述的方法,其中,确定异常得分的步骤还包括:将选择的最高得分的平均值确定为异常得分。
6. 根据权利要求1所述的方法,其中,确定所述可能性的步骤包括:响应于异常得分大于阈值,预测计算系统中未来发生错误。
7. 根据权利要求1所述的方法,还包括:对日志数据执行预处理,其中,预处理包括以下处理中的至少一个:
用表示所有特殊字符或数字的令牌替换日志数据中的特殊字符或数字;
对在替换之后被确定为重复的连续日志消息进行合并;以及
在合并之后,形成按时间顺序连续的日志消息的日志消息组。
8. 根据权利要求1所述的方法,其中,生成所述得分的步骤包括:
通过对包括在日志数据中的按时间顺序连续的日志消息进行分组来形成日志消息组;
以及
通过将每个日志消息组顺序地输入到鉴别器模型,从鉴别器模型获得分别与每个日志消息组对应的异常得分。
9. 根据权利要求1所述的方法,其中,日志数据包括日志消息,并且每个日志消息包括关于日志消息被生成的时间或记录到日志消息的事件发生的时间的相应时间信息。
10. 根据权利要求9所述的方法,其中,每个日志消息包括描述计算系统的状态的状态消息、与计算系统的状态对应的预定义的代码标识符、或与计算系统的状态对应的预定义的数字标识符。
11. 根据权利要求1至10中的任一项所述的方法,其中,鉴别器模型与生成器模型配对以形成鉴别器-生成器模型,并且由鉴别器模型根据来自计算系统的训练日志数据生成用于训练鉴别器模型的训练令牌。
12. 一种用于预测计算系统中的错误的错误预测装置,包括:
一个或多个处理器;以及
存储器,存储指令,所述指令被配置为使所述一个或多个处理器:
接收在计算系统的操作期间由计算系统生成的日志数据,日志数据包括令牌;

将令牌输入到生成各个令牌的得分的鉴别器模型,每个得分对应于相应令牌是异常令牌的概率;

基于所述得分确定异常得分;并且

基于异常得分确定计算系统中未来发生错误的可能性。

13.根据权利要求12所述的错误预测装置,其中,所述指令还被配置为使所述一个或多个处理器:通过调整令牌的得分使得令牌的得分的标准差增大来确定强调得分。

14.根据权利要求13所述的错误预测装置,其中,所述指令还被配置为使所述一个或多个处理器:基于从强调得分之中选择的一组最高得分来确定异常得分。

15.根据权利要求12所述的错误预测装置,其中,所述指令还被配置为使所述一个或多个处理器:响应于异常得分大于阈值,预测计算系统中未来发生错误。

16.根据权利要求12所述的错误预测装置,其中,所述指令还被配置为使所述一个或多个处理器:对日志数据进行预处理,其中,预处理包括:

用预定义的令牌替换日志数据中的日志消息的文本中的特殊字符和数字;

移除日志消息的词令牌之中的与停用词对应的词令牌;

当包括在日志数据中的按时间排序的日志消息之中的相同文本的日志消息连续重复时,对重复的日志消息进行合并;或者

通过对按时间排序的日志消息进行分组来生成日志消息组。

17.根据权利要求12至16中的任一项所述的错误预测装置,其中,所述指令还被配置为使所述一个或多个处理器:

通过对包括在日志数据中的按时间排序的日志消息进行分组来生成日志消息组;并且

通过将每个日志消息组顺序地输入到鉴别器模型,从鉴别器模型获得分别与日志消息组对应的得分。

18.一种训练方法,包括:

收集计算系统的日志数据;

通过对收集的日志数据之中的指示计算系统的正常状态的日志数据执行预处理来生成包括令牌的学习数据;

通过执行掩蔽处理来进一步生成包括掩蔽令牌的学习数据,掩蔽处理阻止令牌中的一些令牌被识别;

使用生成器模型来估计掩蔽令牌的替换令牌;

将包括替换令牌的学习数据输入到鉴别器模型,鉴别器模型对包括替换令牌的学习数据执行推断以:生成指示包括替换令牌的学习数据的令牌是否是替换令牌的第一预测值,并且生成指示在掩蔽处理期间应用了哪种掩蔽模式的第二预测值;以及

基于第一预测值和第二预测值更新生成器模型和鉴别器模型的参数。

19.根据权利要求18所述的训练方法,其中,替换令牌从词典随机选择。

20.根据权利要求18所述的训练方法,其中,生成器模型和鉴别器模型各自包括变换器模型的编码器,并且其中,生成器模型被配置为用从多个候选令牌之中选择的替换令牌来替换掩蔽令牌并输出替换令牌。

预测计算系统中的错误的方法和装置以及训练方法

技术领域

[0001] 以下描述涉及用于预测计算系统中的错误的技术和机器学习技术。

背景技术

[0002] 高性能计算 (HPC) 系统是处理复杂计算的强大计算平台。HPC 系统通常包括高性能硬件和软件。传统上, HPC 系统主要用于工程模拟 (诸如, 核现象、流体动力学、粒子物理以及需要高计算性能的其他模拟)。然而, 人工智能 (AI) 和信息技术 (IT) 的进步已经增加了 HPC 系统在处理和分析海量数据集方面的重要性, 因此已经增加了对 HPC 系统的需求。

[0003] 最近, 已经在开发适用于各种工业领域的 HPC 系统。更先进的 HPC 系统可能由于相当多的组件而具有更复杂的结构和逐渐扩展的尺寸。HPC 系统的复杂结构和增加数量的组件已经倾向于增加 HPC 系统中发生错误和故障的频率。

发明内容

[0004] 提供本发明内容以简要的形式介绍在以下具体实施方式中进一步描述的构思的选择。本发明内容不旨在确定要求保护的主题的关键特征或必要特征, 也不旨在用于帮助确定要求保护的主题的范围。

[0005] 在一个总体方面, 一种预测计算系统中的错误的方法包括: 接收在计算系统的操作期间由计算系统生成的日志数据; 将日志数据令牌化为令牌; 将令牌输入到生成各个令牌的得分的鉴别器模型, 每个得分对应于相应令牌是异常令牌的概率; 基于所述得分确定异常得分; 以及基于异常得分确定计算系统中未来发生错误的可能性。

[0006] 所述方法还可包括: 对所述得分执行增强处理, 其中, 增强处理调整所述得分, 使得所述得分的标准差增大。

[0007] 鉴别器模型可以是神经网络模型, 神经网络模型被配置为基于相应令牌是异常令牌的可能性来确定相应令牌的得分。

[0008] 确定异常得分的步骤可包括: 从所述得分之中选择一组最高得分。

[0009] 确定异常得分的步骤还可包括: 将选择的最高得分的平均值确定为异常得分。

[0010] 确定所述可能性的步骤包括: 响应于异常得分大于阈值, 预测计算系统中未来发生错误。

[0011] 所述方法还可包括: 对日志数据执行预处理, 其中, 预处理包括以下处理中的至少一个: 用表示所有特殊字符或数字的令牌替换日志数据中的特殊字符或数字; 对在替换之后被确定为重复的连续日志消息进行合并; 或者在合并之后, 形成按时间顺序连续的日志消息的日志消息组, 并且通过将日志消息组应用于鉴别器模型来针对相应日志消息组确定计算系统的未来错误的可能性。

[0012] 生成所述得分的步骤可包括: 通过对包括在日志数据中的按时间顺序连续的日志消息进行分组来形成日志消息组; 以及通过将每个日志消息组顺序地输入到鉴别器模型, 从鉴别器模型获得分别与日志消息组对应的异常得分。

[0013] 日志数据可包括日志消息,并且每个日志消息可包括关于日志消息被生成的时间或记录到日志消息的事件发生的时间的相应时间信息。

[0014] 每个日志消息可包括描述计算系统的状态的状态消息、与计算系统的状态对应的预定义的代码标识符、或与计算系统的状态对应的预定义的数字标识符。

[0015] 鉴别器模型可与生成器模型配对以形成鉴别器-生成器模型,并且由鉴别器模型根据来自计算系统的训练日志数据生成用于训练鉴别器的训练令牌。

[0016] 在另一总体方面,一种用于预测计算系统中的错误的计算装置包括:一个或多个处理器;以及存储器,存储指令,所述指令被配置为使所述一个或多个处理器:接收在计算系统的操作期间由计算系统生成的日志数据,日志数据包括令牌;将令牌输入到生成各个令牌的得分的鉴别器模型,每个得分对应于相应令牌是异常令牌的概率;基于所述得分确定异常得分;并且基于异常得分确定计算系统中未来发生错误的可能性。

[0017] 所述指令还可被配置为使所述一个或多个处理器通过调整令牌的得分使得令牌的得分的标准差增大来确定强调得分。

[0018] 所述指令还可被配置为使所述一个或多个处理器基于从增强得分之中选择的一组最高得分来确定异常得分。

[0019] 所述指令还可被配置为使所述一个或多个处理器响应于异常得分大于阈值,预测计算系统中未来发生错误。

[0020] 所述指令还可被配置为使所述一个或多个处理器对日志数据进行预处理,其中,预处理可包括:用预定义的令牌替换日志数据中的日志消息的文本中的特殊字符和数字;移除日志消息的词令牌之中的与停用词对应的词令牌;当包括在日志数据中的按时间排序的日志消息之中的相同文本的日志消息连续重复时,对重复的日志消息进行合并;或者通过对按时间排序的日志消息进行分组来生成日志消息组。

[0021] 所述指令还可被配置为使所述一个或多个处理器:通过对包括在日志数据中的按时间排序的日志消息进行分组来生成日志消息组;并且通过将每个日志消息组顺序地输入到鉴别器,从鉴别器获得分别与日志消息组对应的得分。

[0022] 在另一总体方面,一种方法包括:收集计算系统的日志数据;通过对收集的日志数据之中的指示计算系统的正常状态的日志数据执行预处理来生成包括令牌的训练数据;通过执行掩蔽处理来进一步生成学习数据,掩蔽处理阻止令牌中的一些令牌被识别;使用生成器模型来估计掩蔽令牌的替换令牌;将包括替换令牌的令牌输入到鉴别器,鉴别器对包括替换令牌的令牌执行推断以:生成包括替换令牌的学习数据的令牌是否是替换令牌的第一预测值,并且生成在掩蔽处理期间应用了哪种掩蔽模式的第二预测值;以及基于第一预测值和第二预测值更新生成器模型和鉴别器模型的参数。

[0023] 替换令牌可从字典随机选择。

[0024] 生成器模型和鉴别器模型可各自包括变换器模型的编码器,并且生成器可被配置为用从多个候选令牌之中选择的替换令牌来替换掩蔽令牌并输出替换令牌。

[0025] 根据以下具体实施方式、附图和权利要求,其他特征和方面将是清楚的。

附图说明

[0026] 图1示出根据一个或多个示例实施例的被配置为预测计算系统中的错误的示例错

误预测系统。

[0027] 图2示出根据一个或多个示例实施例的示例训练方法。

[0028] 图3、图4和图5示出根据一个或多个示例实施例的对日志数据执行的数据预处理。

[0029] 图6示出根据一个或多个示例实施例的在学习处理中执行的数据标记的示例。

[0030] 图7示出根据一个或多个示例实施例的使用生成器和鉴别器从学习数据获得第一预测值和第二预测值的示例。

[0031] 图8示出根据一个或多个示例实施例的生成器和鉴别器的示例操作。

[0032] 图9示出根据一个或多个示例实施例的预测计算系统中的错误的示例错误预测方法。

[0033] 图10示出根据一个或多个示例实施例的使用鉴别器从日志数据预测计算系统的错误发生可能性的示例。

[0034] 图11示出根据一个或多个示例实施例的训练装置的示例配置。

[0035] 图12示出根据一个或多个示例实施例的错误预测装置的示例配置。

[0036] 贯穿附图和具体实施方式,除非另外描述或提供,否则相同或相似的附图参考标记将被理解为指代相同或相似的元件、特征和结构。附图可不按比例,并且为了清楚、说明和方便,附图中的元件的相对尺寸、比例和描绘可被夸大。

具体实施方式

[0037] 提供以下详细描述以帮助读者获得对在此描述的方法、设备和/或系统的全面理解。然而,在理解本申请的公开之后,在此描述的方法、设备和/或系统的各种改变、修改和等同物将是清楚的。例如,在此描述的操作的顺序仅是示例,并且不限于在此阐述的那些顺序,而是除了必须以特定次序发生的操作之外,可如在理解本申请的公开之后将是清楚的那样被改变。此外,为了更清楚和简明,可省略对在理解本申请的公开之后已知的特征的描述。

[0038] 在此描述的特征可以以不同的形式体现,并且不应被解释为限于在此描述的示例。相反,提供在此描述的示例仅是为了示出在理解本申请的公开之后将是清楚的实现在此描述的方法、设备和/或系统的许多可行方式中的一些。

[0039] 在此使用的术语仅用于描述各种示例,并且不用于限制公开。除非上下文另外清楚地指示,否则单数形式也旨在包括复数形式。如在此使用的,术语“和/或”包括相关所列项中的任何一个以及任何两个或更多的任何组合。作为非限制性示例,术语“包含”、“包括”和“具有”表明存在陈述的特征、数量、操作、构件、元件和/或它们的组合,但是不排除存在或添加一个或多个其他特征、数量、操作、构件、元件和/或它们的组合。

[0040] 贯穿说明书,当组件或元件被描述为“连接到”、“结合到”或“接合到”另一组件或元件时,它可直接“连接到”、“结合到”或“接合到”所述另一组件或元件,或者可合理地存在介于其间的一个或多个其他组件或元件。当组件或元件被描述为“直接连接到”、“直接结合到”或“直接接合到”另一组件或元件时,可不存在介于其间的其他元件。同样地,例如“在……之间”与“紧接在……之间”以及“与……邻近”与“与……紧邻”的表述也可如前面所描述的那样被解释。

[0041] 尽管在此可使用诸如“第一”、“第二”和“第三”或A、B、(a)、(b)等的术语来描述各

种构件、组件、区域、层或部分,但是这些构件、组件、区域、层或部分不受这些术语限制。这些术语中的每个不用于限定例如对应的构件、组件、区域、层或部分的本质、次序或顺序,而是仅用于将对应的构件、组件、区域、层或部分与其他构件、组件、区域、层或部分区分开。因此,在不脱离示例的教导的情况下,在此描述的示例中所称的第一构件、第一组件、第一区域、第一层或第一部分也可被称为第二构件、第二组件、第二区域、第二层或第二部分。

[0042] 除非另外定义,否则在此使用的所有术语(包括技术术语和科学术语)具有与本公开所属领域的普通技术人员基于对本申请的公开的理解通常理解的含义相同的含义。除非在此明确地如此定义,否则术语(诸如,在通用词典中定义的术语)应被解释为具有与它们在该领域的上下文和本申请的公开中的含义一致的含义,并且不应以理想化或过于形式化的含义进行解释。在此关于示例或实施例的术语“可”的使用(例如,关于示例或实施例可包括或实现什么)表示存在包括或实现这样的特征的至少一个示例或实施例,而所有示例不限于此。

[0043] 在下文中,将参照附图详细描述示例,并且附图中相同的参考标号始终指代相同的元件。

[0044] 图1示出根据一个或多个示例实施例的被配置为预测计算系统中的错误的示例错误预测系统。

[0045] 参照图1,错误预测系统110可预测计算系统100中的错误(或异常)或确定错误的可能性(likelihood),并且可基于计算系统100的日志数据来这样做。关于在此描述的计算系统100,当讨论预测时,术语“错误”和“异常”可互换使用。计算系统100中的错误或异常是指计算系统100与预期或计划不同地操作的情况或者发生故障的情况。

[0046] 计算系统100处理数据和/或执行操作或计算,并且在其处理中生成包括计算系统100的状态和/或情况的日志消息的日志数据。作为非限制性示例,计算系统100可以是高性能计算(HPC)系统、超级计算系统、云系统、数据中心、移动装置、活动装置、个人计算机(PC)等。计算系统100可以是生成日志数据的任何系统或装置。由计算系统100生成的日志数据可包括用于跟踪和分析计算系统100的活动或行为的信息。因此,日志数据可以是用于预测计算系统100中错误的发生的参考。日志数据可包括日志消息和关于生成每个日志消息的时间的时间信息(例如,时间戳)。日志消息可包括例如描述计算系统100的状态的状态消息、与计算系统100的状态对应的预定义的代码标识符和与计算系统100状态对应的预定义的数字标识符中的至少一个。日志数据可由计算系统100中的各种组件(例如,操作系统、应用软件、系统监测服务(例如,性能监测)、一个或多个网络堆栈、装置驱动器等)生成。在一些实施方式中,可由不同的系统组件使用将各种事件记录到相同日志文件的日志记录服务来集中收集日志数据。

[0047] 日志消息可以是例如指示计算系统100的状态或计算系统100的特定组件的状态是正常的正常日志消息。日志消息还可以是指示在计算系统100中或在计算系统100的特定组件中存在异常的异常日志消息。通常,正常日志消息与异常日志消息之间的内容差异在某种程度上是任意的,并且可根据实施方式而变化。然而,日志消息被识别为正常或异常,可使用它们之间的区别来基于日志数据预测计算系统100中的错误。错误预测系统110可将正常日志消息和异常日志消息有效地区分开,并且通过下面详细描述的方法更准确地预测计算系统100中的错误。

[0048] 错误预测系统110可使用基于深度学习的模型(例如,生成器和鉴别器)来预测在未来时间点计算系统100中发生错误的可能性(“错误发生可能性”),基于深度学习的模型基于从包括在计算系统100中的各种组件生成的日志消息进行学习和预测。错误预测系统110可基于包括在日志数据中的日志消息序列中的文本信息来预测计算系统100的错误发生可能性。

[0049] 错误预测系统110可包括日志数据收集模块120、日志数据处理模块130、模型训练模块140和系统异常预测模块150。日志数据收集模块120、日志数据处理模块130、模型训练模块140和系统异常预测模块150的操作可由一个或多个处理器(例如,中央处理器(CPU)和/或图形处理器(GPU))执行。

[0050] 日志数据收集模块120可收集在计算系统100正在操作(例如,执行工作负载)时生成的日志数据。日志数据收集模块120可收集日志数据,该日志数据将被用于测量在计算系统100中是否可能发生错误(响应变量)并构建用于训练基于深度学习的模型的学习数据。日志数据收集模块120可收集在计算系统100的操作期间生成的日志数据。收集的日志数据稍后可用作在计算系统100的错误预测处理中将用于基于深度学习的模型的解释性变量。日志数据可包括例如日志消息、指示日志消息的类型的日志类型(例如,基板管理控制器(BMC)日志、系统日志和组件日志)、标识日志消息的含义的标识符(例如,日志ID)、日志类别值、事件的元数据等。

[0051] 日志数据处理模块130可对计算系统100的日志数据执行数据预处理。日志数据处理模块130可通过对日志消息的文本进行令牌化来将包括在日志数据中的日志消息划分为令牌(token,又称为词元、词向量或词法单元)。例如,令牌可以是词或短语,但是不限于此。总之,令牌可用作处理日志数据的基本单元。

[0052] 从计算系统100收集的日志数据可提供特定于功能或装备的状态信息或通知。日志数据可包括根据计算系统100的硬件特性的特定唯一值,并且唯一值可充当错误预测中的噪声。为了准确的错误预测,可期望泛化或移除包括在日志数据中的这种唯一值和不必要的信息。日志数据处理模块130可用预定义的令牌替换与包括在日志消息中的唯一值对应的令牌。例如,日志数据处理模块130可用泛化词(例如,NUM)替换包括在日志消息中的数字。日志数据处理模块130可确定令牌对于日志消息的分析是不必要的(例如,停用词(stop word))并且移除该令牌。将被视为停用词的令牌可以是预定的,或者可由预定义的规则确定。

[0053] 计算系统100可在相同或接近的时间生成许多日志消息,并且日志数据处理模块130可基于日志消息的含义来对具有相同含义的日志消息进行整合(合并)。对具有重复含义的日志消息进行整合可提高处理速度和效率。另外,日志数据处理模块130可通过在各个时间点将日志消息设置为将被处理的一个单元来将日志消息的顺序属性应用于错误预测。换句话说,日志数据处理模块130可使用各种映射规则将日志消息从其原始形式转换为标准形式。例如,相同类别的日志消息(尽管来自不同的源)可被转换为具有公共格式并且具有类别的标识符的消息/记录。

[0054] 模型训练模块140可基于通过由日志数据处理模块130执行的数据预处理获得的预处理后的日志数据来训练基于深度学习的模型(或基于神经网络的模型)。模型训练模块140可仅使用指示计算系统100正常的日志数据来训练基于深度学习的模型。基于深度学习

的模型可包括生成器和鉴别器。生成器可仅在学习/训练处理中使用,并且可不用于预测计算系统100中的错误(做出推断)。鉴别器可用于预测计算系统100中的错误,并且还可在学习处理中使用。

[0055] 生成器和鉴别器可各自包括变换器(transformer)模型的编码器。变换器模型可在具有编码器-解码器结构(例如,现有的seq2seq结构)的同时利用注意力或自注意力来实现。变换器模型的编码器可包括例如注意力块、相加归一化(Add&Norm)块和前馈块。例如,当编码器被实现为神经网络的层时,编码器可包括与注意力块对应的第一层和与前馈块对应的第二层。注意力块可确定注意力值并且并行地执行自注意力操作。自注意力操作是对自身的操作,并且注意力操作获得注意力值。例如,前馈块可以是通用前馈神经网络。相加归一化块可执行将输入和输出彼此相加的残差连接操作并执行层归一化操作(例如,使用平均值和方差的归一化)。注意力块可确定与给定查询的所有关键字中的每个的相似度,并且将确定的相似度作为权重应用于映射到对应关键字的每个值。注意力块可计算相似度被应用的加权和,并且提供结果作为注意力值。上述查询、关键字和值可以是例如(例如,日志消息的)输入句子中的所有词的向量。在由注意力块执行的自注意力期间,可确定输入句子中的词之间的相似度。

[0056] 模型训练模块140可对从预处理后的日志数据获得的一些令牌执行掩蔽处理。可执行掩蔽处理以确定哪些令牌被选择,并且因此确定哪些(由令牌表示的)信息被选择。生成器可接收包括通过掩蔽处理获得的掩蔽令牌(即,未被掩蔽掉的令牌)的日志数据作为输入,并且可用另一令牌替换掩蔽令牌(或将掩蔽令牌转换为另一令牌)。例如,生成器可通过用词典中的令牌(例如,词)替换掩蔽令牌来生成新句子。

[0057] 鉴别器可接收已被替换令牌的包括替换令牌的日志数据(生成器的输出)作为输入,并且基于此输出指示接收的令牌是替换令牌还是原始令牌的第一预测值。鉴别器还可基于接收的令牌输出与应用于掩蔽处理的掩蔽模式相关联的第二预测值。原始令牌是尚未被生成器用另一令牌进行替换(或转换为另一令牌)的令牌(即,(在预处理之后并且在掩蔽和替换之前)最初包括在日志数据中的令牌)。“由鉴别器预测输入数据中的令牌是替换令牌还是原始令牌”可被称为替换令牌检测(RTD)任务。此外,“由鉴别器预测在掩蔽处理期间(针对接收的日志数据)哪种掩蔽模式被应用”可被称为替换掩蔽检测(RMD)任务。鉴别器可通过RTD任务具有令牌鉴别能力(对令牌单元进行鉴别的能力),并且可通过RMD任务具有词单元鉴别能力。

[0058] 模型训练模块140可确定生成器的输出的第一损失值和鉴别器的输出的第二损失值。可使用损失函数来确定损失。可基于第一损失值和第二损失值来确定总损失值。模型训练模块140可使用例如反向传播算法来更新生成器和鉴别器中的每个的参数(例如,用于实现变换器的编码器的设置值和变量),使得总损失值减小。下面更详细地描述生成器和鉴别器的学习处理。

[0059] 系统异常预测模块150可基于计算系统100的日志数据来预测在计算系统100中发生异常或错误的概率。系统异常预测模块150可使用训练后的鉴别器来预测计算系统100的错误发生可能性。在鉴别器的学习/训练处理完成之后,计算系统100的日志数据可由日志数据收集模块120再次收集,并且可由日志数据处理模块130对收集的日志数据执行数据预处理。系统异常预测模块150可使用鉴别器在特定提前时间(lead time,或前置时间)过去

之后预测计算系统100(是否将发生错误)的错误发生可能性。

[0060] 通过数据预处理获得的预处理后的日志数据(输入令牌)可被输入到鉴别器,并且鉴别器可输出日志数据的相应输入令牌的得分,其中输入令牌的得分对应于输入令牌是异常令牌(即,表示异常或错误)的可能性。例如,鉴别器可输出日志数据的输入令牌不是原始令牌的概率值或日志数据的输入令牌是原始令牌的概率值作为输入令牌的得分。如在此使用的,异常令牌是与计算系统100的异常(或错误)状态相关联的令牌,并且可与对应于处于正常状态的计算系统100的日志数据的令牌区分开。

[0061] 系统异常预测模块150可基于令牌的得分来确定稍后用于预测计算系统100的错误发生可能性的异常得分。系统异常预测模块150可对令牌的得分执行强调处理(highlighting process)以确定强调得分,并且基于强调得分确定异常得分。强调处理可包括将较高值的得分调整为更高并且将较低值的得分调整为更低的增强处理(sharpening process)。通过增强处理,可将在正常状态下获得的日志数据的令牌和在异常状态下获得的日志数据的令牌更有效地区分开。

[0062] 系统异常预测模块150可基于强调得分之中的具有大于或等于预设参考值的值的那些得分或预设数量的高得分(前N个得分)中的那些得分来确定异常得分。例如,系统异常预测模块150可确定(i)大于或等于参考值的得分(例如,增强得分)的平均值或(ii)预设数量的高得分(例如,增强得分)的平均值作为异常得分。系统异常预测模块150可基于异常得分来估计计算系统100的错误发生可能性。例如,当异常得分大于阈值时,系统异常预测模块150可预测在计算系统100中将发生错误。当异常得分小于或等于阈值时,系统异常预测模块150可预测在计算系统100中将不会发生错误。

[0063] 错误预测系统110可解决人类难以解决的问题(即,预测在操作计算系统100的处理中可能发生的错误)。预测可基于例如使用人工智能(AI)技术的超级计算机日志数据。错误预测系统110可使用计算系统100的日志数据来训练基于深度学习的模型,并且使用训练后的基于深度学习的模型来检测计算系统100的异常模式,以提前预测计算系统100中的错误的发生。即使当日志数据另外指示计算系统100的正常操作时(也就是说,当日志数据中没有错误或异常的直接指示时),错误或异常也可被预测。错误预测系统110可通过根据计算系统100的日志数据的特性来优化日志数据(例如,强调处理,仅使用令牌的得分之中的具有高值的一些得分等)来提高错误预测准确度。即使当正常日志消息与异常日志消息之间的文本或表述没有很大差异时以及当没有许多词来区分正常日志消息和异常日志消息时,错误预测也可被改善。

[0064] 图2示出根据一个或多个示例实施例的示例训练方法。训练方法可由在此描述的训练装置(例如,图11的训练装置1100)执行。

[0065] 参照图2,在操作210中,训练装置可收集计算系统(例如,图1的计算系统100)的日志数据。训练装置可收集在计算系统的操作期间(例如,当执行工作负载时)生成的日志数据。日志数据可包括例如日志消息。例如,每个日志消息可包括关于日志消息被生成的时间或者记录到日志消息的事件发生的时间的相应时间信息。每个日志消息可将计算系统的状态表示为文本、生成日志消息的时间、用于标识日志消息的含意的标识符(例如,ID)、日志消息的源、由日志消息的源提供的与日志消息有关的信息或元数据、和/或日志消息的类型,仅举一些示例。

[0066] 在操作220中,训练装置可对日志数据执行数据预处理。训练装置用于(通过预处理)形成学习数据的日志数据中的至少一些可指示计算系统的正常状态。

[0067] 在一个示例实施例中,训练装置可执行例如以下中的至少一个:用预定义的令牌替换日志消息中的特殊字符或数字的文本;移除日志消息的令牌之中的停用词的令牌;当日志数据中具有相同文本的日志消息(例如,相同消息的重复)在按时间生成的日志消息(或按时间的日志消息)之间连续地/邻近地重复时,对重复的日志消息进行整合;或者通过对按时间的日志消息进行分组来生成日志消息组。

[0068] 在日志数据中,可存在具有相同内容的连续日志消息。然而,不需要分析具有相同内容的所有连续日志消息,因此训练装置可将具有这样的重复内容的连续日志消息进行整合(合并)。每个日志消息可包括特殊字符和数字。预测计算系统中的错误的发生基于关于日志消息的信息和日志消息之间的上下文的信息,并且在这种情况下,日志消息中的特殊字符和数字可以是训练基于深度学习的模型的障碍。因此,训练装置可用预定义的令牌替换与日志消息的文本之中的数字对应的文本,并且将该令牌视为停用词以从训练目标移除对于训练或学习而言不必要的令牌。这样的数据预处理处理可减少日志消息中的信息丢失,并且实现基于深度学习的模型的高效训练和学习。通过以这种方式减少日志消息的令牌,具有相同重要性的日志消息在其最终令牌内容方面可变得相同,而不管不显著的文本差异。

[0069] 尽管已经描述了令牌化和令牌替换,但是不需要使用令牌本身。更通用的技术是用泛化信息替换在表示相同事物的日志消息之中可能变化的信息(参数)。令牌替换技术仅是可被用于泛化消息的一种技术。

[0070] 在操作230中,训练装置可对学习数据执行掩蔽处理。训练装置可生成包括通过执行掩蔽处理而获得的掩蔽令牌(已经经过掩蔽的令牌)的学习数据,该掩蔽处理阻止(屏蔽掉)包括在学习数据中的一些令牌被识别。例如,训练装置可将多个预定义的掩蔽模式之中随机选择的掩蔽模式应用于学习数据。掩蔽模式可以是指示包括在学习数据中的令牌之中的哪个令牌将通过掩蔽处理被掩蔽和将被掩蔽的令牌的数量模式。

[0071] 在操作240中,训练装置可使用生成器来生成(估计)用于替换学习数据中的掩蔽令牌的替换令牌。在这种情况下(当执行令牌替换时),仅在计算系统正常(不具有异常或错误)时获得的日志数据可用作学习数据。也就是说,可仅对与计算系统的正常操作相关联的令牌执行令牌替换。例如,其中日志消息的令牌以时间顺序连接的学习数据可被输入到生成器。输入到生成器的学习数据可包括掩蔽令牌。生成器可用从多个候选令牌之中选择的替换令牌来替换掩蔽令牌,并且输出替换令牌。例如,生成器可通过用替换令牌替换掩蔽令牌来新生成句子。替换令牌可以是学习词典中的词。在一些示例中,可随机选择替换令牌。

[0072] 在操作250中,训练装置可获得学习数据(其包括替换令牌)中的每个相应令牌的第一预测值和第二预测值。通过推断获得的第一预测值对应于相应令牌是替换令牌的可能性。第二预测值对应于在掩蔽处理期间给定掩蔽模式已经被应用的可能性。可通过使用对学习数据(其包括替换令牌)执行推断的鉴别器来获得第一预测值和第二预测值。由生成器生成的(令牌形式的)句子可被输入到鉴别器。鉴别器可预测包括在学习数据中的每个令牌是原始令牌还是替换令牌,并且可预测哪种掩蔽模式被应用于学习数据。

[0073] 在操作260中,训练装置可基于第一预测值和第二预测值更新生成器和鉴别器的

参数(例如,节点的权重)。训练装置可基于目标函数连续且重复地更新或优化基于深度学习的模型(例如,生成器和鉴别器)。目标函数可由损失函数或成本函数确定。训练装置可基于生成器的输出确定第一损失,并且分别基于由鉴别器输出的第一预测值和第二预测值确定第二损失和第三损失。训练装置可将最终损失确定为第一损失、第二损失和第三损失的加权和,并且使用反向传播算法更新生成器和鉴别器的参数,使得最终损失降低。这样的学习处理可允许生成器和鉴别器学习与计算系统的正常状态对应的日志数据的特性。

[0074] 图3、图4和图5示出根据一个或多个示例实施例的对日志数据执行的数据预处理。

[0075] 图3示出从将对其执行错误/异常预测的计算系统获得的示例日志数据310。日志数据310包括按照日志消息被生成的时间的顺序的日志消息(按时间的日志消息)。日志消息可包括特殊字符和/或数字。日志消息中的特殊字符和数字可用预定义的令牌替换,并且可移除日志消息的词令牌之中的与停用词对应的词令牌。通过对日志数据310执行的这种数据预处理获得的结果被示出为日志数据320。参照日志数据320,日志消息中的数字(例如,2、3、4和5)被替换为令牌“NUM”。此外,当日志消息以英语生成时,大写字母可共同地被转换为小写字母或者小写字母可共同地被转换为大写字母。在日志数据320中,大写字母被转换为小写字母。随后,当具有相同文本的日志消息是日志数据中的连续副本时,可删除连续重复的日志消息(减少为单个日志消息)。例如,在日志数据320中存在具有相同文本“starting system NUM activity accounting tool”的连续重复的四个日志消息,并且四个日志消息可被整合,并且三个重复的日志消息可被删除。日志数据330是从日志数据320删除这种重复的日志消息的示例。

[0076] 图4示出从计算系统获得的示例日志消息410。日志消息410可包括可被预处理为令牌的句子形式的内容。如图4中所示,可通过用预定义的令牌“NUM”(例如,表示所有特殊字符或数字的令牌)替换包括在日志消息410中的数字并且移除特殊字符(例如,冒号(:)、逗号(,)和等号(=))或用空格替换特殊字符来获得预处理后的日志消息420。此外,包括在日志消息410的令牌中的大写字母可被转换为小写字母。

[0077] 如图4的示例中所示,日志消息可具有指示它们是否与计算系统的正常状态相关联或者它们是否与计算系统的异常(错误/异常)状态相关联的标签。

[0078] 图5示出(如虚线框)针对包括在日志数据500中的按时间顺序的日志消息形成的日志消息组。这样的消息组可作用于处理的数据单元。例如,按时间次序列出的日志消息可被分组成预定义数量的日志消息组,这些日志消息组可按时间顺序彼此重叠。可根据日志消息之间的步幅对日志消息进行分组,以用于确定后续日志消息组。日志消息的分组可被称为“窗口化”,并且每个日志消息组可被称为“窗口”。在这种情况下,包括在日志消息组中的日志消息的数量可被称为“窗口大小”,并且用于确定后续日志消息组的日志消息之间的步幅可被称为“窗口的步幅”。在一些实施方式中,窗口大小可被定义为时间的长度而不是日志消息的数量。

[0079] 如图5中所示,在一个示例中,窗口大小可以是30个消息,并且窗口步幅可以是15个消息。第一日志消息组 x_k 可被形成或定义。随后,通过步幅510(例如,15),特定数量(例如,30)的日志消息可被分组,并且第二日志消息组 x_{k+1} 可被生成。通过该处理,可针对包括在日志数据500中的日志消息形成顺序地重叠的日志消息组。

[0080] 可(i)在预测计算系统中的错误的发生的处理中以及(ii)在学习/训练处理中(两

者都在此被描述)执行以上参照图3、图4和图5描述的日志数据的处理。

[0081] 图6示出根据一个或多个示例实施例的对将在学习(训练)处理中使用的训练数据进行标记的示例。

[0082] 通常,在计算系统的错误发生可能性的预测中,可存在预示着计算系统中发生错误的一些迹象,并且迹象可例如取决于以下方面而发生在不同时间:计算系统中发生的错误或缺陷的类型、计算系统的类型、包括在计算系统中的组件的类型以及日志消息的类型。因此,使用在固定的预定时间过去之后预测错误的模型可降低通用的(高度改变的)应用的错误预测,并且可专用于仅预测特定类型的错误。另外,为了改变用于预测错误发生的时间跨度(即,预测错误的时序),可需要对模型进行重新训练。对于这样的问题,可基于用于预测错误的日志消息的最末时间点来预测多个时间点(多个时间跨度)的错误发生可能性。为此,训练装置可在学习/训练处理中使用两个或更多个固定间隔来形成正确答案或地面真值(GT)。例如,当与GT的时间点的时间差(时间跨度)是提前时间并且提前时间被设置为10秒、30秒、60秒、180秒和300秒时,GT的间隔可分别包括(10、30)秒、(30、60)秒、(60、180)秒、(180、300)秒和(300、420)秒。例如,当提前时间是10秒时,对应的预测可指示在(i)包括在日志消息组中的最末日志消息被生成的时间点之后的10秒与(ii)30秒之间的时间期间是否发生错误。训练装置可将由基于提前时间从按时间的日志消息中之中选择的一个或多个日志消息指示的状态值设置为与学习数据对应的预测结果值的标签。包括在学习/训练数据中的每个实例(训练样本)可包括固定数量(例如,30个)的日志消息的一个日志消息组,并且与样本/实例相关联,可基于在提前时间之后的设置时间间隔内是否存在计算装置中的错误的指示来将目标值(或标签)设置为正常或异常。

[0083] 参照图6,可基于包括在当前日志消息组610中的日志消息来估计(例如,用于训练的)错误发生可能性。“在从包括在日志消息组610中的最末日志消息起的预设提前时间620(例如,10秒)之后的设置时间间隔630内计算装置中是否发生错误”可用作估计的错误发生可能性的GT。例如,“在从(i)包括在当前日志消息组610中的最末日志消息(例如,日志消息30)被生成的时间点之后10秒到(ii)30秒的时间间隔630期间是否发生错误”可用作GT。通过这样的处理,可对学习数据执行标记。

[0084] 图7示出根据一个或多个示例实施例的使用生成器和鉴别器从学习数据获得第一预测值和第二预测值的示例。

[0085] 参照图7,在学习处理中,可对计算装置的日志消息执行数据预处理,并且可确定学习数据710。可对学习数据710执行掩蔽处理720,并且包括在学习数据710中的句子 S_i 中的令牌 $t_1, t_2, \dots, t_{h-1}$ 和 t_h 之中的一些令牌 t_2 和 t_h 可用掩蔽令牌 t_2^{mask} 和 t_h^{mask} 替换,掩蔽令牌 t_2^{mask} 和 t_h^{mask} 本身不能被识别为掩蔽令牌。因此,可生成包括掩蔽令牌 t_2^{mask} 和 t_h^{mask} 的学习数据730。例如,可从令牌 $t_1, t_2, \dots, t_{h-1}$ 和 t_h 之中随机确定将被掩蔽的令牌。从多个预定义的掩蔽模式之中随机选择的掩蔽模式可用于掩蔽处理720。

[0086] 学习数据730可被输入到生成器740,并且生成器740可估计包括在学习数据730中的掩蔽令牌 t_2^{mask} 和 t_h^{mask} 的替换令牌 r_2 和 r_h 。因此,可通过生成器740生成用替换令牌 r_2 和 r_h 替换掩蔽令牌 t_2^{mask} 和 t_h^{mask} 的学习数据750,生成器740用预定义的词典中的词(例如,随机词)替

换掩蔽令牌 t_2^{mask} 和 t_h^{mask} 。学习数据750可以以句子的形式提供。生成器740可包括变换器模型(已知类型的神经网络)的编码器。

[0087] 学习数据750可被输入到鉴别器760,并且鉴别器760可对学习数据750执行推断并输出与学习数据750(包括替换令牌 r_2 和 r_h)的相应令牌 t_1 、 r_2 、 $\dots$ 、 t_{h-1} 和 r_h 中的每个是否对应于替换令牌相关联的推断的第一预测值。鉴别器还可输出与应用于掩蔽处理720的掩蔽模式相关联的相应令牌 t_1 、 r_2 、 $\dots$ 、 t_{h-1} 和 r_h 的第二预测值。对于给定令牌,第一预测值和第二预测值可以是向量值、概率值或得分的形式。鉴别器760可包括变换器模型的编码器,其中第一头输出第一预测值,并且第二头输出第二预测值。第一头和第二头是变换器模型的结构中的头。

[0088] 训练装置可基于第一损失、第二损失和第三损失来定义总损失。第一损失基于生成器740的输出,第二损失基于由鉴别器760输出的第一预测值,并且第三损失基于由鉴别器760输出的第二预测值。训练装置可使用反向传播算法来更新生成器740和鉴别器760的参数(例如,层中的节点的权重),使得最终损失降低。第一损失可反映生成器740确定替换令牌(例如, r_2 和 r_h)与原始令牌(例如, t_2 和 t_h)的接近程度。第二损失可反映鉴别器760针对学习数据750在替换令牌与原始令牌之间进行鉴别的准确程度。第三损失可反映鉴别器760鉴别在掩蔽处理720中使用的掩蔽模式的准确程度。

[0089] 图8示出根据一个或多个示例实施例的生成器和鉴别器的示例操作。

[0090] 参照图8,给出学习数据830以训练生成器810和鉴别器820。学习数据830可被划分为例如以词为单位的令牌。训练装置可根据确定的掩蔽模式执行掩蔽处理,并且生成包括掩蔽令牌的学习数据840。如图8中所示,可对“the”和“root”的令牌执行掩蔽处理。可将包括掩蔽令牌的学习数据840输入到生成器810,并且生成器810可估计掩蔽令牌的替换令牌。例如,生成器810可将“the”估计为通过对“the”执行的掩蔽处理而获得的掩蔽令牌的替换令牌,并且可将“idle”估计为通过对“root”执行的掩蔽处理而获得的掩蔽令牌的替换令牌。对于掩蔽令牌“root”,生成器810可不正确地估计替换令牌。

[0091] 在这种情况下,学习数据850可被输入到鉴别器820,并且鉴别器820可输出“指示包括在学习数据850中的哪些令牌是由生成器810估计的替换令牌以及哪些令牌是原始令牌”的估计860,在学习数据850中,用由生成器810估计的替换令牌来替换包括在学习数据840中的掩蔽令牌。此外,鉴别器820可估计在生成学习数据840时应用的掩蔽模式。例如,鉴别器820可估计在生成学习数据840时从多个预定义的不同掩蔽模式之中选择的掩蔽模式。

[0092] 图9示出根据一个或多个示例实施例的预测计算系统中的错误的示例错误预测方法。错误预测方法可由在此描述的错误预测装置(例如,图12的错误预测装置1200)执行。

[0093] 参照图9,在操作910中,错误预测装置可接收在计算系统的操作期间生成的日志数据。错误预测装置可接收在计算系统中根据时间顺序生成的日志消息,并且存储接收的日志消息。

[0094] 错误预测装置可对日志数据执行数据预处理。数据预处理可包括例如以下处理:用预定义的令牌替换与包括在日志数据中的日志消息的文本之中的特殊字符和数字中的至少一个对应的文本;移除日志消息的词令牌之中的与停用词对应的词令牌;当包括在日志数据中的按时间的日志消息之中的相同文本的日志消息连续重复时,对重复的日志消息进行合并;和/或对按时间的日志消息进行分组并生成日志消息组的处理。对于日志数据的

数据预处理的细节,可参照上面参照图2关于操作220描述的数据预处理。

[0095] 在操作920中,错误预测装置可使用鉴别器获得与包括在日志数据中的每个令牌是否是异常令牌相关联的得分。在这种情况下,日志数据(其可以是数据预处理而获得的日志数据)可包括令牌化的日志消息。例如,得分可以是概率值。错误预测装置可将按时间的日志消息分组为多个日志消息组,并且将日志消息组中的每个顺序地输入到鉴别器,以从鉴别器获得与相应日志消息组对应的得分。例如,连接到包括在日志消息组中的特定数量(例如,30)的日志消息的数据可被输入到鉴别器。当在预定时间间隔内收集日志数据时,可顺序地生成日志数据组,并且可按照生成日志数据组的顺序将日志数据组顺序地输入到鉴别器。

[0096] 在一个示例实施例中,鉴别器可以是确定得分的模型(例如,神经网络),该得分是(或基于)包括在日志数据中的令牌是异常令牌的可能性。该得分可对应于从鉴别器(例如,图7的鉴别器760)的第一头输出的值,鉴别器的第一头输出来自鉴别器的第一预测值。可在学习处理期间训练鉴别器,以将由生成器替换的令牌与当计算系统处于正常状态时获得的日志数据有效地区分开,并且因此,当作为鉴别的目标的令牌被估计为当计算系统处于错误状态或处于预期发生错误的状态时获得的日志数据的令牌(或异常令牌)时,将高值的得分分配给该令牌。相反,当作为鉴别的目标的令牌被估计为当计算系统处于正常状态时获得的日志数据的令牌时,鉴别器可将低值的得分分配给该令牌。

[0097] 在操作930中,错误预测装置可确定通过对令牌的得分执行的强调处理获得的强调(增强)得分。错误预测装置可通过以增大令牌的得分之间的标准差的方式调整令牌的得分来确定强调得分。由于在(i)对应于正常状态的日志消息的词或表述与(ii)对应于错误状态的日志消息的词或表述之间可不存在很大差异,因此可需要执行强调处理以进一步放大差异。强调处理可有助于将与计算系统的正常状态对应的日志消息序列和与计算系统的错误状态(或异常状态)对应的日志消息序列有效地区分开。可基于例如等式1来执行强调处理。

[0098] 等式1

$$[0099] \quad \text{sharpen}(S, T) := \frac{t_i^{\frac{T}{T-1}}}{\sum_{j=1}^h t_j^{\frac{T}{T-1}}}$$

[0100] 在等式1中,S表示强调处理(例如,增强处理)所针对的日志消息组,并且 t_i 表示包括在日志消息组中的第i令牌的异常得分(可针对h个令牌中的每个计算 $\text{sharpen}(S, T)$ 函数)。j是用于对h个令牌中的每个进行求和的索引,并且T表示温度缩放参数。在学习期间,可将温度缩放参数T设置为1,然后可在实际错误预测步骤中确定异常得分以估计给定输入的异常概率。在这种情况下,每个异常得分 t 可被取 $1/T$ 次方(幂)。温度缩放参数可增加分布内样本的异常得分与分布外样本的异常得分之间的距离,从而有助于改善分布外样本的区分度。温度缩放参数值可被预设为适合于特定应用或实施方式的值。 $\text{sharpen}(S, T)$ 表示强调处理的结果。强调处理可将较高的得分值分配给已经具有高得分的令牌,并且将较低的得分值分配给已经具有低得分的令牌。

[0101] 在操作940中,错误预测装置可基于通过强调处理获得的强调得分来确定异常得分。例如,错误预测装置可基于以从强调得分之中的最高得分开始的次序选择的高得分中的一些(例如,得分的总数的1/4)来确定异常得分。错误预测装置可通过仅选择与计算系统

的错误发生可能性紧密相关的一些令牌来确定异常得分。例如,错误预测装置可使用选择的得分的平均值作为异常得分。

[0102] 在操作950中,错误预测装置可基于异常得分来估计计算系统的错误发生可能性。例如,响应于异常得分大于阈值,错误预测装置可预测计算系统中发生错误。响应于异常得分小于或等于阈值,错误预测装置可预测计算系统中未发生错误。进行错误发生可能性的预测的未来时间间隔(即,预测发生错误的时间间隔)可对应于“从包括在用于错误预测的日志数据中的最末日志消息的生成时间起”的特定提前时间之后的定义的时间间隔。

[0103] 图10示出根据一个或多个示例实施例的使用鉴别器从日志数据预测计算系统的错误发生可能性的示例。

[0104] 参照图10,可在预测计算系统中的错误的处理中接收由计算装置生成的日志数据1010。日志数据1010可以通过上面参照图9描述的数据预处理获得并被划分为令牌单元的数据。日志数据1010可被划分为多个令牌 t_1 、 t_2 、……、 t_{h-1} 和 t_h 。

[0105] 日志数据1010可被输入到鉴别器1020,并且鉴别器1020可输出与包括在日志数据1010中的令牌是否是异常令牌相关联的得分1030。得分1030可以是例如指示令牌 t_1 、 t_2 、……、 t_{h-1} 和 t_h 分别对应于异常令牌的概率值。例如,当令牌被估计为非常可能对应于异常令牌时,可分配高值作为令牌的得分,并且当令牌被估计为不太可能对应于异常令牌时,可分配低值作为令牌的得分。

[0106] 错误预测装置可确定通过对令牌的得分1030执行的强调处理1040获得的强调得分1050。错误预测装置可通过调整令牌的得分1030来确定强调得分1050,使得得分1030的标准差增大。通过强调处理,可将较高的得分值分配给具有高得分的令牌(例如, t_3 、 t_4 和 t_{h-1}),并且可将较低的得分值分配给具有低得分的令牌(例如, t_1 、 t_2 、 t_5 和 t_h)。

[0107] 错误预测装置可通过将强调得分1050从最高值到最低值排序来确定排列的得分1060。在步骤1070中,错误预测装置可基于得分1060确定异常得分。更具体地,错误预测装置可基于从强调得分1050的高值之中选择的高得分来确定异常得分。例如,错误预测装置可将具有高得分的令牌(例如, t_{h-1} 、 t_4 和 t_3)的得分的平均值确定为异常值。在步骤1080中,错误预测装置可基于确定的异常得分来预测计算系统的错误发生可能性。例如,响应于异常得分大于阈值,错误预测装置可预测在特定未来时间间隔(例如,从包括在用于当前错误预测的日志数据中的最末日志消息的生成时间延续的特定提前时间之后的定义的时间间隔)内在计算系统中发生错误。

[0108] 图11示出根据一个或多个示例实施例的训练装置的示例配置。

[0109] 参照图11,训练装置1100可训练生成器和鉴别器,生成器和鉴别器是在此描述的基于深度学习的模型。训练装置1100可包括处理器1110、存储器1120和通信模块1130,并且训练装置1100的每个组件可通过通信总线1140彼此通信。在一个示例实施例中,可省略这些组件中的一些(例如,通信模块1130),或者可将其他组件添加到训练装置1100。

[0110] 处理器1110可控制训练装置1100的其他组件(例如,硬件或软件组件),并且可执行各种类型的数据处理或操作。在一个示例实施例中,作为数据处理或操作的至少一部分,处理器1110可将另一组件接收的指令或数据存储在存储器1120中,处理存储在存储器1120中的指令或数据,并且将得到的数据存储在存储器1120中。

[0111] 在一个示例实施例中,处理器1110可以是主处理器(例如,CPU或应用处理器(AP))

和/或可独立于主处理器进行操作或与主处理器结合操作的辅助处理器(例如,GPU、神经处理器(NPU)、图像信号处理器(ISP)、传感器集线器处理器或通信处理器(CP))。

[0112] 存储器1120可存储由训练装置1100的至少一个组件(例如,处理器1110或通信模块1130)使用的各种数据。数据可包括例如程序(例如,应用)以及用于与其相关的命令的输入数据或输出数据。存储器1120可存储可由处理器1110执行的指令(或程序)。存储器1120可包括例如易失性存储器或非易失性存储器。

[0113] 通信模块1130可支持训练装置1100与另一装置之间的直接(或有线)通信信道或者无线通信信道的建立,并且支持通过建立的通信信道的通信。例如,通信模块1130可从计算系统接收日志数据。通信模块1130可包括用于执行通信功能的通信电路。通信模块1130可包括可独立于处理器1110进行操作并且支持直接(例如,有线)通信或无线通信的CP。通信模块1130可包括执行无线通信的无线通信模块(例如,蓝牙TM通信模块、蜂窝通信模块、Wi-Fi通信模块或全球导航卫星系统(GNSS)通信模块)或者有线通信模块(例如,局域网(LAN)通信模块或电力线通信(PLC)模块)。

[0114] 处理器1110可收集计算系统的日志数据。处理器1110可通过对收集的日志数据之中的指示计算系统为正常的日志数据执行数据预处理来确定学习数据。由处理器1110执行的数据预处理可包括例如以下处理中的至少一个:用预定义的令牌替换与包括在日志数据中的日志消息的文本之中的特殊字符或数字中的至少一个对应的文本的处理;移除日志消息的词令牌之中的与停用词对应的词令牌的处理;当包括在日志数据中的按时间的日志消息之中的相同文本的日志消息连续重复时,对重复的日志消息进行整合的处理;对按时间的日志消息进行分组并生成日志消息组的处理;以及将由基于提前时间从按时间的日志消息之中选择的一个或多个日志消息指示的状态值设置为与学习数据对应的预测结果值的标签的处理。

[0115] 处理器1110可通过执行掩蔽处理来生成包括掩蔽令牌的学习数据,该掩蔽处理阻止对包括在学习数据中的多个令牌之中的一些令牌的识别。处理器1110可使用生成器来估计掩蔽令牌的替换令牌。生成器可用从多个候选令牌之中选择的替换令牌来替换掩蔽令牌,并且输出替换令牌。处理器1110可使用将包括替换令牌的学习数据用作输入的鉴别器,来获得与学习数据(其包括替换令牌)的每个令牌是否对应于替换令牌相关联的第一预测值以及与应用于掩蔽处理的掩蔽模式相关联的第二预测值。处理器1110可基于第一预测值和第二预测值来更新生成器和鉴别器的参数(例如,节点之间的连接的权重)。处理器1110可基于生成器的输出值以及鉴别器的第一预测值和第二预测值来确定损失,并且可使用反向传播算法来更新生成器和鉴别器的参数,使得损失降低。

[0116] 图12示出根据一个或多个示例实施例的错误预测装置的示例配置。

[0117] 参照图12,错误预测装置1200可使用在此描述的鉴别器来预测计算系统的错误发生可能性,该错误发生可能性是计算系统中发生错误的可能性。错误预测装置1200可包括处理器1210、存储器1220和通信模块1230,并且错误预测装置1200的每个组件可通过通信总线1240彼此通信。在一个示例实施例中,可省略组件中的一些(例如,通信模块1230),或者可将其他组件添加到错误预测装置1200。

[0118] 处理器1210可控制错误预测装置1200的其他组件(例如,硬件或软件组件)并执行各种类型的数据处理或操作。在一个示例实施例中,作为数据处理或操作的至少一部分,处

理器1210可将从另一组件接收的指令或数据存储在存储器1220中,处理存储在存储器1220中的指令或数据,并且将得到的数据存储在存储器1220中。

[0119] 处理器1210可以是主处理器(例如,CPU或AP)和/或可独立于主处理器进行操作或与主处理器结合操作的辅助处理器(例如,GPU、NPU、ISP、传感器集线器处理器或CP)。

[0120] 存储器1220可存储由错误预测装置1200的至少一个组件(例如,处理器1210或通信模块1230)使用的各种数据。数据可包括例如程序(例如,应用)以及用于与其相关的命令的输入数据或输出数据。存储器1220可存储可由处理器1210执行的指令(或程序)。存储器1220可包括易失性存储器或非易失性存储器。

[0121] 通信模块1230可支持错误预测装置1200与另一装置之间的直接(或有线)通信信道或者无线通信信道的建立,并且支持通过建立的通信信道的通信。例如,通信模块1230可从计算系统接收日志数据。通信模块1230可包括用于执行通信功能的通信电路。通信模块1230可包括可独立于处理器1210进行操作并且支持直接(例如,有线)通信或无线通信的CP。通信模块1230可包括执行无线通信的无线通信模块(例如,蓝牙™通信模块、蜂窝通信模块、Wi-Fi通信模块或GNSS通信模块)或者有线通信模块(例如,LAN通信模块或PLC模块)。

[0122] 处理器1210可接收在计算系统的操作期间生成的日志数据。在一个示例实施例中,处理器1210可对日志数据执行数据预处理。由处理器1210执行的数据预处理可包括例如以下处理中的至少一个:用预定义的令牌替换与包括在日志数据中的日志消息的文本之中的特殊字符或数字中的至少一个对应的文本的处理;移除日志消息的词令牌之中的与停用词对应的词令牌的处理;当包括在日志数据中的按时间的日志消息之中的相同文本的日志消息连续重复时,对重复的日志消息进行整合的处理;对按时间的日志消息进行分组并生成日志消息组的处理;以及将由基于提前时间从按时间的日志消息之中选择的一个或多个日志消息指示的状态值设置为与学习数据对应的预测结果值的标签的处理。

[0123] 可使用鉴别器获得与包括在日志数据中的每个令牌是否是异常令牌相关联的得分。处理器1210可对包括在日志数据中的按时间的日志消息进行分组并生成多个日志消息组,并且将多个日志消息组中的每个日志消息组顺序地输入到鉴别器,以从鉴别器获得与日志消息组对应的得分。

[0124] 处理器1210可确定通过对令牌的得分执行的强调处理获得的强调得分。处理器1210可通过调整令牌的得分使得令牌的得分的标准差增大来确定强调得分。处理器1210可基于强调得分来确定异常得分。例如,处理器1210可基于以从强调得分之中的最高得分开始的次序选择的高得分中的一些来确定异常得分。处理器1210可基于异常得分来估计计算系统的错误发生可能性。例如,响应于异常得分大于阈值,处理器1210可预测计算系统中发生错误。

[0125] 在此描述的示例实施例可使用硬件组件、软件组件和/或它们的组合来实现。处理装置可使用一个或多个通用或专用计算机(诸如,以处理器、控制器、算术逻辑单元(ALU)、数字信号处理器、微型计算机、现场可编程门阵列(FPGA)、可编程逻辑单元(PLU)、微处理器或能够以限定的方式响应并执行指令的任何其他装置为例)来实现。处理装置可运行操作系统(OS)和在OS上运行的一个或多个软件应用。处理装置还可响应于软件的执行来访问、存储、操纵、处理和创建数据。为了简单起见,处理装置的描述被用作单数形式;然而,本领域技术人员将理解,处理装置可包括多个处理元件和多种类型的处理元件。例如,处理装置

可包括多个处理器或者处理器和控制器。另外,不同的处理配置(诸如,并行处理器)是可行的。

[0126] 在此关于图1至图12描述的计算设备、电子装置、处理器、存储器、显示器、信息输出系统和硬件、存储装置以及其他设备、装置、单元、模块和组件由硬件组件实现或表示硬件组件。可用于执行在本申请中描述的操作的硬件组件的示例在适当的情况下包括:控制器、传感器、生成器、驱动器、存储器、比较器、算术逻辑单元、加法器、减法器、乘法器、除法器、积分器以及被配置为执行在本申请中描述的操作的任何其他电子组件。在其他示例中,执行在本申请中描述的操作的一个或多个硬件组件由计算硬件实现(例如,由一个或多个处理器或计算机实现)。处理器或计算机可由一个或多个处理元件(诸如,逻辑门阵列、控制器和算术逻辑单元、数字信号处理器、微型计算机、可编程逻辑控制器、现场可编程门阵列、可编程逻辑阵列、微处理器或被配置为以限定的方式响应并执行指令以实现期望结果的任何其他装置或装置的组合)实现。在一个示例中,处理器或计算机包括或连接到存储由处理器或计算机执行的指令或软件的一个或多个存储器。由处理器或计算机实现的硬件组件可执行指令或软件(诸如,操作系统(OS)和在OS上运行的一个或多个软件应用),以执行在本申请中描述的操作。硬件组件还可响应于指令或软件的执行来访问、操纵、处理、创建和存储数据。为了简单起见,可在本申请中描述的示例的描述中使用单数术语“处理器”或“计算机”,但是在其他示例中,可使用多个处理器或计算机,或者处理器或计算机可包括多个处理元件、或多种类型的处理元件、或两者。例如,单个硬件组件或者两个或更多个硬件组件可由单个处理器、或者两个或更多个处理器、或者处理器和控制器来实现。一个或多个硬件组件可由一个或多个处理器或者处理器和控制器实现,并且一个或多个其他硬件组件可由一个或多个其他处理器或者另一处理器和另一控制器实现。一个或多个处理器或者处理器和控制器可实现单个硬件组件、或者两个或更多个硬件组件。硬件组件可具有不同处理配置中的任何一个或多个,不同处理配置的示例包括单个处理器、独立处理器、并行处理器、单指令单数据(SISD)多处理、单指令多数据(SIMD)多处理、多指令单数据(MISD)多处理和多指令多数据(MIMD)多处理。

[0127] 在图1至图12中示出的执行在本申请中描述的操作的方法由计算硬件执行(例如,由一个或多个处理器或计算机执行),其中,计算硬件被实现为如上所述地执行指令或软件,以执行在本申请中描述的由所述方法所执行的操作。例如,单个操作或者两个或更多个操作可由单个处理器、或者两个或更多个处理器、或者处理器和控制器执行。一个或多个操作可由一个或多个处理器或者处理器和控制器执行,并且一个或多个其他操作可由一个或多个其他处理器或者另一处理器和另一控制器执行。一个或多个处理器或者处理器和控制器可执行单个操作、或者两个或更多个操作。

[0128] 用于控制计算硬件(例如,一个或多个处理器或计算机)以实现硬件组件并执行如上所述的方法的指令或软件可被编写为计算机程序、代码段、指令或它们的任何组合,以单独地或共同地指示或配置一个或多个处理器或计算机作为用于执行由硬件组件执行的操作和如上所述的方法的机器或专用计算机进行操作。在一个示例中,指令或软件包括由一个或多个处理器或计算机直接执行的机器代码(诸如,由编译器产生的机器代码)。在另一示例中,指令或软件包括由一个或多个处理器或计算机使用解释器执行的高级代码。指令或软件可基于附图中示出的框图和流程图以及在此的相应描述使用任何编程语言来编写,

附图中示出的框图和流程图以及在此的相应描述公开了用于执行由硬件组件执行的操作和如上所述的方法的算法。

[0129] 用于控制计算硬件(例如,一个或多个处理器或计算机)实现硬件组件并执行如上所述的方法的指令或软件、以及任何相关联的数据、数据文件和数据结构可被记录、存储或固定在一个或多个非暂时性计算机可读存储介质中或被记录、存储或固定在一个或多个非暂时性计算机可读存储介质上。非暂时性计算机可读存储介质的示例包括:只读存储器(ROM)、随机存取可编程只读存储器(PROM)、电可擦除可编程只读存储器(EEPROM)、随机存取存储器(RAM)、动态随机存取存储器(DRAM)、静态随机存取存储器(SRAM)、闪存、非易失性存储器、CD-ROM、CD-R、CD+R、CD-RW、CD+RW、DVD-ROM、DVD-R、DVD+R、DVD-RW、DVD+RW、DVD-RAM、BD-ROM、BD-R、BD-R LTH、BD-RE、蓝光或光盘存储设备、硬盘驱动器(HDD)、固态驱动器(SSD)、闪存、卡式存储器(诸如,多媒体卡或微型卡(例如,安全数字(SD)或极端数字(XD)))、磁带、软盘、磁光数据存储装置、光学数据存储装置、硬盘、固态盘和任何其他装置,任何其他装置被配置为以非暂时方式存储指令或软件、以及任何相关联的数据、数据文件和数据结构,并向一个或多个处理器或计算机提供指令或软件、以及任何相关联的数据、数据文件和数据结构,使得一个或多个处理器或计算机可执行指令。在一个示例中,指令或软件以及任何相关联的数据、数据文件和数据结构分布在联网的计算机系统上,使得指令和软件以及任何相关联的数据、数据文件和数据结构通过一个或多个处理器或计算机以分布式方式存储、访问和执行。

[0130] 虽然本公开包括特定的示例,但是在理解本申请的公开之后将清楚的是,在不脱离权利要求及其等同物的精神和范围的情况下,可在这些示例中进行形式和细节上的各种改变。在此描述的示例仅被认为是描述性的,而不是出于限制的目的。每个示例中的特征或方面的描述将被认为适用于其他示例中的类似特征或方面。如果描述的技术以不同的次序执行,和/或如果描述的系统、架构、装置或电路中的组件以不同的方式组合,和/或由其他组件或它们的等同物替换或补充,则可实现合适的结果。

[0131] 因此,除了以上公开之外,公开的范围也可由权利要求及其等同物限定,并且在权利要求及其等同物的范围内的所有变化都应被解释为包括在公开中。

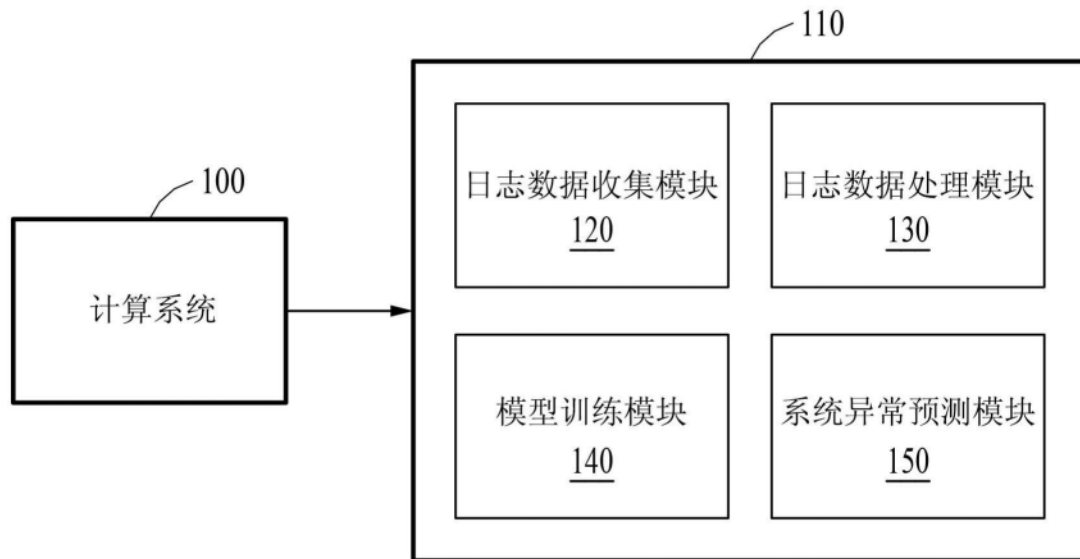


图1

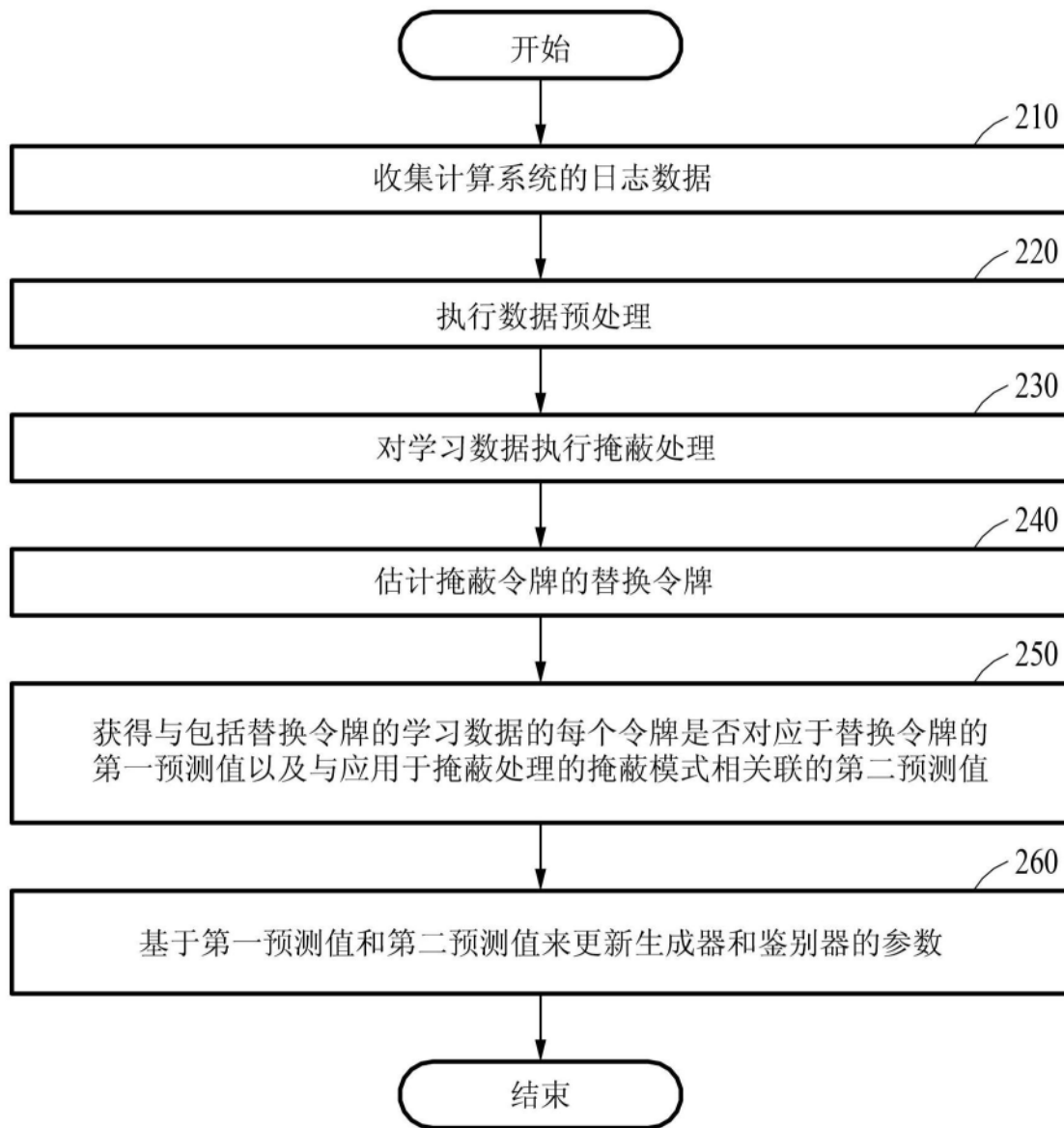


图2

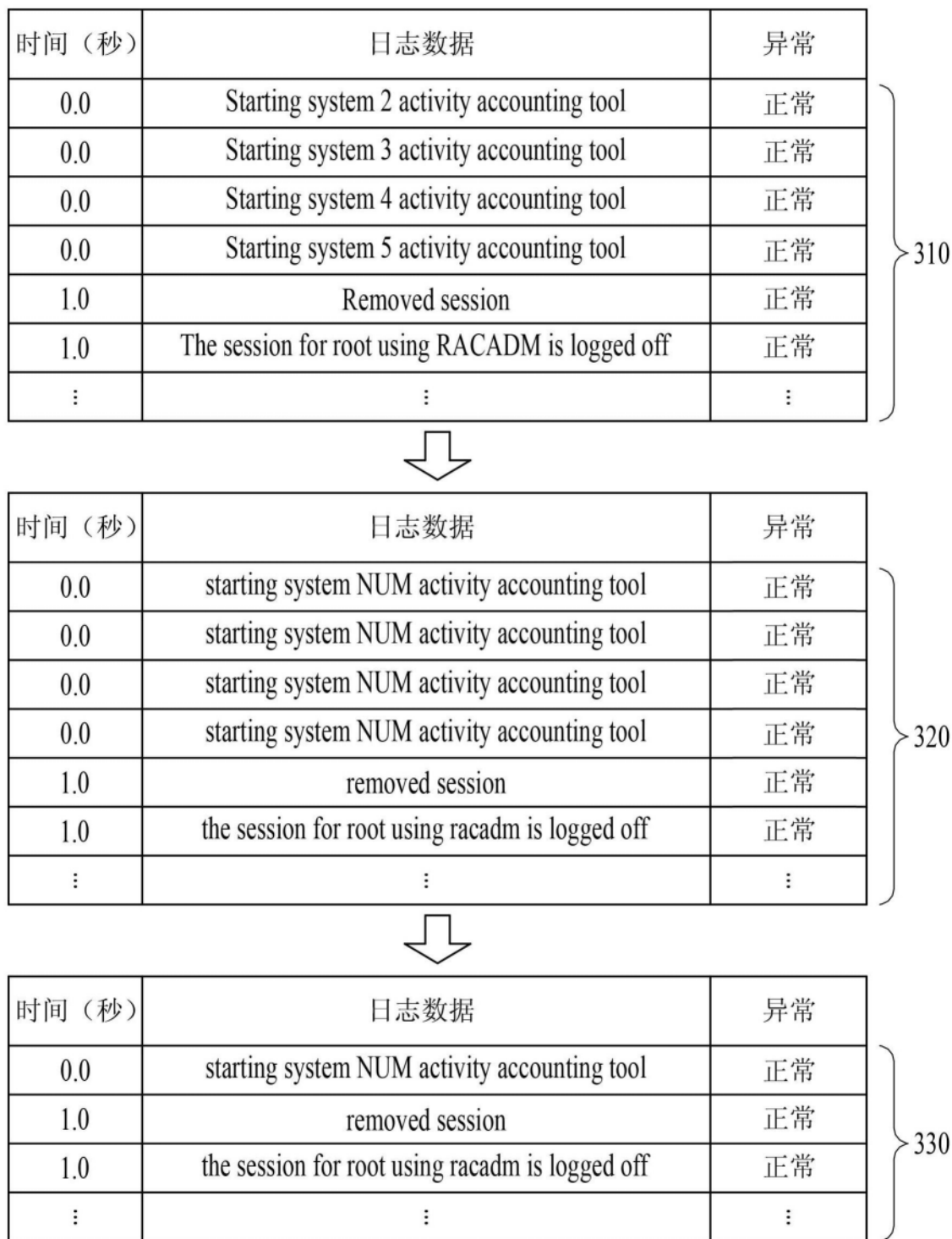


图3

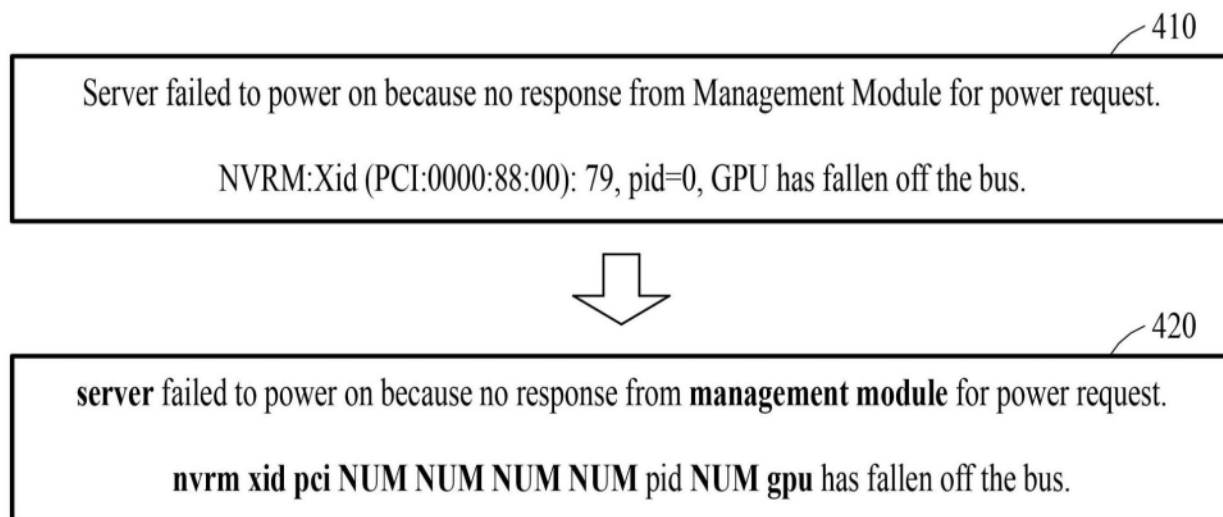


图4

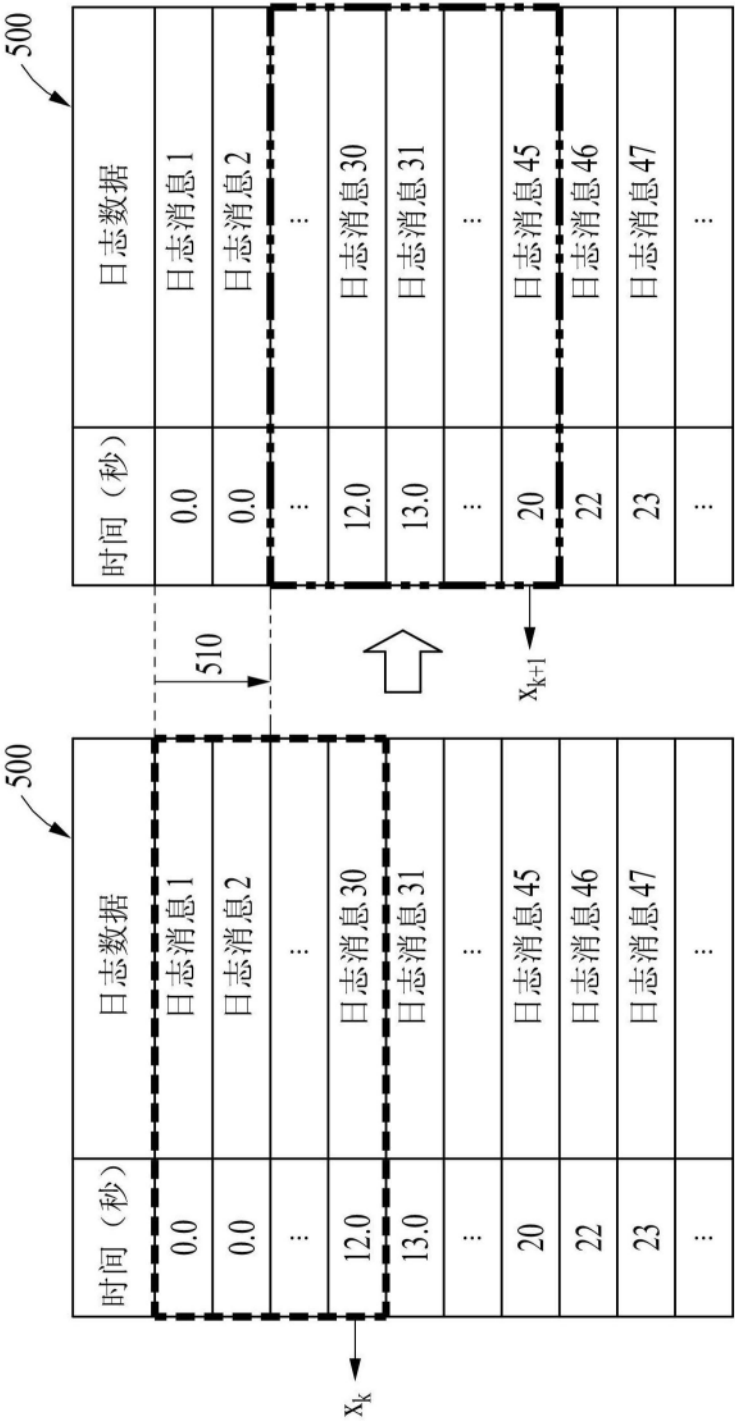


图5

时间（秒）	日志数据	异常	
0.0	日志消息 1	正常	610
1.0	日志消息 2	正常	
⋮	⋮	⋮	
31.0	日志消息 29	错误	
50.0	日志消息 30	正常	
提前时间10秒			
51.0	日志消息 31	正常	620
51.0	日志消息 32	正常	
52.0	日志消息 33	正常	
55.0	日志消息 34	正常	
60.0	日志消息 35	正常	630
61.0	日志消息 36	正常	
64.0	日志消息 37	正常	
70.0	日志消息 38	错误	
83.0	日志消息 39	正常	
88.0	日志消息 40	正常	
90.0	日志消息 41	正常	

图6

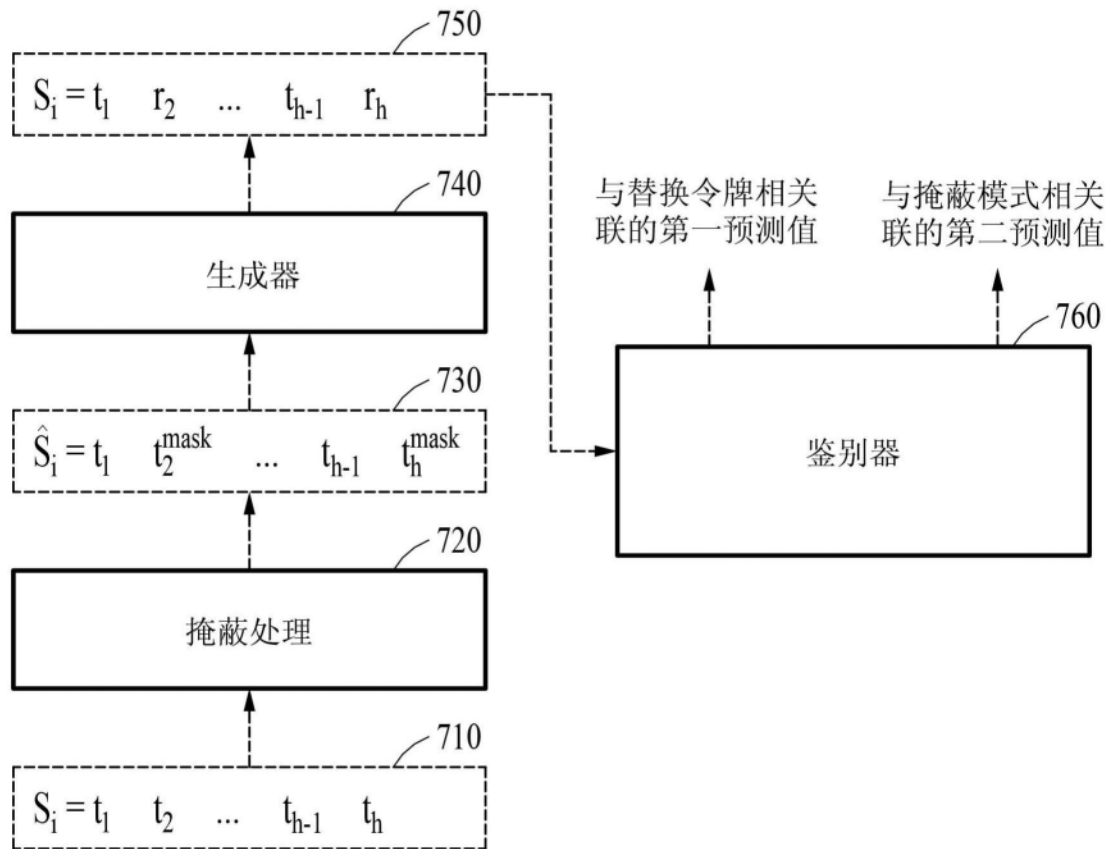


图7

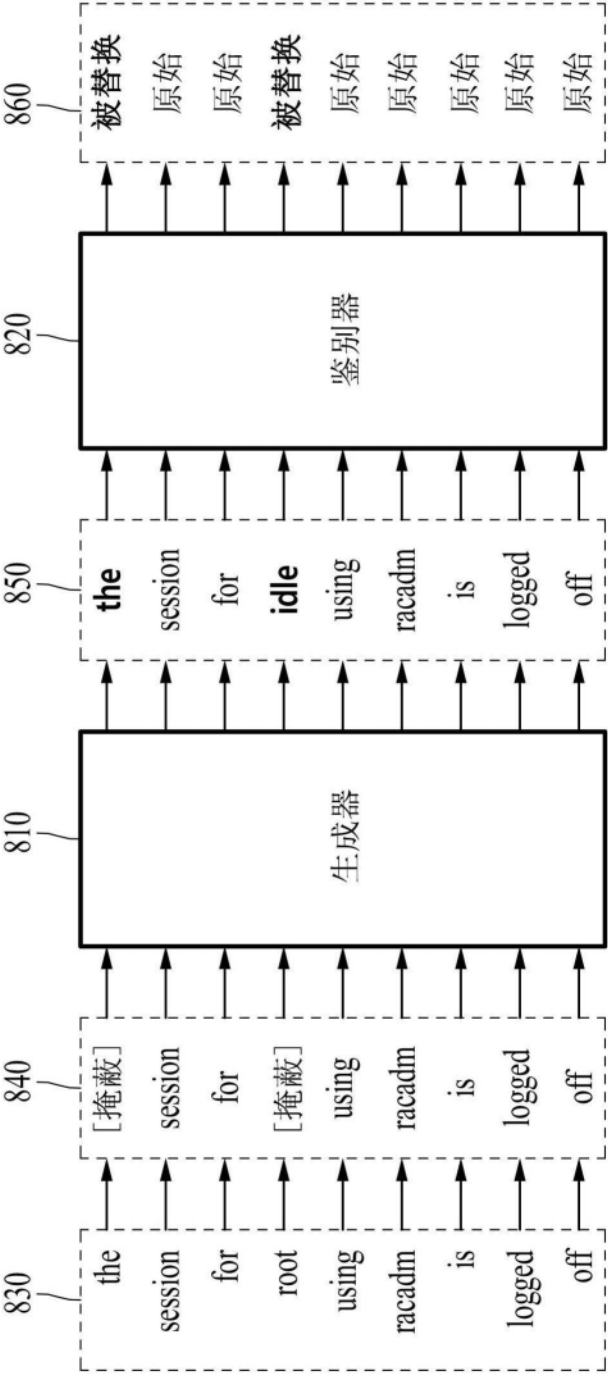


图8

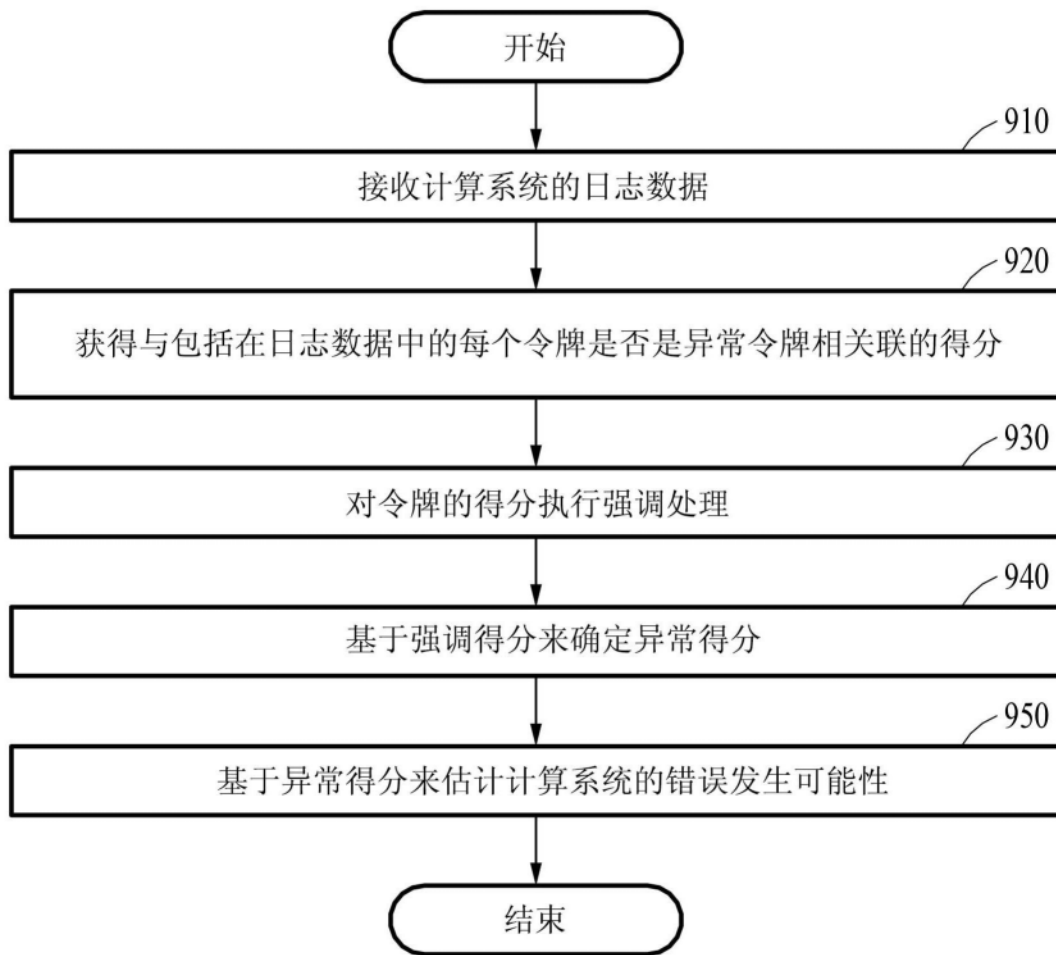


图9

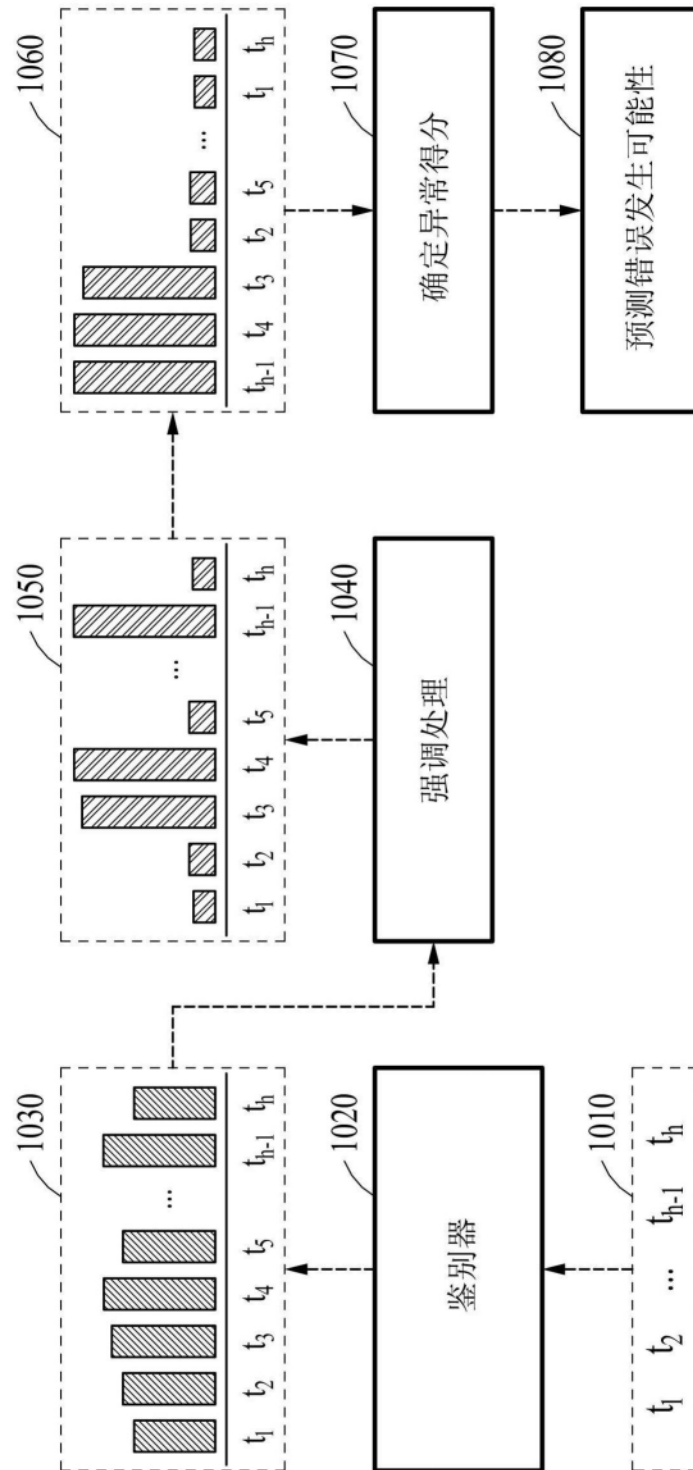


图10

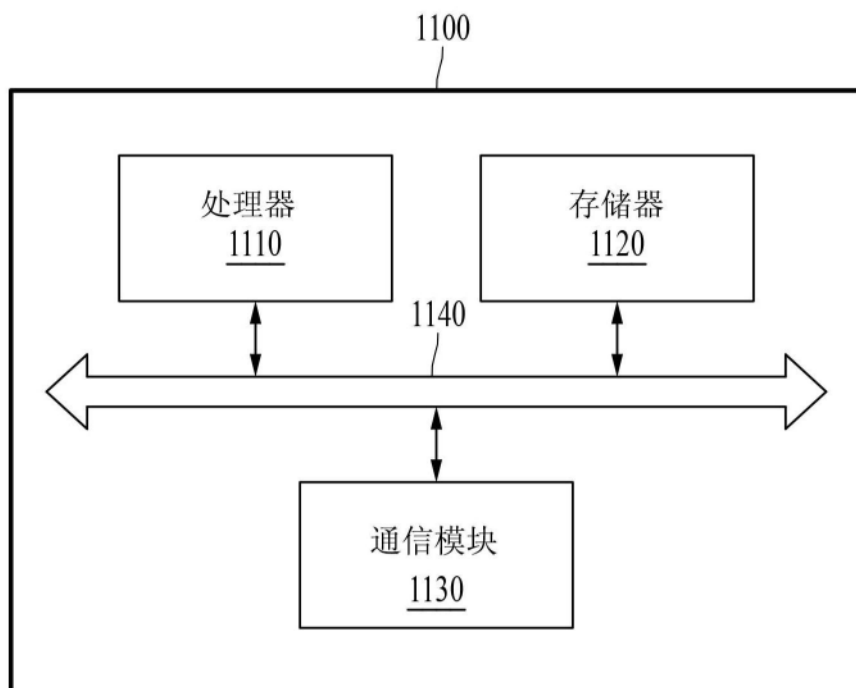


图11

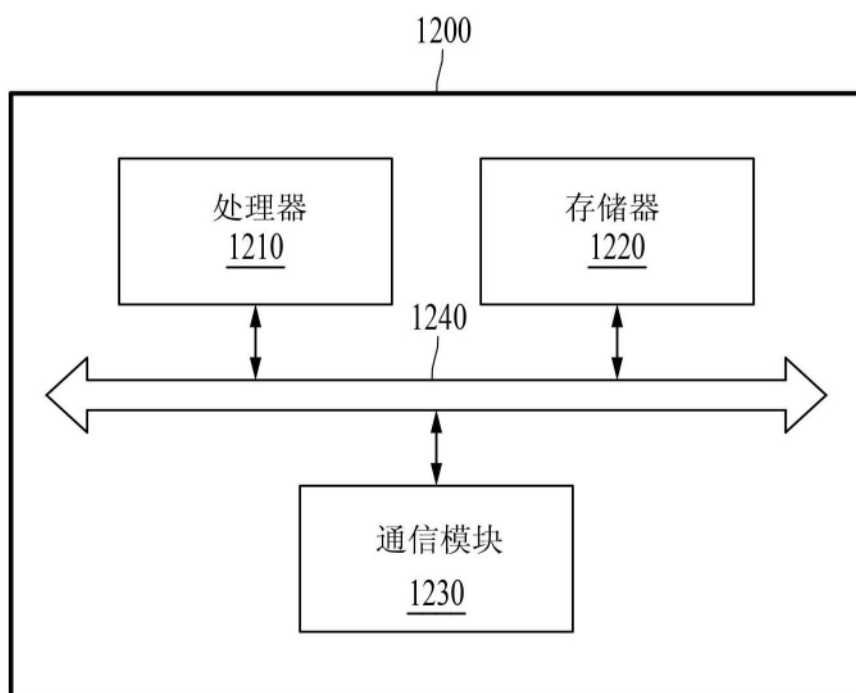


图12