(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 118871930 A
(43)申请公布日 2024.10.29

(21)申请号 202280093638.7
(22)申请日 2022.11.21
(30)优先权数据
17/654,891 2022.03.15 US
(85)PCT国际申请进入国家阶段日
2024.09.13
(86)PCT国际申请的申请数据
PCT/US2022/050531 2022.11.21
(87)PCT国际申请的公布数据
WO2023/177426 EN 2023.09.21
(71)申请人 甲骨文国际公司
地址 美国
(72)发明人 A·H·雷扎伊安 A·波莱里
(74)专利代理机构 中国贸促会专利商标事务所有限公司 11038
专利代理师 冯薇
(51)Int.Cl.
G06N 3/0895(2006.01)
G06N 3/044(2006.01)
G06N 3/045(2006.01)
G06N 3/0464(2006.01)
G06N 3/084(2006.01)

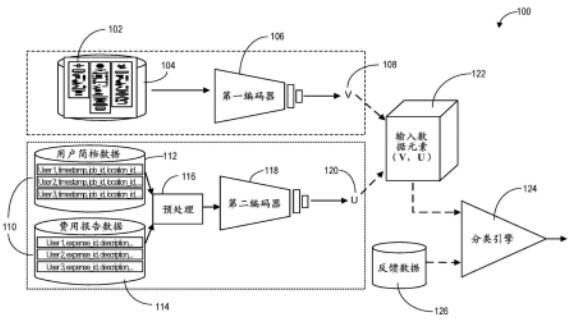
权利要求书4页 说明书29页 附图11页

(54)发明名称

使用对比学习进行异常事件预测

(57)摘要

各种技术可以包括用于使用对比学习来预测数据处理系统中的异常事件的系统和方法。该方法可以包括访问非结构化数据文件和与非结构化数据文件相关联的情境数据。该方法还可以包括为非结构化数据文件生成事件数据输入元素。事件数据输入元素可以包括特征向量集合。特征向量集合可以包括通过使用第一编码器处理非结构化文件而生成的第一特征向量和通过使用第二编码器处理情境数据而生成的第二特征向量。该方法还可以包括通过使用机器学习模型处理事件数据输入元素来生成非结构化数据文件的分类结果,其中分类结果包括对特定事件是否对应于异常事件的预测。



CN 118871930 A

1.一种方法,包括:

访问非结构化数据文件,其中非结构化数据文件识别特定事件的发生;

访问与非结构化数据文件相关联的情境数据,其中情境数据包括描述所述特定事件发生的环境的信息;

为非结构化数据文件生成事件数据输入元素,其中事件数据输入元素包括特征向量集合,并且其中事件数据输入元素通过以下方式生成:

通过使用第一编码器处理非结构化数据文件来生成所述特征向量集合中的第一特征向量,其中第一特征向量表示第一特征空间中的非结构化数据文件;

通过使用第二编码器处理情境数据来生成所述特征向量集合中的第二特征向量,其中第二特征向量表示第二特征空间中的情境数据,

其中第一编码器的一个或多个参数和第二编码器的一个或多个参数是通过以下方式学习的:

(i)变换第一特征空间中的第一训练特征向量以在共享特征空间中生成表示训练非结构化数据文件的第一训练变换特征向量,其中训练非结构化数据文件识别训练事件;

(ii)变换第二特征空间中的第二训练特征向量以在共享特征空间中生成表示训练情境数据的第二训练变换特征向量,其中训练情境数据识别相同的训练事件;以及

(iii)基于第一训练变换特征向量和第二训练变换特征向量在共享特征空间中生成向量,其中所述生成包括组合来自第一训练变换特征向量和第二训练变换特征向量的信息以产生与单独的第一训练特征向量或第二训练特征向量相比具有改进的投影预测性的该向量;

通过使用机器学习模型处理事件数据输入元素来生成非结构化数据文件的分类结果,其中分类结果包括对所述特定事件是否对应于异常事件的预测;以及

输出非结构化数据文件的分类结果。

2.如权利要求1所述的方法,还包括:

确定第一特征向量或第二特征向量偏离预期向量值;

响应于确定第一特征向量或第二特征向量偏离预期向量值,接收与情境数据相关联的反馈数据,其中生成分类结果还包括使用所述机器学习模型处理反馈数据。

3.如权利要求1所述的方法,其中第一编码器是使用训练非结构化数据文件集合训练的,其中第一编码器的训练包括:

增强所述训练非结构化数据文件集合中的训练非结构化数据文件以生成第一训练增强非结构化数据文件和第二训练增强非结构化数据文件;

在第一特征空间中生成第一训练增强非结构化数据文件的第三训练特征向量;

在第一特征空间中生成第二训练增强非结构化数据文件的第四训练特征向量;

确定第三训练特征向量和第四训练特征向量之间的损失;以及

基于所确定的损失调整第一编码器的一个或多个参数。

4.如权利要求1所述的方法,其中第二编码器是使用训练情境数据对象集合训练的,其中第二编码器的训练包括:

增强所述训练情境数据对象集合中的训练情境数据对象以生成第一训练增强情境数据对象和第二训练增强情境数据对象;

在第二特征空间中生成第一训练增强情境数据对象的第五训练特征向量;

在第二特征空间中生成第二训练增强情境数据对象的第六训练特征向量;

确定第五训练特征向量和第六训练特征向量之间的另一个损失;以及

基于所述另一个损失调整第二编码器的一个或多个参数。

5.如权利要求1所述的方法,其中第一编码器包括卷积神经网络。

6.如权利要求1所述的方法,其中所述机器学习模型包括基于树的分类器。

7.如权利要求1所述的方法,其中所述情境数据包括从第一数据存储库访问的第一数据记录集合和从第二数据存储库访问的第二数据记录集合,其中所述情境数据是通过应用一个或多个预处理操作来合并第一数据记录集合和第二数据记录集合中的一个或多个重叠数据记录而生成的。

8.如权利要求1所述的方法,其中所述分类结果还包括所述特定事件的严重性分数,其中所述严重性分数预测所述特定事件是否对应于特定类型的异常事件。

9.一种有形地实施在非暂态机器可读存储介质中的计算机程序产品,包括被配置为使得一个或多个数据处理器执行操作的指令,所述操作包括:

访问非结构化数据文件,其中非结构化数据文件识别特定事件的发生;

访问与非结构化数据文件相关联的情境数据,其中情境数据包括描述所述特定事件发生的环境的信息;

为非结构化数据文件生成事件数据输入元素,其中事件数据输入元素包括特征向量集合,并且其中事件数据输入元素通过以下方式生成:

通过使用第一编码器处理非结构化数据文件来生成所述特征向量集合中的第一特征向量,其中第一特征向量表示第一特征空间中的非结构化数据文件;

通过使用第二编码器处理情境数据来生成所述特征向量集合中的第二特征向量,其中第二特征向量表示第二特征空间中的情境数据,

其中第一编码器的一个或多个参数和第二编码器的一个或多个参数是通过以下方式学习的:

(i)变换第一特征空间中的第一训练特征向量以在共享特征空间中生成表示训练非结构化数据文件的第一训练变换特征向量,其中训练非结构化数据文件识别训练事件;

(ii)变换第二特征空间中的第二训练特征向量以在共享特征空间中生成表示训练情境数据的第二训练变换特征向量,其中训练情境数据识别相同的训练事件;以及

(iii)基于第一训练变换特征向量和第二训练变换特征向量在共享特征空间中生成向量,其中所述生成包括组合来自第一训练变换特征向量和第二训练变换特征向量的信息以产生与单独的第一训练特征向量或第二训练特征向量相比具有改进的投影预测性的该向量;

通过使用机器学习模型处理事件数据输入元素来生成非结构化数据文件的分类结果,其中分类结果包括对所述特定事件是否对应于异常事件的预测;以及

输出非结构化数据文件的分类结果。

10.如权利要求9所述的计算机程序产品,其中所述指令还使得一个或多个数据处理器执行包括以下的操作:

确定第一特征向量或第二特征向量偏离预期向量值;

响应于确定第一特征向量或第二特征向量偏离预期向量值,接收与情境数据相关联的反馈数据,其中生成分类结果还包括使用所述机器学习模型处理反馈数据。

11.如权利要求9所述的计算机程序产品,其中第一编码器是使用训练非结构化数据文件集合训练的,其中第一编码器的训练包括:

增强所述训练非结构化数据文件集合中的训练非结构化数据文件以生成第一训练增强非结构化数据文件和第二训练增强非结构化数据文件;

在第一特征空间中生成第一训练增强非结构化数据文件的第三训练特征向量;

在第一特征空间中生成第二训练增强非结构化数据文件的第四训练特征向量;

确定第三训练特征向量和第四训练特征向量之间的损失;以及

基于所确定的损失调整第一编码器的一个或多个参数。

12.如权利要求9所述的计算机程序产品,其中第二编码器是使用训练情境数据对象集合训练的,其中第二编码器的训练包括:

增强所述训练情境数据对象集合中的训练情境数据对象以生成第一训练增强情境数据对象和第二训练增强情境数据对象;

在第二特征空间中生成第一训练增强情境数据对象的第五训练特征向量;

在第二特征空间中生成第二训练增强情境数据对象的第六训练特征向量;

确定第五训练特征向量和第六训练特征向量之间的另一个损失;以及

基于所述另一个损失调整第二编码器的一个或多个参数。

13.如权利要求9所述的计算机程序产品,其中第一编码器包括卷积神经网络。

14.如权利要求9所述的计算机程序产品,其中所述分类结果还包括所述特定事件的严重性分数,其中所述严重性分数预测所述特定事件是否对应于特定类型的异常事件。

15.一种系统,包括:

一个或多个数据处理器;以及

包含指令的非暂态计算机可读存储介质,所述指令在所述一个或多个数据处理器上执行时,使得所述一个或多个数据处理器执行包括以下的操作:

访问非结构化数据文件,其中非结构化数据文件识别特定事件的发生;

访问与非结构化数据文件相关联的情境数据,其中情境数据包括描述所述特定事件发生的环境的信息;

为非结构化数据文件生成事件数据输入元素,其中事件数据输入元素包括特征向量集合,并且其中事件数据输入元素通过以下方式生成:

通过使用第一编码器处理非结构化数据文件来生成所述特征向量集合中的第一特征向量,其中第一特征向量表示第一特征空间中的非结构化数据文件;

通过使用第二编码器处理情境数据来生成所述特征向量集合中的第二特征向量,其中第二特征向量表示第二特征空间中的情境数据,

其中第一编码器的一个或多个参数和第二编码器的一个或多个参数是通过以下方式学习的:

(i)变换第一特征空间中的第一训练特征向量以在共享特征空间中生成表示训练非结构化数据文件的第一训练变换特征向量,其中训练非结构化数据文件识别训练事件;

(ii)变换第二特征空间中的第二训练特征向量以在共享特征空间中生成表示训练情

境数据的第二训练变换特征向量,其中训练情境数据识别相同的训练事件;以及

(iii)基于第一训练变换特征向量和第二训练变换特征向量在共享特征空间中生成向量,其中所述生成包括组合来自第一训练变换特征向量和第二训练变换特征向量的信息以产生与单独的第一训练特征向量或第二训练特征向量相比具有改进的投影预测性的该向量;

通过使用机器学习模型处理事件数据输入元素来生成非结构化数据文件的分类结果,其中分类结果包括对所述特定事件是否对应于异常事件的预测;以及

输出非结构化数据文件的分类结果。

16.如权利要求15所述的系统,其中所述指令还使得一个或多个数据处理器执行包括以下的操作:

确定第一特征向量或第二特征向量偏离预期向量值;

响应于确定第一特征向量或第二特征向量偏离预期向量值,接收与情境数据相关联的反馈数据,其中生成分类结果还包括使用所述机器学习模型处理反馈数据。

17.如权利要求15所述的系统,其中第一编码器是使用训练非结构化数据文件集合训练的,其中第一编码器的训练包括:

增强所述训练非结构化数据文件集合中的训练非结构化数据文件以生成第一训练增强非结构化数据文件和第二训练增强非结构化数据文件;

在第一特征空间中生成第一训练增强非结构化数据文件的第三训练特征向量;

在第一特征空间中生成第二训练增强非结构化数据文件的第四训练特征向量;

确定第三训练特征向量和第四训练特征向量之间的损失;以及

基于所确定的损失调整第一编码器的一个或多个参数。

18.如权利要求15所述的系统,其中第二编码器是使用训练情境数据对象集合训练的,其中第二编码器的训练包括:

增强所述训练情境数据对象集合中的训练情境数据对象以生成第一训练增强情境数据对象和第二训练增强情境数据对象;

在第二特征空间中生成第一训练增强情境数据对象的第五训练特征向量;

在第二特征空间中生成第二训练增强情境数据对象的第六训练特征向量;

确定第五训练特征向量和第六训练特征向量之间的另一个损失;以及

基于所述另一个损失调整第二编码器的一个或多个参数。

19.如权利要求15所述的系统,其中第一编码器包括卷积神经网络。

20.如权利要求15所述的系统,其中所述分类结果还包括所述特定事件的严重性分数,其中所述严重性分数预测所述特定事件是否对应于特定类型的异常事件。

使用对比学习进行异常事件预测

[0001] 相关申请的交叉引用

[0002] 本申请要求于2022年3月15日提交的编号为17/654,891的标题为“Anomalous Event Prediction Using Contrastive Learning”的美国临时专利申请的优先权,该申请的内容出于所有目的在此通过引用全文并入本文。

技术领域

[0003] 本公开一般而言涉及预测数据处理系统中的异常事件,并且更具体地,涉及使用对比学习(contrastive learning)来预测数据处理系统中的异常事件的技术。

背景技术

[0004] 异常事件可以指与系统已经辨别出的任何现有类别或分类无关联的事件。异常事件(例如,不可辨别的交易)可以基于各种因素来预测,其中因素诸如事件发生的位置和时间、关于触发事件的实体的信息、以及事件期间交换的通信内容。但是,使用现有的机器学习技术预测异常事件可能具有挑战性。现有的机器学习技术可以包括使用由其对应的训练标签识别的训练数据来训练机器学习模型。在给定域中存在各种类型的异常事件的情况下,训练标签可能无法表示所有这些类型的异常事件。实际上,机器学习模型的预测准确性可能不是最优的。

发明内容

[0005] 在一些实施例中,提供使用对比学习来预测数据处理系统中的异常事件的系统和方法。该方法可以包括访问非结构化数据文件。非结构化数据文件可以识别特定事件的发生。该方法还可以包括访问与非结构化数据文件相关联的情境(context)数据。情境数据可以包括描述特定事件发生的环境的信息。

[0006] 该方法还可以包括为非结构化数据文件生成事件数据输入元素。事件数据输入元素可以包括特征向量集合。事件数据输入元素可以通过以下方式生成:(1)通过使用第一编码器处理非结构化数据文件来生成特征向量集合中的第一特征向量;以及(2)通过使用第二编码器处理情境数据来生成特征向量集合中的第二特征向量。在一些情况下,第一特征向量表示第一特征空间中的非结构化数据文件,并且第二特征向量表示第二特征空间中的情境数据。

[0007] 可以通过以下方式来学习第一编码器的一个或多个参数和第二编码器的一个或多个参数:(i)变换第一特征空间中的第一训练特征向量以在共享特征空间中生成第一训练变换特征向量,该第一训练变换特征向量表示训练非结构化数据文件,其中训练非结构化数据文件识别训练事件;(ii)变换第二特征空间中的第二训练特征向量以在共享特征空间中生成第二训练变换特征向量,该第二训练变换特征向量表示训练情境数据,其中训练情境数据识别相同的训练事件;以及(iii)基于第一训练变换特征向量和第二训练变换特征向量在共享特征空间中生成向量。向量的生成可以包括将来自第一训练变换特征向量和

第二训练变换特征向量的信息组合以产生与单独的第一训练特征向量或第二训练特征向量相比具有改进的投影预测性的该向量。

[0008] 该方法还可以包括通过使用机器学习模型处理事件数据输入元素来生成非结构化数据文件的分类结果。分类结果可以包括对特定事件是否对应于异常事件的预测。该方法还可以包括输出非结构化数据文件的分类结果。

[0009] 本公开的一些实施例包括一种包括一个或多个数据处理器的系统。在一些实施例中,该系统包括非暂态计算机可读存储介质,该非暂态计算机可读存储介质包含指令,该指令在一个或多个数据处理器上执行时,使得该一个或多个数据处理器执行本文公开的一个或多个方法的部分或全部和/或一个或多个处理的部分或全部。本公开的一些实施例包括有形地实施在非暂态机器可读存储介质中的计算机程序产品,该非暂态机器可读存储介质包括被配置为使得一个或多个数据处理器执行本文公开的一个或多个方法的部分或全部和/或一个或多个处理的部分或全部的指令。

[0010] 参考以下说明书、权利要求书和附图,前述内容以及其他特征和实施例将变得更加明显。

附图说明

[0011] 图1图示了根据一些实施例的使用对比学习进行异常事件预测的示例计算环境。

[0012] 图2图示了根据一些实施例的使用卷积神经网络生成非结构化文件的特征向量的示例示意图。

[0013] 图3图示了根据一些实施例的使用对比学习训练第一编码器的示意图。

[0014] 图4图示了根据一些实施例的循环神经网络的示例示意图。

[0015] 图5图示了根据一些实施例的使用对比学习训练第二编码器的示意图。

[0016] 图6图示了根据一些实施例的使用对比学习附加地训练第一编码器和第二编码器的示意图。

[0017] 图7图示了根据一些实施例的训练分类引擎以预测异常事件的示意图。

[0018] 图8包括流程图,该流程图图示了根据一些实施例的使用对比学习预测异常事件的方法的示例。

[0019] 图9图示了简化图,该简化图图示了用于实现实施例中的一个实施例的分布式系统。

[0020] 图10图示了简化框图,该简化框图图示了系统环境的一个或多个组件。

[0021] 图11图示了其中可以实现本发明的各种实施例的示例性计算机系统。

[0022] 根据某些示例实施方式,各个附图中的相同附图标记指示相同的元素。此外,元素的多个实例可以通过在该元素的第一个数字后跟字母或连字符和第二个数字来指示。

具体实施方式

[0023] I.概述

[0024] 以下描述中给出了具体细节,以提供对实施例的透彻理解。但是,将理解的是,可以在没有这些具体细节的情况下实践实施例。例如,电路、系统、网络、处理和其他组件可以以框图形式示出为组件,以免不必要的细节使实施例变得模糊。在其他情况下,可以示出众

所周知的电路、处理、算法、结构和技术，而无需不必要的细节，以避免使实施例变得模糊。

[0025] 此外，应注意的是，各个实施例可以被描述为被描绘为流程图、流图、数据流图、结构图或框图的处理。虽然流程图或图可以将操作描述为顺序处理，但是许多操作可以并行或并发地执行。此外，可以重新布置操作的顺序。当处理的操作完成时，该处理终止，但可能具有图中未包括的附加步骤。处理可以对应于方法、函数、过程、子例程、子程序等。当处理对应于函数时，其终止可以对应于函数返回到调用函数或主函数。

[0026] 后续描述仅提供优选示例性实施例，并非旨在限制本公开的范围、适用性或配置。相反，对优选示例性实施例的后续描述将为本领域技术人员提供实现各种实施例的可行描述。应该理解的是，在不脱离所附权利要求中阐述的精神和范围的情况下，可以对元素的功能和布置进行各种改变。

[0027] 以下描述中给出了具体细节，以提供对实施例的透彻理解。但是，将理解的是，可以在没有这些具体细节的情况下实践实施例。例如，电路、系统、网络、处理和其他组件可以以框图形式示出为组件，以免不必要的细节使实施例变得模糊。在其他情况下，可以示出众所周知的电路、处理、算法、结构和技术，而无需不必要的细节，以避免使实施例变得模糊。

[0028] 此外，应注意的是，各个实施例可以被描述为被描绘为流程图、流图、数据流图、结构图或框图的处理。虽然流程图或图可以将操作描述为顺序处理，但是许多操作可以并行或并发地执行。此外，可以重新布置操作的顺序。当处理的操作完成时，该处理终止，但可能具有图中未包括的附加步骤。处理可以对应于方法、函数、过程、子例程、子程序等。当处理对应于函数时，其终止可以对应于函数返回到调用函数或主函数。

[0029] 提供了使用对比学习预测异常事件的技术、系统和方法。对比学习是自监督机器学习技术，其中可以使用既未标记也未注释的训练数据来训练机器学习模型。对于训练数据中的每个训练元素，可以应用数据增强，通过应用变换来生成“相似”训练元素。然后，可以训练编码器以为每个个体训练元素生成特征向量，其中定义损失以惩罚相似训练元素的特征向量位置在给定特征空间中相距远的情况和/或不同训练元素的特征向量位置在给定特征空间中彼此靠近的情况。然后可以使用经训练的编码器为一个或多个非训练数据元素中的每一个生成特征向量，并且可以基于特定特征向量与每个现有特征向量之间的损失来确定每个个体训练数据元素是否表示异常事件。对比学习的优势在于可以预测各种类型的异常事件，同时减少为训练机器学习模型而附加地需要生成目标标签。

[0030] 异常事件预测系统可以访问识别特定事件发生的非结构化数据文件。异常事件预测系统可以是与相同或不同通信网络中的其他系统和/或客户端设备通信以访问非结构化数据文件的计算机服务器。非结构化数据文件可以包括非结构化数据。非结构化数据可以包括不具有预定义数据模型或未以预定义方式组织的信息。非结构化数据文件可以对应于特定文件格式，诸如电子邮件文件格式、图像文件格式、网页格式、文档文件格式、音频文件格式、视频文件格式、以及本领域普通技术人员可以设想的其他类型的文件格式。

[0031] 异常事件预测系统然后可以解析或以其他方式分析来自非结构化数据文件的内容以识别特定事件的发生。在一些情况下，异常事件预测系统通过使用光学字符辨别来处理图像来提取文本数据，其中文本数据识别特定事件的发生。特定事件可以是识别涉及实体(例如，组织的员工)的一个或多个动作的任何类型的事件。例如，特定事件可以是实体对特定物品的购买，并且非结构化数据文件可以是与此类购买相关联的收据的图像。在另一

个示例中,特定事件可以识别实体与另一个实体之间的电话交谈的发生,并且非结构化数据文件可以是此类电话交谈的语音记录。

[0032] 异常事件预测系统还可以访问与非结构化数据文件相关联的情境数据。情境数据包括描述特定事件发生的环境的任何信息。情境数据可以包括非结构化数据,包括非结构化文本。在一些情况下,情境数据包括根据对应数据库模式或文档对象模型格式化的结构化数据。情境数据可以与非结构化数据一起使用以进一步增强机器学习模型的准确性,因为情境数据提供了可能与预测异常事件相关的附加数据点。情境数据还可以促进对异常事件的预测,而无需为训练数据(例如,训练非结构化数据文件)生成目标标签。异常事件预测系统可以从一个或多个数据存储库访问情境数据,一个或多个数据存储库包括存储特定事件中涉及的实体的用户简档数据的用户简档数据存储库。情境数据可以包括但不限于异常事件检测系统可以访问的以下信息:(i)特定事件发生的位置;(ii)特定事件中涉及的实体;(iii)被预测为与特定事件相关的另一个先前或随后的事件;和/或(iv)与和特定事件相关联的交易相关联的类型或类别。

[0033] 异常事件预测系统然后可以生成事件数据输入元素。事件数据输入元素可以是数据结构(例如,元组),其包括特征向量集合,该特征向量集合可以用于预测特定事件是否对应于异常事件。异常事件预测系统可以通过使用一个或多个机器学习模型处理非结构化数据文件和情境数据来生成事件数据输入元素的特征向量集合。

[0034] 在一些实施例中,异常事件预测系统通过使用第一编码器处理非结构化数据文件来生成表示第一特征空间中的非结构化数据文件的第一特征向量。编码器可以包括机器学习模型,该机器学习模型处理输入以生成可以在对应特征空间中表示的特征向量。在一些情况下,第一编码器是卷积神经网络或多层感知器神经网络。如本文所使用的,特征向量是指共同表示对象的值的n维向量。特征空间可以是特征向量可以通过其对应位置识别的n维空间。例如,第一特征向量可以是包括表示非结构化数据文件的值集合的n维向量,其中该集合中的每个值识别第一特征向量在第一特征空间的维度上的对应位置。可以基于在第一特征空间内确定的第一训练特征向量集合中的两个训练特征向量之间的损失来学习第一编码器的一个或多个参数,其中第一训练特征向量集合中的训练特征向量可以表示与第一集合中的另一个训练特征向量的训练非结构化数据文件相同或相似的训练非结构化数据文件。因此,第一编码器的训练可以基于确定的第一特征向量与第一训练特征向量集合中的每一个向量之间的损失来促进非结构化数据文件的特性的识别。在一些情况下,第一编码器包括用于确定损失的第一对比损失函数,其中该损失可以用于调整第一编码器的一个或多个参数。

[0035] 异常事件预测系统还可以通过使用第二编码器处理情境数据来生成表示第二特征空间中的情境数据的第二特征向量。第二特征向量可以是包括表示情境数据的值集合的n维向量,其中该集合中的每个值识别第二特征向量在第二特征空间的维度上的对应位置。第二编码器可以具有与第一编码器的体系架构相同或不同的体系架构。因此,可以训练第一编码器和第二编码器中的每一个以在其相应的特征空间中生成特征向量,并且第一编码器和第二编码器可以不共享任何参数。可以基于在第二特征空间内确定的第二训练特征向量集合中的两个训练特征向量之间的另一个损失来学习第二编码器的一个或多个参数,其中第二训练特征向量集合中的训练特征向量可以表示与第二集合中的另一个训练特征向

量的训练情境数据相同或相似的训练情境数据。因此,第二编码器的训练可以基于第二特征向量与第一训练特征向量集合中的每一个向量之间的损失来促进情境数据的特性的识别。第二编码器可以包括第二对比损失函数,该第二对比损失函数可以用于确定第二训练特征向量集合中的两个训练特征向量之间的另一个损失,其中该另一个损失可以用于调整第二编码器的一个或多个参数。

[0036] 但是,对于每个特征向量使用单独的特征空间可能导致各种问题。例如,若干特征向量可能在相应的特征空间中被视为"相似",不是因为它们对应于相似的事件,而是因为它们共享切向特性(例如,相似的购买商店、相似的日期和时间)。为了确保第一特征向量和第二特征向量在其相应的特征空间中表示相同的特定事件,可以实现共享特征空间。第一编码器和第二编码器的训练可以附加地基于在共享特征空间中确定的识别相同事件的两个训练特征向量之间的损失来执行。第一编码器和第二编码器在共享特征空间中的对比学习可以减少依赖于切向特性的特征向量的过度拟合。此外,共享特征空间中的对比学习可以促进对指示异常事件的特定数据项(例如,非结构化数据文件、情境数据)的预测。

[0037] 在一些实施例中,进一步学习第一编码器的一个或多个参数和第二编码器的一个或多个参数,以基于第一训练变换特征向量和第二训练变换特征向量在共享特征空间中生成向量。可以通过组合来自第一训练变换特征向量和第二训练变换特征向量的信息来生成向量,从而产生与单独的第一训练特征向量或第二训练特征向量相比具有改进的投影预测性的该向量。可以通过将第一投影函数应用于表示识别训练事件的训练非结构化数据文件的第一训练特征向量来生成第一训练变换特征向量。类似地,可以通过将第二投影函数应用于表示识别相同训练事件的训练情境数据的第二训练特征向量来生成第二训练变换特征向量。在一些情况下,第一对比损失函数和第二对比损失函数被组合并通过一个或多个加权因子加权,使得可以使用组合损失函数来确定损失。然后可以使用确定的损失来调整第一编码器和第二编码器的相应参数。加权因子可以是可以在第一编码器和第二编码器的训练期间进行微调的超参数。因此,第一编码器和第二编码器的对比学习可以基于在共享特征空间内确定的损失来确保与非结构化数据文件和情境数据相关联的特征向量对应于相同的特定事件。

[0038] 异常事件预测系统可以通过使用机器学习模型处理事件数据输入元素来生成分类结果。分类结果包括对特定事件是否对应于异常事件的预测。在一些情况下,机器学习模型是神经网络或基于树的分类器模型。分类结果还可以包括具有置信区间的异常事件的严重性分数。严重性分数可以预测特定事件是否对应于需要后续动作(例如,审计)的特定类型的异常事件。然后可以使用严重性分数对预测的异常事件进行分类或聚类。例如,可以对异常事件进行聚类以基于预测的严重性分数对异常事件进行排名。下游管道任务可以使用此信息进行后续动作。异常事件预测系统因此可以预料各种类型的异常事件,即使编码器和机器学习模型是在没有标记的训练数据的情况下进行训练的。在一些情况下,异常事件预测系统进一步处理反馈数据,以增强机器学习模型预测分类结果的准确性。在一些情况下,反馈数据用于调整第一特征向量和/或第二特征向量的值。异常事件预测系统可以输出分类结果。输出的分类结果可以由另一个数据处理系统用于执行后续动作,以发起非结构化数据文件中识别出的特定事件的审计处理。在一些情况下,另一个数据处理系统使用输出的分类结果来拒绝或以其他方式暂停对非结构化数据文件的进一步处理。

[0039] II.使用对比学习进行异常事件预测的计算环境

[0040] 图1图示了根据一些实施例的使用对比学习进行异常事件预测的示例计算环境。计算环境可以包括异常事件预测系统100,其执行一个或多个操作以预测非结构化数据文件中识别出的特定事件是否对应于异常事件。异常事件预测系统100可以是与其他系统和/或客户端设备通信以访问非结构化数据文件的计算机服务器。

[0041] 作为说明性示例,异常事件预测系统100可以与企业资源规划(ERP)费用审计系统进行交互或以其他方式进行通信。ERP系统可以从各种数据源访问与购买闪存存储器设备相关联的数据。示例数据源可以是存储多个收据图像的图像储存库。另一个示例数据源可以是存储多个扫描采购订单的文档储存库。因此,ERP系统可以处理所访问的数据以确定是批准、拒绝、还是进一步处理闪存存储器设备的购买。

[0042] 异常事件预测系统100可以访问非结构化数据文件102。非结构化数据文件102可以包括非结构化数据。非结构化数据可以包括没有预定义数据模型或未以预定义方式组织的信息。非结构化数据文件102可以对应于特定文件格式,包括电子邮件文件格式(例如,MSG)、图像文件格式(例如,JPG、JPEG、PNG、BMP)、网页格式(例如,HTML)、文档文件格式(例如,DOC、PDF、XPS、EPUB)、音频文件格式(例如,MP3、M4A)、视频文件格式(例如,MPEG、MOV)以及本领域普通技术人员可以设想的其他类型的文件格式。继续该示例,非结构化数据文件102可以是识别由实体购买闪存存储器设备的收据图像文件。异常事件检测系统可以从存储在图像储存库104中的其他收据图像访问该收据图像。

[0043] 异常事件预测系统100可以解析或以其他方式分析非结构化数据文件中的内容以识别特定事件的发生。继续以上示例,异常事件预测系统100可以通过使用光学字符辨别来处理非结构化数据文件102来提取文本数据。文本数据可以识别购买的发生,包括购买的位置、购买的时间和日期、以及用于购买闪存存储器设备的支付信息。

[0044] 异常事件预测系统100的第一编码器106可以生成第一特征向量108,其表示第一特征空间中的非结构化数据文件102。第一编码器106可以包括处理非结构化数据文件102以生成第一特征向量108的第一编码器。第一编码器106可以被训练以生成非结构化数据文件102的特征向量108,其中定义损失以便惩罚相似训练元素(例如,训练非结构化数据文件)的特征向量的位置在第一特征空间中相距远的情况和/或不同训练元素的特征向量的位置在第一特征空间中彼此靠近的情况。

[0045] 第一编码器106可以是卷积神经网络或多层感知器神经网络。第一特征向量108可以是n维向量,其包括表示非结构化数据文件102的值集合,其中该集合中的每个值识别第一特征向量108在第一特征空间的维度上的对应位置。可以基于第一特征向量108与第一特征空间中第一训练特征向量集合中的每一个向量之间的损失来训练第一编码器,以促进非结构化数据文件102的特性的识别。继续以上示例,第一编码器106可以是卷积神经网络,其通过处理从非结构化数据文件102中提取的文本数据来生成第一特征向量108。

[0046] 异常事件预测系统100还可以访问与非结构化数据文件102相关联的情境数据110。情境数据110包括描述特定事件发生的环境的任何信息(例如,特定事件发生的位置、被预测与特定事件相关的另一个先前或后续事件)。情境数据110可以提供附加的数据点,这些附加的数据点可能与处理非结构化数据文件102相关,以预测异常事件。

[0047] 继续以上示例,情境数据110可以包括从用户简档数据存储库112访问的一个或多

个数据记录,其中该一个或多个数据记录识别授权购买闪存存储器设备的实体。情境数据110还可以包括从费用报告数据存储库114访问的一个或多个数据记录,其中该一个或多个数据记录包括代表授权购买闪存存储器设备的实体提交的费用报告。在一些实施例中,从用户简档数据存储库112访问的数据记录包括也可以在从费用报告数据存储库114访问的数据记录中识别的信息(例如,实体识别符)。

[0048] 在一些情况下,情境数据110包括根据对应数据库模式或文档对象模型格式化的结构化数据。例如,情境数据110可以根据用户简档数据存储库112的SQL数据库模式进行格式化,其中该模式定义一个或多个数据字段,诸如“user_id”、“timestamp”、“job_id”、“location_id”、“role_id”、“duty_id”、“previlage_id”和“assignment_id”。在另一个示例中,情境数据110可以根据费用报告数据存储库114的XML文档对象模型进行格式化,其中该文档对象模型定义一个或多个数据字段,诸如“user_id”、“expense_id”和“description”、“justification”、“expense_type”和“expense_category”。

[0049] 在一些情况下,异常事件预测系统100的预处理模块116进一步处理情境数据110。预处理模块116可以对情境数据110应用一个或多个数据操纵(例如,规范化、过滤),以增强处理情境数据110的编码器的性能。例如,预处理模块116可以包括平坦化操作,该操作通过将从一个或多个数据存储库(例如,用户简档数据存储库112、费用报告数据存储库114)访问的数据记录合并到单个或减少数量的表中并移除任何冗余数据来处理情境数据110。继续以上示例,预处理情境数据110可以包括合并来自用户简档数据存储库112和费用报告数据存储库114的数据记录的信息。来自合并数据记录的信息可以由不同的数据库模式定义,该数据库模式包括一个或多个数据字段,诸如“user_id”、“timestamp”、“job_id”、“location_id”、“role_id”、“duty_id”、“previlage_id”、“assignment_id”、“expense_id”以及“description”、“justification”、“expense_type”和“expense_category”。

[0050] 异常事件预测系统100的第二编码器118可以通过处理情境数据110来生成表示第二特征空间中的情境数据110的第二特征向量120。可以训练第二编码器118以生成情境数据110的特征向量120,其中定义另一个损失以便惩罚相似训练元素(例如,训练情境数据)的特征向量的位置在第二特征空间中相距远的情况和/或不同训练元素的特征向量的位置在第二特征空间中彼此靠近的情况。因此,第二编码器118的训练可以基于第二特征向量120与第一训练特征向量集合中的每一个向量之间的损失来促进情境数据110的特性的识别。

[0051] 第二编码器118可以是卷积神经网络或多层感知器神经网络。第二编码器118可以具有与第一编码器106的体系架构相同或不同的体系架构。因此第一编码器106和第二编码器118中的每一个都可以被训练以在其相应的特征空间中生成特征向量108和120,其中第一编码器106和第二编码器118可以不共享任何参数。

[0052] 第二特征向量120可以是n维向量,其包括表示情境数据110的值集合,其中该集合中的每个值识别第二特征向量120在第二特征空间的维度上的对应位置。可以训练第二编码器118以基于第二特征向量120与第二特征空间中第一训练特征向量集合中的每一个向量之间的损失来促进情境数据110的特性的识别。继续以上示例,第二编码器118可以是多层感知器神经网络,其通过处理预处理的情境数据110来生成第二特征向量120。

[0053] 第一编码器106和第二编码器118可以被进一步训练,使得第一特征向量108和第

二特征向量120在其相应的特征空间中表示相同的特定事件。例如,第一编码器106和第二编码器118可以通过在共享特征空间上生成其相应的向量来进一步训练。第一编码器106和第二编码器118的训练可以附加地基于在共享特征空间中识别相同事件的两个训练特征向量之间的损失来执行。在一些情况下,训练包括基于以下各项在共享特征空间中生成向量:(i)通过将第一投影函数应用于第一特征向量108而生成的第一训练变换特征向量;以及(ii)通过将第二投影函数应用于第二特征向量110而生成的第二训练变换特征向量。可以通过组合来自第一训练变换特征向量和第二训练变换特征向量的信息来生成向量,从而得到与单独的第一训练特征向量或第二训练特征向量相比具有改进的投影预测性的该向量。第一编码器和第二编码器在共享特征空间中的对比学习可以减少依赖于切向特性的特征向量的过度拟合。因此,第一编码器106和第二编码器118的对比学习可以确保与非结构化数据文件102和情境数据110相关联的特征向量对应于相同的事件(例如,购买闪存存储器设备)。

[0054] 异常事件预测系统100可以基于第一特征向量108和第二特征向量120生成非结构化数据文件102的事件数据输入元素122。非结构化数据文件102的事件数据输入元素122可以是至少包括第一特征向量108和第二特征向量120的数据结构(例如,元组),其可以由机器学习模型(例如,分类器模型)处理以预测非结构化数据文件102中识别出的特定事件是否对应于异常事件。继续该示例,事件数据输入元素122可以由由“event(u,v)”定义的元组来识别,其中:(1)变量名称“event”识别事件数据输入元素对应于非结构化数据文件102;(2)“u”表示第一特征向量108;并且(3)“v”表示第二特征向量120。因此,事件数据输入元素122的第一特征向量108和第二特征向量120可以表示与和非结构化数据文件102相关联的特定事件(例如,购买闪存存储器设备)相关联的一个或多个特性。

[0055] 分类引擎124可以被配置为使用另一个机器学习模型来处理事件数据输入元素122以生成分类结果。分类结果可以包括对非结构化数据文件102中识别出的特定事件是否对应于异常事件的预测。分类结果还可以包括具有置信区间的异常事件的严重性分数。然后可以使用严重性分数对预测的异常事件进行分类或聚类。例如,可以对异常事件进行聚类以基于预测的严重性分数对异常事件进行排名。下游管道任务可以使用此信息进行后续动作。在一些情况下,机器学习模型是神经网络或基于树的分类器模型。

[0056] 在一些情况下,分类引擎124的机器学习模型附加地处理反馈数据126,以增强分类引擎124的机器学习模型预测分类结果的准确性。在一些情况下,反馈数据126包括响应于异常事件预测系统100预测第一特征向量108和第二特征向量120中的一个偏离预期向量值而由另一个用户手动输入的数据。例如,可以基于以下各项来预测与第一特征向量的第一预期向量值的第一偏差:(1)生成训练第一特征空间的第一特征向量的分布;(2)确定第一特征向量对应于所生成分布的离群值(outlier);以及(3)确定第一特征向量偏离预期向量值。

[0057] 分类引擎124可以输出分类结果。输出的分类结果可以由另一个数据处理系统使用来执行后续动作,以发起非结构化数据文件102中识别出的特定事件的审计处理。例如,审计系统可以对闪存存储器设备的购买进行审查,以确定此类购买是否是授权交易。在一些情况下,另一个数据处理系统使用输出的分类结果来拒绝或以其他方式暂停对非结构化数据文件102的进一步处理。

[0058] III.异常事件预测的训练技术

[0059] 训练用于预测异常事件的机器学习模型可以包括两个处理:(1)训练一个或多个编码器以生成表示给定非结构化数据文件的事件数据输入元素;以及(2)训练另一个机器学习模型以预测非结构化数据文件中识别出的特定事件是否对应于异常事件。关于第一阶段,可以训练一个或多个编码器以生成其相应的特征向量,这些特征向量对与非结构化数据文件(例如,图像收据、视频文件)中识别出的特定事件相关联的一个或多个特性进行编码。可以使用其中训练数据未被标记或注释的无监督学习来训练每个编码器。可以使用来自对应数据源(例如,图像储存库、用户简档数据存储库)的训练数据来训练每个编码器。特征向量可以用于生成事件数据输入元素。

[0060] 关于第二阶段,可以训练机器学习模型来处理事件数据输入元素以生成分类结果。分类结果包括关于特定事件是否对应于异常事件的预测。在一些情况下,机器学习模型的训练包括将反馈数据追加到事件数据输入元素以增强另一个机器学习模型的预测准确性。

[0061] 可以使用不同类型的训练数据来训练编码器和机器学习模型,以预测各个领域的异常事件。例如,可以使用收据图像和费用报告来训练编码器和机器学习模型,以预测费用审计系统的异常事件。在另一个示例中,可以使用传感器数据(例如,文本、语音、图像)来训练编码器和机器学习模型,以预测视频监控系统的异常事件。其他示例可以包括:(1)使用语音记录来训练编码器和机器学习模型,以预测呼叫中心系统的异常事件;以及(2)使用电子邮件文件来训练编码器和机器学习模型,以预测客户支持系统的异常事件。

[0062] A.用于处理非结构化数据文件的第一编码器

[0063] 可以使用未标记的训练数据训练第一编码器以处理特定的非结构化数据文件并生成特征向量。特征向量可以表示第一特征空间中的非结构化数据文件。例如,第一特征向量可以是n维向量,其包括表示特定非结构化数据文件的值集合,其中该集合中的每个值识别第一特征向量在第一特征空间的维度上的对应位置。由经训练的第一编码器生成的第一特征向量可以编码或以其他方式表示与非结构化数据文件中识别出的特定事件相关联的一个或多个特性。

[0064] 1.训练数据

[0065] 在一些实施例中,第一编码器的训练数据包括非结构化数据文件的第一训练集。第一训练集的每个非结构化数据文件可以包括非结构化数据,其包括不具有预定义数据模型或未以预定义方式组织的信息。在一些情况下,第一训练集的每个非结构化数据文件不包括任何训练标签或注释。因此,第一编码器的训练处理包括使用未标记的训练数据训练第一编码器。

[0066] 第一训练集的非结构化数据文件可以对应于特定文件格式,包括电子邮件文件格式(例如,MSG)、图像文件格式(例如,JPG、JPEG、PNG、BMP)、网页格式(例如,HTML)、文档文件格式(例如,DOC、PDF、XPS、EPUB)、音频文件格式(例如,MP3、M4A)、视频文件格式(例如,MPEG、MOV)、以及本领域普通技术人员可以设想的其他类型的文件格式。基于文件格式,可以对第一训练集的每个非结构化数据文件执行一个或多个数据预处理。例如,如果第一训练集的非结构化数据文件对应于语音记录文件格式,那么可以对第一训练集的每个非结构化数据文件应用语音转文本算法以将音频数据转换成文本数据。在另一个示例中,如果第

一训练集的非结构化数据文件对应于视频文件格式,那么可以将视频帧提取算法应用于第一训练集的每个非结构化数据文件,以从视频数据生成图像数据。

[0067] 第一训练集的每个非结构化数据文件可以识别特定领域的对应事件的发生。例如,费用报告领域的第一训练集的非结构化数据文件可以识别实体对网络监视服务的订阅。在另一个示例中,视频监控领域的第一训练集的非结构化数据文件可以识别视频文件中一个或多个对象的外观。

[0068] 2.机器学习模型选择

[0069] 然后可以选择第一编码器的一个或多个机器学习模型进行训练。第一编码器的机器学习模型可以是卷积神经网络、初始神经网络、残差神经网络、循环神经网络、长短期记忆模型、门控循环单元模型或深度神经网络的其他变体。机器学习模型还可以是可以被训练以从非结构化数据文件生成特征向量的任何其他合适的模型,诸如朴素贝叶斯分类器、线性分类器、支持向量机、诸如随机森林模型的装袋模型、增强模型、浅层神经网络、或者一种或多种此类技术的组合——例如,多尺度卷积神经网络。在一些情况下,第一编码器具有暹罗(Siamese)网络体系架构,其中机器学习模型在每次迭代中被训练两次,以惩罚相似训练特征向量的位置在第一特征空间中相距远的情况以及不同训练特征向量的位置在第一特征空间中彼此靠近的情况。

[0070] 第一编码器可以包括卷积神经网络。图2图示了根据一些实施例的使用卷积神经网络生成非结构化文件的特征向量的示例示意图200。为了说明目的,参考图1中所示的组件描述了示意图200中实现的处理,但其他实现也是可能的。例如,图1的存储在非暂态计算机可读介质中的第一编码器106的程序代码由一个或多个处理设备执行,以使这些设备执行本文描述的操作。

[0071] 可以将非结构化数据文件205提供给被配置为应用深度卷积神经网络模型的一个或多个层的第一编码器。非结构化数据文件205可以是收据的图像文件,其可以由像素集合表示,其中每个像素包括颜色值(例如,RGB值)集合。在一些情况下,与给定像素对应的颜色值与(例如)一般强度值和/或与一种或多种颜色中的每一个颜色相关的强度值相关联。

[0072] 第一编码器可以对与非结构化数据文件205对应的像素值执行一系列操作,这些操作形成用于若干类型的卷积神经网络体系架构的单个卷积层:(1)卷积;(2)非线性转换(例如,ReLU);(3)池化或子采样;以及(4)全连接层。在一些情况下,第一编码器的卷积操作通过使用小方块的输入数据来学习图像特征,从而保留像素之间的空间关系。例如,非结构化数据文件205被视为像素值的矩阵,其中矩阵的每个像素区域都可以被指派值。此外,非结构化数据文件205可以是具有范围从0到1的像素值的黑白图像。非结构化数据文件205可以替代地或附加地包括具有三个指派的RGB值的彩色图像,每个RGB值的范围从0到255。

[0073] 在访问非结构化数据文件205之后,第一编码器可以基于非结构化数据文件205生成特征图210。在一些实施例中,特征提取器包括一个或多个内核或过滤器,其包括可以“滑过”与非结构化数据文件205对应的矩阵的值矩阵,以便生成特征图210。例如,内核被配置为检测边缘、锐化图像或识别表示主体的像素集合,以便预处理数据并提供与分类器的图像分析相关的细节。在一些情况下,(例如,基于手动输入)预定义与特征提取对应的超参数集合。特征提取超参数可以识别(例如)每个过滤器的步幅(例如,1步、2步)、填充大小、内核大小和/或内核形状。

[0074] 继续图2的示例,第一编码器可以执行非线性(例如,ReLU)转换,其将特征图210中的所有负像素值替换为零。第一编码器的ReLU操作可以在卷积神经网络模型中引入非线性,因为卷积操作从表示非结构化数据文件205的矩阵值生成线性函数输出。非线性转换可以附加地或替代地包括使用sigmoid函数和/或tanh函数。

[0075] 在ReLU操作之后,第一编码器可以对特征图210执行池化操作,以减小特征图210的空间大小,并且随后生成池化特征图215。在一些实施例中,池化降低了由特征提取器的内核生成的每个特征图的维数,但保留了在非结构化数据文件205中捕获的图像定义特征(例如,图像的边缘)。在一些情况下,第一编码器执行最大池化操作以访问特征图内的区域(例如,特征图内的5x5矩阵)并选择与最高值相关联的元素。可以迭代此操作以遍历整个特征图210,此时最大池化操作完成池化特征图215的生成。第一编码器可以替代地或附加地执行平均池化操作来代替最大池化操作,该最大池化操作选择特征图210内区域中捕获的元素的总和或平均值。通过执行池化操作,第一编码器可以实现若干技术优势,包括能够生成允许减少卷积神经网络模型内参数和计算数量的非结构化数据文件205的输入表示,以及缩放非结构化数据文件205的不变表示。

[0076] 继续图2的示例,第一编码器可以继续对非结构化数据文件205应用一个或多个附加的卷积层,其中可以对池化特征图215执行卷积、ReLU和池化操作。在一些情况下,第二卷积层生成特征图220、池化特征图225等。重复这些操作,直到提取的特征变得与全连接层的缩放和平移等变(equivariant)。

[0077] 在应用若干卷积层之后,第一编码器可以应用全连接层230,该全连接层可以是在输出层235处结合softmax激活函数或其他类型的线性或非线性函数的多层感知器网络。在一些实施例中,第一编码器使用访问提取的特征并且生成表示非结构化数据文件205的第一特征向量的全连接层。在一些情况下,第一特征向量对与从非结构化数据文件205中识别出的特定事件相关联的一个或多个特性(例如,时间、位置、价格)进行编码。第一特征向量可以用于生成事件数据输入元素,该事件数据输入元素后面由另一个机器学习模型处理,该机器学习模型预测在非结构化数据文件205中识别出的特定事件是否对应于异常事件。

[0078] 在一些情况下,第一特征向量包括用于在第一特征空间中识别非结构化数据文件的位置(例如,位置坐标)的图形坐标。在一些情况下,将嵌入层应用于第一特征向量以降低其维数。例如,第一特征向量可以与具有20个维度的值相关联,其中嵌入层可以将值的维度降低到10。

[0079] 将认识到的是,虽然图2描绘了使用两个卷积层来处理非结构化数据文件205,但是可以使用不同数量的卷积层(例如,这可能具有重复这些操作的效果,可以由第一编码器重复一次或多次)。例如,通过应用10个或更多个卷积层来从非结构化数据文件205中提取特征来确定输出。在一些实施例中,对于分类器应用的一个或多个卷积层,省略了池化操作。在一些实施例中,分类器使用不同版本的卷积神经网络模型体系架构,包括但不限于AlexNet、ZFNet、GoogLeNet、VGGNet、ResNets、DenseNet等。

[0080] 3.机器学习模型的训练

[0081] 在选择机器学习模型之后,可以使用训练数据训练第一编码器。可以使用对比学习训练第一编码器以生成表示非结构化数据文件的第一特征向量。因此,对比学习允许使用既未标记也未注释的训练数据训练第一编码器。

[0082] 图3图示了根据一些实施例的使用对比学习训练第一编码器的示意图300。如图3中所示,第一编码器的训练至少包括以下阶段:(1)增强训练非结构化数据文件以生成第一训练增强非结构化数据文件和第二训练增强非结构化数据文件;(2)在第一特征空间中生成第一训练增强非结构化数据文件的第一训练特征向量;(3)在第一特征空间中生成第二训练增强非结构化数据文件的第二训练特征向量;(4)确定第一训练特征向量和第二训练特征向量之间的损失;以及(5)基于确定的损失调整第一编码器的一个或多个参数。

[0083] 可以通过将数据增强302应用于训练数据的训练非结构化数据文件304以生成第一增强训练非结构化数据文件306和第二增强训练非结构化数据文件308来发起第一编码器的训练。数据增强302可以包括变换训练非结构化数据文件304以生成训练非结构化数据文件的两个或更多个修改版本。数据增强确保两个或更多个修改版本与训练非结构化数据文件304“相似”。

[0084] 修改版本可以用于执行第一编码器310的对比学习。例如,训练图像(例如,收据图像)的数据增强可以包括几何变换、翻转、颜色修改、裁剪、旋转、噪声注入和随机擦除。在另一个示例中,训练音频文件(例如,客户表示的语音记录)的数据增强可以包括噪声注入、时间移位、音调修改和回放速度修改。在又一个示例中,训练视频文件的数据增强可以包括分段仿射变换、超像素处理、高斯模糊、颜色反转、随机旋转、随机调整大小、平移、帧添加、帧移位、翻转、下采样和上采样。在图3中,第一增强训练非结构化数据文件306描绘了训练非结构化数据文件304的旋转版本,并且第二增强训练非结构化数据文件308描绘了训练非结构化数据文件304的颜色反转版本。

[0085] 第一编码器310可以将第一增强训练非结构化数据文件306处理为第一特征空间316中的第一训练特征向量312。第一训练特征向量312可以是n维值向量,这些值共同表示第一增强训练非结构化数据文件306。第一特征空间316可以是n维空间,其中第一训练特征向量312可以通过其对应位置来识别。在一些情况下,第一编码器的卷积神经网络应用于第一增强训练非结构化数据文件306以生成第一训练特征向量312。

[0086] 第一编码器310还可以将第二增强训练非结构化数据308文件处理为第一特征空间316中的第二训练特征向量314。与第一训练特征向量312相似,第二训练特征向量314可以是n维值向量,这些值共同表示第二增强训练非结构化数据文件308。在一些情况下,第一编码器310的卷积神经网络还应用于第二增强训练非结构化数据文件308以生成第二训练特征向量。第一编码器的训练可以被执行为使得第一训练特征向量312和第二训练特征向量均可以对训练非结构化数据文件的一个或多个特性进行编码。

[0087] 可以确定第一训练特征向量312和第二训练特征向量314之间的损失318。所确定的损失318可以识别第一编码器310在训练阶段期间的性能水平。例如,如果第一编码器预测第一训练特征向量312和第二训练特征向量314彼此相似,那么所确定的损失318可以减小。相反,如果第一编码器预测第一训练特征向量312和第二训练特征向量314彼此不同,那么所确定的损失318可以增加。损失可以通过以下方式确定:(1)确定第一训练特征向量和第二训练特征向量之间的损失(例如,余弦相似度、距离);以及(2)将第一对比损失函数应用于所确定的损失。在一些情况下,第一对比损失函数是交叉熵损失函数,其被定义为惩罚第一训练特征向量312和第二训练特征向量314的位置在第一特征空间316中相距远的情况。

[0088] 第一编码器可以基于确定的损失318调整其参数320中的一个或多个。可以执行调整第一编码器的参数以最小化第一特征空间316内第一训练特征向量和第二训练特征向量之间的损失。通过识别没有标签或注释的“相似”非结构化数据文件,第一编码器的训练可以促进识别特定非结构化数据文件的特性。

[0089] B.用于处理情境数据的第二编码器

[0090] 与第一编码器相似,第二编码器可以使用未标记的训练数据进行训练,以处理情境数据并生成另一个特征向量。另一个特征向量可以表示第二特征空间中的情境数据。例如,第二特征向量可以是n维向量,其包括表示情境数据的值集合,其中该集合中的每个值识别第二特征向量在第二特征空间的维度上的对应位置。因此,由经训练的第二编码器生成的第二特征向量可以编码或以其他方式表示与在对应的非结构化数据文件中识别出的相同事件相关联的一个或多个特性。

[0091] 1.训练数据

[0092] 在一些实施例中,第二编码器的训练数据包括情境数据对象的第二训练集。情境数据包括描述与第一训练集的对应非结构化数据文件识别出的特定事件相关联的环境的任何信息。训练情境数据对象可以包括非结构化数据,包括非结构化文本。在一些情况下,第二训练集的情境数据对象包括根据对应数据库模式或文档对象模型格式化的结构化数据。第二训练集的情境数据对象提供可以代替训练标签或注释使用的附加数据点。第二训练集的情境数据对象可以用于训练第二编码器以在第二特征空间中生成特征向量。异常事件检测系统可以使用第二特征空间中的特征向量来预测异常事件,而无需为训练数据生成目标标签。

[0093] 第二训练集的情境数据对象可以从一个或多个数据存储库(例如,用户简档数据存储库)获得。第二训练集的情境数据对象可以包括但不限于异常事件检测系统可以访问的以下信息:(i)特定事件发生的位置;(ii)特定事件中涉及的实体;(iii)被预测与特定事件相关的另一个先前或后续事件;和/或(iv)与和特定事件相关联的交易相关联的类型或类别。

[0094] 2.机器学习模型选择

[0095] 然后可以选择第二编码器的一个或多个机器学习模型进行训练。与第一编码器类似,第二编码器的机器学习模型可以是卷积神经网络、初始神经网络、残差神经网络、循环神经网络、长短期记忆模型、门控循环单元模型或深度神经网络的其他变体。机器学习模型也可以是可以被训练以从非结构化数据文件生成特征向量的任何其他合适的模型,诸如朴素贝叶斯分类器、线性分类器、支持向量机、诸如随机森林模型的装袋模型、增强模型、浅层神经网络、或一种或多种此类技术的组合——例如,多尺度卷积神经网络。在一些情况下,第二编码器具有暹罗网络体系架构,其中机器学习模型在每次迭代中被训练两次,以惩罚相似训练特征向量的位置在第二特征空间中相距远的情况以及不同训练特征向量的位置在第二特征空间中彼此靠近的情况。

[0096] 在一些情况下,第二编码器具有与第一编码器的体系架构相同或不同的体系架构。因此,第一编码器和第二编码器中的每一个都可以被训练以在其相应的特征空间中生成特征向量,其中第一编码器和第二编码器可以不共享任何参数。

[0097] 第二编码器可以是卷积神经网络或多层感知器神经网络。在一些情况下,第二编

码器是循环神经网络。图4描绘了根据一些实施例的循环神经网络的示例示意图400。循环神经网络包括神经网络的重复模块(“单元格(cell)”)链。具体而言,循环神经网络的操作包括重复由情境数据对象的文本部分内的文本部分(t)的位置进行索引的单个单元格。为了提供其循环行为,循环神经网络维持隐藏状态s_t,该隐藏状态s_t被提供作为网络下一次迭代的输入。如图4的左侧部分中所示,循环神经网络接收文本部分x_t的特征表示和使用先前文本部分的输入特征集合确定的隐藏状态值s_{t-1}。以下等式提供了如何确定隐藏状态s_t:

[0098] $$s_t=\varphi\ (Ux_t+Ws_{t-1}),$$

[0099] 其中U和W是分别应用于x_t和s_{t-1}的权重值,并且φ是非线性函数,诸如tanh或ReLU。

[0100] 循环神经网络的输出被表达为:

[0101] $o_t=\mathrm{softmax}(Vs_t)$,

[0102] 其中V是应用于隐藏状态值s_t的权重值。

[0103] 因此,隐藏状态s_t可以被称为网络的记忆。隐藏状态s_t可以取决于与使用或以其他方式从一个或多个先前文本部分中导出的输入和/或输出相关联的信息。步骤o_t处的输出是用于识别文本部分的部分识别符(section identifier)的值集合,其至少部分地基于文本部分位置t处的记忆进行计算。

[0104] 3.机器学习模型的训练

[0105] 在选择机器学习模型(例如,卷积神经网络、循环神经网络)之后,可以使用训练数据训练第二编码器。与第一编码器相似,可以使用对比学习训练第二编码器,以生成表示情境数据的第二特征向量。对比学习允许使用既未标记也未注释的第二训练集训练第二编码器。

[0106] 图5图示了根据一些实施例的使用对比学习训练第二编码器的示意图500。如图5中所示,第二编码器的训练至少包括以下阶段:(1)增强训练情境数据对象;(2)在第二特征空间中生成第一训练特征向量;(3)在第二特征空间中生成第二训练特征向量;(4)确定第一训练特征向量和第二训练特征向量之间的损失;以及(5)基于确定的损失调整第二编码器的一个或多个参数。

[0107] 可以通过将数据增强502应用于训练数据集的训练情境数据对象504以生成第一增强训练情境数据对象506和第二增强训练非结构化情境数据对象508来发起第二编码器的训练。数据增强502可以包括变换训练情境数据对象504以生成训练情境数据对象的两个或更多个修改版本。数据增强确保两个或更多个修改版本与训练情境数据对象504“相似”。修改版本可以用于执行第二编码器510的对比学习。例如,情境数据对象的数据增强可以包括文本数据的反向翻译,其中情境数据对象可以从原始源语言(例如,英语)翻译成目标语言(例如,德语),然后再翻译回原始源语言。情境数据对象的文本增强的其他示例可以包括训练情境数据对象504中的一个或多个单词的随机交换,以及情境数据对象504中的一个或多个单词的随机同义词替换。

[0108] 第二编码器510可以将第一增强训练情境数据对象506处理为第二特征空间516中的第一训练特征向量512。第一训练特征向量512可以是n维值向量,这些值共同表示第一增强训练情境数据对象506。第二特征空间516可以是n维空间,其中可以通过其对应位置来识别训练情境数据对象的第一训练特征向量512。在一些情况下,将第二编码器的残差神经网络应用于第一增强训练情境数据对象506以生成第一训练特征向量512。

[0109] 第二编码器510还可以将第二增强训练情境数据对象508处理为第二特征空间516中的第二训练特征向量514。与第一训练特征向量512相似,第二训练特征向量514可以是n维值向量,这些值共同表示第二增强训练情境数据对象508。在一些情况下,第二编码器510的残差神经网络还应用于第二增强训练情境数据对象508以生成第二训练特征向量514。第二编码器510的训练可以被执行为使得第一训练特征向量512和第二训练特征向量514均可以对训练情境数据对象504的一个或多个特性进行编码。

[0110] 可以确定第一训练特征向量512和第二训练特征向量514之间的损失518。所确定的损失518可以识别第二编码器510在训练阶段期间的性能水平。例如,如果第二编码器510预测第一训练特征向量512和第二训练特征向量514彼此相似,那么所确定的损失518可以减小。相反,如果第二编码器510预测第一训练特征向量512和第二训练特征向量514彼此不同,那么所确定的损失518可以增加。损失518可以通过以下方式确定:(1)确定第一训练特征向量512和第二训练特征向量514之间的损失(例如,余弦相似度、距离);以及(2)将第二对比损失函数应用于所确定的损失。在一些情况下,第二对比损失函数是交叉熵损失函数,其被定义为惩罚第一训练特征向量512和第二训练特征向量514的位置在第二特征空间中相距远的情况。

[0111] 第二编码器可以基于确定的损失调整其参数520中的一个或多个。可以执行调整第二编码器510的参数520以最小化在第二特征空间516内确定的第一训练特征向量512和第二训练特征向量514之间的损失。通过识别没有标签或注释的“相似”情境数据对象,第二编码器510的训练可以促进识别情境数据的特性。

[0112] C.第一编码器和第二编码器的附加训练

[0113] 除了单独训练第一编码器和第二编码器之外,训练阶段还可以包括在共享特征空间中一起训练第一编码器和第二编码器。附加训练可以确保第一特征向量和第二特征向量在其相应的特征空间中表示相同的特定事件。因此,可以基于在共享特征空间中由第一编码器生成的第一训练特征向量与由第二编码器生成的第二训练特征向量之间的确定的损失来执行第一编码器和第二编码器的训练。在一些情况下,训练包括基于以下各项在共享特征空间中生成向量:(i)通过将第一投影函数应用于第一特征向量108而生成的第一训练变换特征向量;以及(ii)通过将第二投影函数应用于第二特征向量110而生成的第二训练变换特征向量。可以通过组合来自第一训练变换特征向量和第二训练变换特征向量的信息来生成向量,从而产生与单独的第一训练特征向量或第二训练特征向量相比具有改进的投影预测性的该向量。

[0114] 图6图示了根据一些实施例的使用对比学习附加地训练第一编码器和第二编码器的示意图600。如图6中所示,第一编码器和第二编码器的训练可以至少包括以下阶段:(1)选择匹配的训练非结构化数据文件和训练情境数据对象;(2)在共享特征空间中生成训练非结构化数据文件的第一变换训练特征向量;(3)在共享特征空间中生成训练情境数据对象的第二变换训练特征向量;以及(4)确定第一变换训练特征向量和第二变换训练特征向量之间的损失;以及(5)基于确定的损失调整第一编码器的一个或多个参数和第二编码器的一个或多个参数。

[0115] 可以通过选择匹配的训练非结构化数据文件602和训练情境数据对象604来发起附加训练。选择可以包括确定匹配的训练非结构化数据文件602和训练情境数据对象604之

间存在重叠数据。例如,训练情境数据对象604中识别出的费用报告数据可以包括购买描述、实体信息、购买日期和时间,其中许多与训练非结构化数据文件602中识别出的对应信息匹配。在一些情况下,可以通过使用光学字符辨别从训练非结构化数据文件602中提取文本数据来识别匹配信息。

[0116] 第一编码器606可以将训练非结构化数据文件602处理为第一特征空间中的第一训练特征向量610。第二编码器608可以将训练情境数据对象604处理为第二特征空间中的第二训练特征向量612。为了训练第一编码器606辨别训练非结构化数据文件602和训练情境数据对象604对应于相同的特定事件,可以应用第一投影函数614将第一特征向量610变换成共享特征空间622中的第一变换训练特征向量618。类似地,为了训练第二编码器608辨别训练非结构化数据文件602和训练情境数据对象604对应于相同的特定事件,可以应用第二投影函数616将第二特征向量612变换成共享特征空间622中的第二变换训练特征向量620。共享特征空间622可以是n维空间,其中第一变换训练特征向量618和第二变换训练特征向量620均可以通过其相应的位置来识别。第一投影函数614和第二投影616可以变换训练特征向量,使得可以在对比学习期间在相同的共享特征空间622内直接将变换特征向量进行比较。

[0117] 第一变换训练特征向量618与第二变换训练特征向量620之间的损失624。确定的损失624可以识别第一编码器606和第二编码器608两者在预测训练非结构化数据文件和训练情境数据对象是否识别相同的特定事件方面的性能水平。例如,如果预测第一变换训练特征向量618和第二变换训练特征向量620被生成为彼此相似,那么确定的损失624可以减少。相反,如果预测第一变换训练特征向量618和第二变换训练特征向量620彼此不同,那么确定的损失618可以增加。损失624可以通过以下方式确定:(1)确定第一变换训练特征向量618与第二变换训练特征向量620之间的损失(例如,余弦相似度、距离);以及(2)将组合对比损失函数应用于所确定的损失。在一些情况下,组合对比损失函数是交叉熵损失函数,其被定义为惩罚第一变换训练特征向量618和第二变换训练特征向量620的位置在共享特征空间622中相距远的情况。

[0118] 在一些情况下,第一对比损失函数和第二对比损失函数被组合并通过一个或多个加权因子加权,使得可以在附加训练期间使用组合损失函数。加权因子可以是可以在第一编码器606和第二编码器608的附加训练期间进行微调的超参数。因此,第一编码器606和第二编码器608的对比学习可以基于在共享特征空间622内确定的损失确保非结构化数据文件和对应的情境数据识别相同的特定事件。

[0119] 例如,相对于应用于第二变换训练特征向量620的第二加权因子,可以增加应用于第一变换训练特征向量618的第一加权因子。可以响应于对非结构化数据文件更能预测事件的确定来执行第一加权因子的增加。相反,可以相对于第一加权因子增加第二加权因子。可以响应于对情境数据更能预测事件的另一个确定来执行第二加权因子的增加。可以在训练第一编码器和第二编码器之前或期间执行加权因子的调整。因此,组合损失函数可以确定加权损失,该加权损失可以用于调整第一编码器和第二编码器的参数,使得第一编码器和第二编码器可以提高其检测异常事件的预测能力。

[0120] 在另一个示例中,第一加权因子和第二加权因子可以被配置为相同的值。第一加权因子和第二加权因子的配置允许第一编码器和第二编码器的训练在无监督的情况下进

行训练。此外,可以执行配置直到检测异常事件的预测能力能够达到预期结果(例如,损失最小化到预定阈值以下)。

[0121] 附加地或替代地,一个或多个加权因子是可以基于由组合损失函数确定的损失进行调整的参数。实际上,可以结合第一编码器和第二编码器的参数来学习一个或多个加权因子。在训练期间对加权因子的学习可以在没有监督的情况下执行。此外,对加权因子的学习可以促进基于最初可能已经被视为无关紧要的信息来改进对异常事件的预测。

[0122] 然后,第一编码器606可以基于确定的损失624调整其参数626中的一个或多个。第二编码器608可以基于确定的损失624调整其参数628中的一个或多个。可以执行第一编码器606的参数626的调整和第二编码器608的参数628的调整两者,以在共享特征空间622内最小化第一变换训练特征向量618和第二变换训练特征向量620之间确定的损失。实际上,通过学习"相似"事件,第一编码器606和第二编码器608的附加训练可以促进在没有标签或注释的情况下识别事件的特性。

[0123] D.预测异常事件的机器学习技术

[0124] 由第一编码器生成的第一特征向量和由第二编码器生成的第二特征向量可以被另一个机器学习模型使用,以预测非结构化数据文件中识别出的特定事件是否对应于异常事件。在一些情况下,分类引擎(例如,图1的分类引擎124)被训练以通过以下方式生成分类结果:(1)处理与特定事件相关联的训练情境数据对象和/或训练非结构化数据文件以生成中间输出;(2)确定中间输出与包括训练第一特征向量和训练第二特征向量的事件数据输入元素之间的损失;以及(3)基于确定的损失调整分类引擎的机器学习模型的一个或多个参数。事件数据输入元素可以被配置为情境数据对象的目标标签。在一些情况下,在分类引擎的训练期间,第一编码器和第二编码器的参数保持不修改。在一些情况下,分类引擎进一步处理反馈数据,以增强机器学习模型预测分类结果的准确性。在一些情况下,反馈数据用于调整事件数据输入元素的第一特征向量和/或第二特征向量的值。

[0125] 1.训练数据

[0126] 在一些实施例中,用于分类引擎的训练数据包括带标签的情境数据对象的第三训练集。训练情境数据对象包括描述与由第一训练集的对应训练非结构化数据文件识别出的特定事件相关联的环境的任何信息。训练情境数据对象可以包括非结构化数据,包括非结构化文本。在一些情况下,训练情境数据对象包括根据对应数据库模式或文档对象模型进行格式化的结构化数据。附加地或替代地,每个带标签的情境数据对象也可以与对应的训练非结构化数据文件相关联。

[0127] 在一些情况下,第三训练集的每个训练情境数据对象都标记有对应的事件数据输入元素。事件数据输入元素可以是包括特征向量集合的数据结构(例如,元组),该特征向量集合可以用于预测特定事件是否对应于异常事件。该特征向量集合可以包括由第一编码器通过处理训练非结构化数据文件生成的第一特征向量和由第二编码器通过处理训练情境数据对象生成的第二特征向量。事件数据输入元素可以用作情境数据对象的目标标签,使得可以训练分类引擎来预测异常事件,而无需手动标记或注释每个训练情境数据对象。

[0128] 附加地或替代地,第三训练集的每个训练情境数据对象可以与训练情境数据对象的反馈数据相关联。反馈数据可以包括用户手动输入的用于调整事件数据输入元素的第一特征向量和/或第二特征向量的值的数据。反馈数据可以被配置为增强或更新对应训练情

境数据对象的训练标签。训练情境数据对象可以响应于对事件数据输入元素的第一训练特征向量和第二训练特征向量中的一个偏离预期向量值的另一个预测而与反馈数据相关联。

[0129] 2.机器学习模型选择

[0130] 然后可以选择分类引擎的一个或多个机器学习模型进行训练。第一编码器的机器学习模型可以是人工神经网络、卷积神经网络、初始神经网络、残差神经网络、循环神经网络、长短期记忆模型、门控循环单元模型或深度神经网络的其他变体。机器学习模型还可以是可以被训练以从非结构化数据文件生成特征向量的任何其他合适的模型，诸如朴素贝叶斯分类器、线性分类器、支持向量机、诸如随机森林模型的装袋模型、增强模型、浅层神经网络、或一种或多种此类技术的组合——例如，多尺度卷积神经网络。在一些情况下，分类引擎具有多类分类器模型，其中分类结果对应于与最高预测分数相关联的类。

[0131] 3.机器学习模型的训练

[0132] 图7图示了根据一些实施例的用于训练分类引擎以预测异常事件的示意图700。如图7中所示，分类引擎的机器学习模型的训练可以包括使用第三训练集的训练情境数据对象和对应的事件数据输入元素作为输入。可以通过处理训练情境数据对象和对应的事件数据输入元素从机器学习模型生成中间输出。中间输出可以预测由训练情境数据对象识别出的事件对应于特定类型事件(例如，正常事件、异常事件)的概率。可以通过将损失函数(例如，交叉熵损失函数)应用于中间输出和被配置为训练情境数据对象的目标标签的事件数据输入元素来确定损失。确定的损失可以用于调整分类引擎的机器学习模型的参数，使得可以通过使用训练数据将损失最小化。

[0133] 可以通过访问训练情境数据对象702和训练事件数据输入元素706来发起分类引擎的训练。训练事件数据输入元素706包括第一训练特征向量708和第二训练特征向量710。可以通过使用第一编码器(例如，图4的第一编码器410)处理对应的训练非结构化数据文件来生成第一训练特征向量708。此外，可以通过使用第二编码器(例如，图6的第二编码器610)处理情境数据对象702来生成第二训练特征向量710。事件数据输入元素可以用作情境数据对象的目标标签，使得可以训练分类引擎来预测异常事件，而无需手动标记或注释每个训练情境数据对象。

[0134] 在一些情况下，训练情境数据对象702可以与反馈数据704相关联。可以响应于对事件数据输入元素706的第一训练特征向量708和第二训练特征向量710中的一个偏离预期向量值的另一个预测来执行反馈数据的关联。例如，可以通过以下方式预测与第一特征向量708的第一预期向量值的第一偏差：(1)生成训练第一特征空间的第一特征向量的分布；(2)确定第一特征向量708对应于所生成分布的离群值；以及(3)确定第一特征向量708偏离预期向量值。在一些情况下，通过确定所生成分布的四分位数范围值并确定第一特征向量低于或超过四分位数范围值来确定离群值。还可以通过以下方式确定离群值：(1)确定第一特征向量的z分数；以及(2)确定z分数低于第一预定阈值(例如，-3)或超过第二预定阈值(例如，3)。

[0135] 分类引擎712可以处理训练情境数据对象702以生成中间结果714。中间结果714可以包括表示训练情境数据对象702的一个或多个值。在一些情况下，分类引擎可以包括生成中间结果714的人工神经网络。

[0136] 可以确定中间结果714和训练事件数据输入元素706之间的损失716。确定的损失

716可以识别分类引擎712关于它是否可以预测特定事件是否对应于异常事件的性能水平。例如,如果分类引擎712预测中间结果714和训练事件数据输入元素706彼此相似,那么确定的损失716可以减少。相反,如果分类引擎712预测中间结果714和训练事件数据输入元素706彼此不同,那么确定的损失716可以增加。

[0137] 分类引擎可以基于确定的损失调整其参数718中的一个或多个。可以执行调整分类引擎712的参数718以最小化中间结果714和训练事件数据输入元素706之间确定的损失716。通过使用事件数据输入元素识别“相似”情境数据对象,如果给定情境数据和对应事件数据输入元素的确定损失超过预定阈值,那么分类引擎712的训练可以用于预测特定事件对应于异常事件。

[0138] IV.使用对比学习预测异常事件

[0139] 图8包括图示根据一些实施例的使用对比学习预测异常事件的方法的示例的流程图800。流程图800中描述的一些操作可以由计算机系统执行。虽然流程图800可以将操作描述为顺序处理,但在各种实施例中,许多操作可以并行或并发执行。此外,可以重新布置操作的次序。操作可以具有图中未示出的附加步骤。此外,方法的一些实施例可以通过硬件、软件、固件、中间件、微代码、硬件描述语言或它们的任何组合来实现。当用软件、固件、中间件或微代码实现时,用于执行相关联任务的程序代码或代码段可以存储在计算机可读介质(诸如存储介质)中。

[0140] 在步骤802处,异常事件预测系统(例如,异常事件预测系统100)可以访问非结构化数据文件。非结构化数据文件识别特定事件的发生。非结构化数据文件102可以对应于特定文件格式,包括电子邮件文件格式(例如,MSG)、图像文件格式(例如,JPG、JPEG、PNG、BMP)、网页格式(例如,HTML)、文档文件格式(例如,DOC、PDF、XPS、EPUB)、音频文件格式(例如,MP3、M4A)、视频文件格式(例如,MPEG、MOV)和其他类型的文件格式。

[0141] 在步骤804处,异常事件预测系统可以访问与非结构化数据文件相关联的情境数据。情境数据包括描述特定事件发生的环境的信息。例如,情境数据可以识别特定事件发生的位置或被预测与特定事件相关的另一个先前或后续事件。在一些情况下,异常事件预测系统将一个或多个预处理操作(例如,规范化、过滤)应用于情境数据,以增强处理情境数据的编码器的性能。例如,可以通过将从一个或多个数据存储库访问的数据记录(例如,重叠数据记录)合并到单个或减少数量的表中并移除任何冗余数据来将平坦化操作应用于情境数据。

[0142] 在步骤806处,异常事件预测系统可以通过使用第一编码器处理非结构化数据文件来生成特征向量集合中的第一特征向量。第一特征向量表示第一特征空间中的非结构化数据文件。在一些情况下,第一编码器包括残差神经网络或卷积神经网络。可以训练第一编码器以生成非结构化数据文件的第一特征向量,其中定义损失以惩罚相似训练元素(例如,训练非结构化数据文件)的特征向量的位置在第一特征空间中相距远的情况和/或不同训练元素的特征向量的位置在第一特征空间中彼此靠近的情况。

[0143] 在步骤808处,异常事件预测系统可以通过使用第二编码器处理情境数据来生成特征向量集合中的第二特征向量。第二特征向量表示第二特征空间中的情境数据。可以训练第二编码器以生成情境数据的第二特征向量,其中定义另一个损失以便惩罚相似训练元素(例如,训练情境数据)的特征向量的位置在第二特征空间中相距远的情况和/或不同训

练元素的特征向量的位置在第二特征空间中彼此靠近的情况。

[0144] 在一些情况下,第二编码器包括另一个卷积神经网络。第二编码器可以具有与第一编码器的体系架构相同或不同的体系架构。因此,第一编码器和第二编码器中的每一个都可以被训练以在其相应的特征空间中生成特征向量,其中第一编码器和第二编码器可以不共享任何参数。

[0145] 在一些情况下,通过以下方式学习第一编码器的一个或多个参数和第二编码器的一个或多个参数:(i)变换第一特征空间中的第一训练特征向量以在共享特征空间中生成表示训练非结构化数据文件的第一训练变换特征向量,其中训练非结构化数据文件识别训练事件;(ii)变换第二特征空间中的第二训练特征向量以在共享特征空间中生成表示训练情境数据的第二训练变换特征向量,其中训练情境数据识别相同的训练事件;以及(iii)基于第一训练变换特征向量和第二训练变换特征向量在共享特征空间中生成第三训练向量,其中生成包括组合来自第一训练变换特征向量和第二训练变换特征向量的信息以产生与单独的第一训练特征向量或第二训练特征向量相比具有改进的投影预测性的第三训练向量。

[0146] 在一些情况下,第一编码器和第二编码器的训练包括使用组合损失函数来确定第一变换训练特征向量和第二变换训练特征向量之间的损失,并基于确定的损失调整第一编码器的一个或多个参数和第二编码器的一个或多个参数。组合损失函数可以包括第一对比损失函数和第二对比损失函数。组合损失函数可以由一个或多个加权因子加权,该加权因子可以基于训练非结构化数据文件和/或情境数据的相对预测能力进行调整。

[0147] 在步骤810处,异常事件预测系统可以为非结构化数据文件生成事件数据输入元素。事件数据输入元素可以包括特征向量集合。特征向量集合可以包括第一特征向量和第二特征向量,它们由机器学习模型(例如,分类器模型)处理以预测非结构化数据文件中识别出的特定事件是否对应于异常事件。

[0148] 在步骤812处,异常事件预测系统可以通过使用机器学习模型处理事件数据输入元素来生成非结构化数据文件的分类结果。在一些情况下,分类结果包括对特定事件是否对应于异常事件的预测。分类结果还可以包括具有置信区间的异常事件的严重性分数。然后可以使用严重性分数对预测的异常事件进行分类或聚类。例如,可以对异常事件进行聚类以基于预测的严重性分数对异常事件进行排名。下游管道任务可以使用此信息进行后续动作。可以训练机器学习模型来处理情境数据和/或非结构化数据文件以生成中间输出。可以将中间输出与用作目标标签的事件数据输入元素进行比较。基于该比较,机器学习模型可以生成对特定事件是否对应于异常事件的预测。机器学习模型可以包括人工神经网络或基于树的分类器。

[0149] 在一些情况下,异常事件预测系统确定在第二特征空间中,第二特征向量与另一个特征向量之间的另一个损失超过预定阈值。响应于确定另一个损失超过预定阈值,异常事件预测系统可以接收与情境数据相关联的反馈数据。可以使用机器学习模型处理反馈数据以生成分类结果。

[0150] 在步骤814处,异常事件预测系统可以输出分类结果。例如,分类结果可以在本地呈现或传输到另一个设备。结果可以与非结构化数据文件的识别符一起输出。处理800此后终止。

[0151] V.说明性系统

[0152] 图9描绘了用于实现实施例中的一个实施例的分布式系统900的简化图。在所示的实施例中,分布式系统900包括一个或多个客户端计算设备902、904、906和908,该一个或多个客户端计算设备被配置为通过一个或多个网络910执行和操作客户端应用,诸如web浏览器、专有客户端(例如,Oracle Forms)等。服务器912可以经由网络910与远程客户端计算设备902、904、906和908通信地耦合。

[0153] 在各种实施例中,服务器912可以适于运行由系统的一个或多个组件提供的一个或多个服务或软件应用。在一些实施例中,这些服务可以作为基于web的服务或云服务或者在软件即服务(SaaS)模型下被供应给客户端计算设备902、904、906和/或908的用户。操作客户端计算设备902、904、906和/或908的用户进而可以利用一个或多个客户端应用来与服务器912进行交互以利用由这些组件提供的服务。

[0154] 在图中描绘的配置中,系统900的软件组件918、920和922被示出为在服务器912上实现。在其他实施例中,系统900的一个或多个组件和/或由这些组件提供的服务也可以由客户端计算设备902、904、906和/或908中的一个或多个来实现。然后,操作客户端计算设备的用户可以利用一个或多个客户端应用来使用由这些组件提供的服务。这些组件可以用硬件、固件、软件或其组合来实现。应该认识到的是,各种不同的系统配置是可能的,这些配置可能与分布式系统900不同。图中所示的实施例因此是用于实现实施例系统的分布式系统的一个示例,而不旨在是限制性的。

[0155] 客户端计算设备902、904、906和/或908可以是便携式手持设备(例如,**iPhone®**(蜂窝电话)、**iPad®**(计算平板电脑)、个人数字助理(PDA))或可穿戴设备(例如,Google **Glass®**头戴式显示器),运行诸如Microsoft Windows **Mobile®**和/或各种移动操作系统(诸如iOS、Windows Phone、Android、BlackBerry 10、Palm OS等)的软件,并且启用互联网、电子邮件、短消息服务(SMS)、**Blackberry®**或其他通信协议。客户端计算设备可以是通用个人计算机,作为示例,包括运行各种版本的Microsoft **Windows®**、Apple **Macintosh®**和/或Linux操作系统的个人计算机和/或膝上型计算机。客户端计算设备可以是运行各种可商业获得的**UNIX®**或类UNIX操作系统(包括但不限于各种GNU/Linux操作系统,诸如例如Google Chrome OS)中的任何操作系统的工作站计算机。替代地或附加地,客户端计算设备902、904、906和908可以是能够通过(一个或多个)网络910进行通信的任何其他电子设备,诸如瘦客户端计算机、启用互联网的游戏系统(例如,具有或不具有**Kinect®**手势输入设备的Microsoft Xbox游戏控制台)和/或个人消息传递设备。

[0156] 虽然示出了具有四个客户端计算设备的示例性分布式系统900,但是可以支持任何数量的客户端计算设备。其他设备(诸如具有传感器等的设备等)可以与服务器912进行交互。

[0157] 分布式系统900中的(一个或多个)网络910可以是本领域技术人员熟悉的、可以使用各种可商业获得的协议中的任何协议来支持数据通信的任何类型的网络,其中协议包括但不限于TCP/IP(传输控制协议/互联网协议)、SNA(系统网络体系架构)、IPX(互联网分组交换)、AppleTalk等。仅仅作为示例,(一个或多个)网络910可以是局域网(LAN),诸如基于以太网、令牌环等的LAN。(一个或多个)网络910可以是广域网和互联网。它可以包括虚拟网络,包括但不限于虚拟专用网络(VPN)、内联网、外联网、公共交换电话网(PSTN)、红外网络、

无线网络(例如,依据电气和电子协会(IEEE)802.11协议套件、蓝牙(Bluetooth®)和/或任何其他无线协议中的任何协议而操作的网络);和/或这些网络和/或其他网络的任何组合。

[0158] 服务器912可以由以下组成:一个或多个通用计算机、专用服务器计算机(作为示例,包括PC(个人计算机)服务器、UNIX®服务器、中档服务器、大型计算机、机架安装的服务器等)、服务器场、服务器集群或任何其他适当的布置和/或组合。在各种实施例中,服务器912可以适于运行在前述公开中描述的一个或多个服务或软件应用。例如,服务器912可以与用于执行以上根据本公开的实施例描述的处理的服务器对应。

[0159] 服务器912可以运行包括以上讨论的任何操作系统的操作系统,以及任何可商业获得的服务器操作系统。服务器912还可以运行各种附加的服务器应用和/或中间层应用中的任何应用,包括HTTP(超文本传输协议)服务器、FTP(文件传输协议)服务器、CGI(公共网关接口)服务器、JAVA®服务器、数据库服务器等。示例性数据库服务器包括但不限于可从Oracle、Microsoft、Sybase、IBM(国际商业机器)等商业获得的那些数据库服务器。

[0160] 在一些实施方式中,服务器912可以包括一个或多个应用,以分析和整合从客户端计算设备902、904、906和908的用户接收到的数据馈送和/或事件更新。作为示例,数据馈送和/或事件更新可以包括但不限于:Twitter®馈送、Facebook®更新或者从一个或多个第三方信息源和连续数据流接收到的实时更新,实时更新可以包括与传感器数据应用、金融价格收报机(financial ticker)、网络性能测量工具(例如,网络监视和流量管理应用)、点击流分析工具、汽车交通监视等相关的实时事件。服务器912还可以包括一个或多个应用,以经由客户端计算设备902、904、906和908的一个或多个显示设备来显示数据馈送和/或实时事件。

[0161] 分布式系统900还可以包括一个或多个数据库914和916。数据库914和916可以驻留在各种位置。作为示例,数据库914和916中的一个或多个可以驻留在服务器912本地(和/或驻留在服务器912中)的非暂态存储介质上。可替代地,数据库914和916可以远离服务器912并且经由基于网络的连接或专用的连接与服务器912进行通信。在一组示例中,数据库914和916可以驻留在存储区域网络(SAN)中。类似地,用于执行服务器912所具有的功能的任何必要文件都可以适当地本地存储在服务器912上和/或远程存储。在一组实施例中,数据库914和916可以包括适于响应SQL格式的命令而存储、更新和检索数据的关系数据库,诸如由Oracle提供的数据库。

[0162] 图10是根据本公开的一些实施例的可以将由实施例系统的一个或多个组件提供的服务作为云服务供应的系统环境1000的一个或多个组件的简化框图。在所示的实施例中,系统环境1000包括可以由用户使用以与提供云服务的云基础设施系统1002进行交互的一个或多个客户端计算设备1004、1006和1008。客户端计算设备可以被配置为操作客户端应用,诸如web浏览器、专有客户端应用(例如,Oracle Forms)或某种其他应用,这些应用可以由客户端计算设备的用户用来与云基础设施系统1002进行交互以使用由云基础设施系统1002提供的服务。

[0163] 应当认识到的是,图中描绘的云基础设施系统1002可以具有与所描绘的组件不同的其他组件。另外,图中所示的实施例仅仅是可以结合本发明的实施例的云基础设施系统的一个示例。在一些其他实施例中,云基础设施系统1002可以具有比图中所示更多或更少

的组件,可以组合两个或更多个组件,或者可以具有组件的不同配置或布置。

[0164] 客户端计算设备1004、1006和1008可以是与上面针对902、904、906和908所描述的设备类似的设备。

[0165] 虽然示例性系统环境1000被示出具有三个客户端计算设备,但是可以支持任何数量的客户端计算设备。其他设备(诸如具有传感器的设备等)可以与云基础设施系统1002进行交互。

[0166] (一个或多个)网络1010可以促进客户端1004、1006和1008与云基础设施系统1002之间的数据的通信和交换。每个网络可以是本领域技术人员所熟悉的可以使用各种商业上可获得的协议(包括上面针对(一个或多个)网络910所描述的那些协议)中的任何协议支持数据通信的任何类型的网络。

[0167] 云基础设施系统1002可以包括一个或多个计算机和/或服务器,其可以包括上面针对服务器912所描述的那些计算机和/或服务器。

[0168] 在某些实施例中,由云基础设施系统提供的服务可以包括按需对云基础设施系统的用户可用的许多服务,诸如在线数据存储和备份解决方案、基于Web的电子邮件服务、被托管的办公室(office)套件和文档协作服务、数据库处理、受管理的技术支持服务等。由云基础设施系统提供的服务可以动态扩展以满足云基础设施系统的用户的需要。由云基础设施系统提供的服务的具体实例化在本文中被称为“服务实例”。一般而言,从云服务提供商的系统经由通信网络(诸如互联网)对用户可用的任何服务被称为“云服务”。通常,在公共云环境中,构成云服务提供商的系统的服务器和系统与客户自己的本地服务器和系统不同。例如,云服务提供商的系统可以托管应用,并且用户可以经由诸如互联网的通信网络按需订购和使用应用。

[0169] 在一些示例中,计算机网络云基础设施中的服务可以包括对存储装置、被托管的数据库、被托管的web服务器、软件应用或由云供应商向用户提供的其他服务的受保护的计算机网络访问,或者如本领域中另外已知的受保护的计算机网络访问。例如,服务可以包括通过互联网对云上的远程存储装置的密码保护的访问。作为另一个示例,服务可以包括基于web服务的被托管的关系数据库和脚本语言中间件引擎,以供联网的开发人员私有使用。作为另一个示例,服务可以包括对在云供应商的网站上托管的电子邮件软件应用的访问。

[0170] 在某些实施例中,云基础设施系统1002可以包括以自助服务、基于订阅、弹性可扩展、可靠、高度可用和安全的方式递送给客户的应用、中间件和数据库服务供应的套件。这种云基础设施系统的示例是由本受让方提供的Oracle公共云。

[0171] 在各种实施例中,云基础设施系统1002可以适于自动供给、管理和跟踪客户对云基础设施系统1002供应的服务的订阅。云基础设施系统1002可以经由不同的部署模型来提供云服务。例如,可以依据公共云模型提供服务,其中云基础设施系统1002被销售云服务的组织拥有(例如,被Oracle拥有),并且服务对一般公众或不同行业的企业可用。作为另一个示例,可以依据私有云模型来提供服务,其中云基础设施系统1002仅针对单个组织操作,并且可以为该组织内的一个或多个实体提供服务。还可以依据社区云模型来提供云服务,其中云基础设施系统1002和由云基础设施系统1002提供的服务由相关社区中的若干组织共享。还可以依据混合云模型提供云服务,该混合云模型是两个或更多个不同模型的组合。

[0172] 在一些实施例中,由云基础设施系统1002提供的服务可以包括在软件即服务

(SaaS)类别、平台即服务(PaaS)类别、基础设施即服务(IaaS)类别或包括混合服务的其他服务类别下提供的一个或多个服务。客户经由订阅订单可以订购由云基础设施系统1002提供的一个或多个服务。云基础设施系统1002然后执行处理以提供客户的订阅订单中的服务。

[0173] 在一些实施例中,由云基础设施系统1002提供的服务可以包括但不限于应用服务、平台服务和基础设施服务。在一些示例中,应用服务可以由云基础设施系统经由SaaS平台提供。SaaS平台可以被配置为提供落入SaaS类别的云服务。例如,SaaS平台可以提供在集成开发和部署平台上构建和递送按需应用套件的能力。SaaS平台可以管理和控制用于提供SaaS服务的底层软件和基础设施。通过利用由SaaS平台提供的服务,客户可以利用在云基础设施系统上执行的应用。客户可以获取应用服务,而无需客户购买单独许可证和支持。可以提供各种不同的SaaS服务。示例包括但不限于为大型组织提供销售绩效管理、企业集成和灵活性的解决方案的服务。

[0174] 在一些实施例中,平台服务可以由云基础设施系统经由PaaS平台提供。PaaS平台可以被配置为提供落入PaaS类别的云服务。平台服务的示例可以包括但不限于使组织(诸如Oracle)能够在共享的公共体系架构上整合现有应用以及充分利用平台提供的共享服务来构建新应用的能力的服务。PaaS平台可以管理和控制用于提供PaaS服务的底层软件和基础设施。客户可以获取由云基础架构系统提供的PaaS服务,而无需客户购买单独许可证和支持。平台服务的示例包括但不限于Oracle Java云服务(JCS)、Oracle数据库云服务(DBCS)等。

[0175] 通过利用由PaaS平台提供的服务,客户可以采用由云基础设施系统支持的编程语言和工具,并且还控制所部署的服务。在一些实施例中,由云基础设施系统提供的平台服务可以包括数据库云服务、中间件云服务(例如,Oracle融合中间件服务)和Java云服务。在一个实施例中,数据库云服务可以支持共享服务部署模型,该模型使得组织能够池化数据库资源并且以数据库云的形式向客户供应数据库即服务。中间件云服务可以为客户提供开发和部署各种云应用的平台,并且Java云服务可以为客户提供在云基础设施系统中部署Java应用的平台。

[0176] 各种不同的基础设施服务可以由云基础设施系统中的IaaS平台提供。基础设施服务促进底层计算资源(诸如存储装置、网络和其他基础计算资源)的管理和控制,以供客户利用由SaaS平台和PaaS平台提供的服务。

[0177] 在某些实施例中,云基础设施系统1002还可以包括基础设施资源1030,用于向云基础设施系统的客户提供用于提供各种服务的资源。在一个实施例中,基础设施资源1030可以包括预先集成和优化的硬件(诸如服务器、存储装置和联网资源)的组合,以执行由PaaS平台和SaaS平台提供的服务。

[0178] 在一些实施例中,云基础设施系统1002中的资源可以由多个用户共享并且根据需要动态重新分配。此外,可以将资源分配给在不同时区的用户。例如,云基础设施系统1030可以使在第一时区中的第一组用户能够在指定的小时数内利用云基础设施系统的资源,并且然后使相同资源能够被重新分配给位于不同时区的另一组用户,从而使资源的利用率最大化。

[0179] 在某些实施例中,可以提供由云基础设施系统1002的不同组件或模块以及由云基

础设施系统1002提供的服务共享的多个内部共享服务1032。这些内部共享服务可以包括但不限于:安全和身份服务、集成服务、企业储存库服务、企业管理器服务、病毒扫描和白名单服务、高可用性、备份和恢复服务、启用云支持的服务、电子邮件服务、通知服务、文件传输服务等。

[0180] 在某些实施例中,云基础设施系统1002可以提供云基础设施系统中的云服务(例如,SaaS、PaaS和IaaS服务)的综合管理。在一个实施例中,云管理功能可以包括用于供给、管理和跟踪由云基础设施系统1002接收到的客户订阅等的能力。

[0181] 在一个实施例中,如图中所描绘的,云管理功能可以由一个或多个模块提供,其中模块诸如订单管理模块1020、订单编排模块1022、订单供给模块1024、订单管理和监视模块1026、以及身份管理模块1028。这些模块可以包括一个或多个计算机和/或服务器或者使用一个或多个计算机和/或服务器来提供,这些计算机和/或服务器可以是通用计算机、专用服务器计算机、服务器场、服务器集群或任何其他适当的布置/或组合。

[0182] 在示例性操作1034中,使用客户端设备(诸如客户端设备1004、1006或1008)的客户可以通过请求由云基础设施系统1002提供的一个或多个服务并且下订阅由云基础设施系统1002供应的一个或多个服务来的订单来与云基础设施系统1002进行交互。在某些示例中,客户可以访问云用户界面(UI)、云UI 1012、云UI 1014和/或云UI 1016并经由这些UI下订阅订单。云基础设施系统1002响应于客户下订单而接收到的订单信息可以包括识别客户以及客户想要订阅的云基础设施系统1002供应的一个或多个服务的信息。

[0183] 在客户已经下订单之后,经由云UI 1012、1014和/或1016来接收订单信息。

[0184] 在操作1036处,订单存储在订单数据库1018中。订单数据库1018可以是由云基础设施系统1018操作和与其他系统元件一起操作的若干数据库之一。

[0185] 在操作1038处,订单信息被转发到订单管理模块1020。在一些情况下,订单管理模块1020可以被配置为执行与订单相关的计费和记账功能,诸如验证订单,以及在验证后预订订单。

[0186] 在操作1040处,将关于订单的信息传送到订单编排模块1022。订单编排模块1022可以利用订单信息为客户下的订单编排服务和资源的供给。在一些情况下,订单编排模块1022可以使用订单供给模块1024的服务来编排资源的供给以支持订阅的服务。

[0187] 在某些实施例中,订单编排模块1022使得能够管理与每个订单相关联的流程并应用逻辑来确定订单是否应该进行到供给。在操作1042处,在接收到新订阅的订单时,订单编排模块1022向订单供给模块1024发送请求以分配资源并配置履行订阅订单所需的那些资源。订单供给模块1024使得能够为客户订购的服务分配资源。订单供给模块1024提供由云基础设施系统1000提供的云服务和用于供给用于提供所请求的服务的资源的物理实现层之间的抽象层。因此,订单编排模块1022可以与实施细节隔离,诸如服务和资源是否实际上即时供给或预先供给并且仅在请求后才分配/指派。

[0188] 在操作1044处,一旦供给了服务和资源,就可以通过云基础设施系统1002的订单供给模块1024向客户端设备1004、1006和/或1008上的客户发送所提供服务的通知。

[0189] 在操作1046处,订单管理和监视模块1026可以管理和跟踪客户的订阅订单。在一些情况下,订单管理和监视模块1026可以被配置为收集订阅订单中的服务的使用统计信息,诸如所使用的存储量、传输的数据量、用户的数量、以及系统运行时间和系统停机时间

的量。

[0190] 在某些实施例中,云基础设施系统1000可以包括身份管理模块1028。身份管理模块1028可以被配置为提供身份服务,诸如云基础设施系统1000中的访问管理和授权服务。在一些实施例中,身份管理模块1028可以控制关于希望利用由云基础设施系统1002提供的服务的客户的信息。这样的信息可以包括认证这些客户的身份的信息以及描述这些客户被授权相对于各种系统资源(例如,文件、目录、应用、通信端口、存储器段等)执行哪些动作的信息。身份管理模块1028还可以包括对关于每个客户的描述性信息以及关于如何和由谁来访问和修改该描述性信息的管理。

[0191] 图11图示了其中可以实现本发明的各种实施例的示例性计算机系统1100。系统1100可以用于实现上述计算机系统中的任何计算机系统。如图所示,计算机系统1100包括经由总线子系统1102与多个外围子系统进行通信的处理单元1104。这些外围子系统可以包括处理加速单元1106、I/O子系统1108、存储子系统1118和通信子系统1124。存储子系统1118包括有形计算机可读存储介质1122和系统存储器1110。

[0192] 总线子系统1102提供用于让计算机系统1100的各种组件和子系统按意图彼此进行通信的机制。虽然总线子系统1102被示意性地示出为单条总线,但是总线子系统的替代实施例可以利用多条总线。总线子系统1102可以是若干种类型的总线结构中的任何类型的总线,包括存储器总线或存储器控制器、外围总线、以及使用各种总线体系架构中的任何体系架构的局部总线。例如,这种体系架构可以包括工业标准体系架构(ISA)总线、微通道体系架构(MCA)总线、增强型ISA(EISA)总线、视频电子标准协会(VESA)局部总线和外围组件互连(PCI)总线,这些总线可以被实现为按IEEE P1386.1标准制造的Mezzanine总线。

[0193] 可以被实现为一个或多个集成电路(例如,常规微处理器或微控制器)的处理单元1104控制计算机系统1100的操作。一个或多个处理器可以被包括在处理单元1104中。这些处理器可以包括单核处理器或多核处理器。在某些实施例中,处理单元1104可以被实现为一个或多个独立的处理单元1132和/或1134,其中在每个处理单元中包括单核或多核处理器。在其他实施例中,处理单元1104也可以被实现为通过将两个双核处理器集成到单个芯片中形成的四核处理单元。

[0194] 在各种实施例中,处理单元1104可以响应于程序代码执行各种程序并且可以维护多个并发执行的程序或进程。在任何给定的时间,要被执行的程序代码中的一些或全部代码可以驻留在(一个或多个)处理器1104中和/或存储子系统1118中。通过适当的编程,(一个或多个)处理器1104可以提供上述各种功能。计算机系统1100可以附加地包括处理加速单元1106,该处理加速单元1106可以包括数字信号处理器(DSP)、专用处理器等。

[0195] I/O子系统1108可以包括用户接口输入设备和用户接口输出设备。用户接口输入设备可以包括键盘、诸如鼠标或轨迹球的指向设备、结合到显示器中的触摸板或触摸屏、滚动轮、点击轮、拨盘、按钮、开关、小键盘、具有语音命令辨识系统的音频输入设备、麦克风以及其他类型的输入设备。用户接口输入设备可以包括例如运动感测和/或手势辨识设备,诸如**Microsoft Kinect®**运动传感器,其使得用户能够通过使用手势和语音命令的自然用户接口来控制诸如Microsoft **Xbox®**360游戏控制器的输入设备以及与该输入设备进行交互。用户接口输入设备也可以包括眼睛姿势辨识设备,诸如从用户检测眼睛活动(例如,当拍摄照片和/或进行菜单选择时的“眨眼”)并且将眼睛姿势变换为进入输入设备(例

如,Google Glass®)中的输入的Google Glass®眨眼检测器。此外,用户接口输入设备可以包括使用户能够通过语音命令与语音辨识系统(例如,Siri®导航器)进行交互的语音辨识感测设备。

[0196] 用户接口输入设备也可以包括但不限于:三维(3D)鼠标、操纵杆或指向棒(pointing stick)、游戏面板和绘图板、以及音频/视觉设备,诸如扬声器、数码相机、数码摄录机、便携式媒体播放器、网络摄像头(webcam)、图像扫描仪、指纹扫描仪、条形码阅读器3D扫描仪、3D打印机、激光测距仪和视线跟踪设备。此外,用户接口输入设备可以包括例如医学成像输入设备,诸如计算机断层扫描、磁共振成像、正电子发射断层显像、医疗超声设备。用户接口输入设备也可以包括例如音频输入设备(诸如MIDI键盘、数字乐器等)。

[0197] 用户接口输出设备可以包括显示子系统、指示灯,或者诸如音频输出设备的非可视显示器等。显示子系统可以是阴极射线管(CRT)、诸如使用液晶显示器(LCD)或等离子显示器的平板设备、投影设备、触摸屏等。一般而言,术语“输出设备”的使用旨在包括用于从计算机系统1100向用户或其他计算机输出信息的所有可能类型的设备和机制。例如,用户接口输出设备可以包括但不限于:可视地传达文本、图形和音频/视频信息的各种显示设备,诸如监视器、打印机、扬声器、耳机、汽车导航系统、绘图仪、语音输出设备以及调制解调器。

[0198] 计算机系统1100可以包括包含软件元件的被示出为当前位于系统存储器1110内的存储子系统1118。系统存储器1110可以存储在处理单元1104上可执行和可加载的程序指令,以及在这些程序的执行期间所生成的数据。

[0199] 取决于计算机系统1100的配置和类型,系统存储器1110可以是易失性的(诸如随机存取存储器(RAM))和/或非易失性的(诸如只读存储器(ROM)、闪存存储器等)。RAM通常包含可被处理单元1104立即访问和/或目前正在被处理单元1104操作和执行的数据和/或程序模块。在一些实施方式中,系统存储器1110可以包括多种不同类型的存储器,诸如静态随机存取存储器(SRAM)或动态随机存取存储器(DRAM)。在一些实施方式中,诸如在启动期间,包含有助于在计算机系统1100内的元件之间传送信息的基本例程的基本输入/输出系统(BIOS)通常可以被存储在ROM中。作为示例而非限制,系统存储器1110也示出了可以包括客户端应用、Web浏览器、中间层应用、关系数据库管理系统(RDBMS)等的应用程序1112,程序数据1114,以及操作系统1116。作为示例,操作系统1116可以包括各种版本的Microsoft Windows®、Apple Macintosh®和/或Linux操作系统、各种可商业获得的UNIX®或类UNIX操作系统(包括但不限于各种GNU/Linux操作系统、Google Chrome®OS等)和/或诸如iOS、Windows®Phone、Android®OS、BlackBerry®10OS和Palm®OS操作系统的移动操作系统。

[0200] 存储子系统1118也可以提供用于存储提供一些实施例的功能的基本编程和数据构造的有形计算机可读存储介质。当被处理器执行时提供上述功能的软件(程序、代码模块、指令)可以被存储在存储子系统1118中。这些软件模块或指令可以被处理单元1104执行。存储子系统1118也可以提供用于存储根据本发明使用的数据的储存库。

[0201] 存储子系统1100也可以包括可进一步连接到计算机可读存储介质1122的计算机可读存储介质读取器1120。与系统存储器1110一起并且可选地与系统存储器1110相结合,计算机可读存储介质1122可以全面地表示用于临时和/或更持久地包含、存储、发送和检索

计算机可读信息的远程、本地、固定和/或可移除的存储设备加存储介质。

[0202] 包含代码或代码的一部分的计算机可读存储介质1122也可以包括本领域已知或使用的任何适当的介质,包括存储介质和通信介质,诸如但不限于:以用于信息的存储和/或传输的任何方法或技术实现的易失性和非易失性的、可移除和不可移除的介质。这可以包括有形的计算机可读存储介质,诸如RAM、ROM、电可擦除可编程ROM(EEPROM)、闪存存储器或其他存储器技术、CD-ROM、数字多功能盘(DVD)或其他光学存储装置、磁带盒、磁带、磁盘存储装置或其他磁存储设备、或者其他有形计算机可读介质。这也可以包括非有形的计算机可读介质,诸如数据信号、数据传输、或者可以用于发送期望信息并且可以被计算系统1100访问的任何其他介质。

[0203] 作为示例,计算机可读存储介质1122可以包括从不可移除的非易失性磁介质读取或写入到该不可移除的非易失性磁介质的硬盘驱动器、从可移除的非易失性磁盘读取或写入到该可移除的非易失性磁盘的磁盘驱动器、以及从可移除的非易失性光盘(诸如CD ROM、DVD和**Blu-Ray®**(蓝光)盘或其他光学介质)读取或写入到该可移除的非易失性光盘的光盘驱动器。计算机可读存储介质1122可以包括但不限于:**Zip®**驱动器、闪存存储器卡、通用串行总线(USB)闪存驱动器、安全数字(SD)卡、DVD盘、数字音频带等。计算机可读存储介质1122也可以包括:基于非易失性存储器的固态驱动器(SSD)(诸如基于闪存存储器的SSD、企业闪存驱动器、固态ROM等)、基于易失性存储器的SSD(诸如基于固态RAM、动态RAM、静态RAM、DRAM的SSD、磁阻RAM(MRAM)SSD)、以及使用基于DRAM和闪存存储器的SSD的组合的混合SSD。盘驱动器及其相关联的计算机可读介质可以为计算机系统1100提供计算机可读指令、数据结构、程序模块以及其他数据的非易失性存储。

[0204] 通信子系统1124提供到其他计算机系统和网络的接口。通信子系统1124用作用于从其他系统接收数据和从计算机系统1100向其他系统发送数据的接口。例如,通信子系统1124可以使计算机系统1100能够经由互联网连接到一个或多个设备。在一些实施例中,通信子系统1124可以包括用于访问无线语音和/或数据网络的射频(RF)收发器组件(例如,使用蜂窝电话技术、先进数据网络技术,诸如3G、4G或EDGE(用于全球演进的增强型数据速率)、Wi-Fi(IEEE 1202.11系列标准)或其他移动通信技术、或者其任何组合)、全球定位系统(GPS)接收器组件、和/或其他组件。在一些实施例中,作为无线接口的附加或者替代,通信子系统1124可以提供有线网络连接(例如,以太网)。

[0205] 在一些实施例中,通信子系统1124也可以代表可以使用计算机系统1100的一个或多个用户来接收以结构化和/或非结构化的数据馈送1126、事件流1128、事件更新1130等形式的输入通信。

[0206] 作为示例,通信子系统1124可以被配置为实时地从社交网络和/或其他通信服务的用户接收数据馈送1126,诸如**Twitter®**馈送、**Facebook®**更新、诸如丰富站点概要(RSS)馈送的web馈送、和/或来自一个或多个第三方信息源的实时更新。

[0207] 此外,通信子系统1124也可以被配置为接收以连续数据流形式的数据,该数据可以包括本质上可以是连续的或无界的没有明确终止的实时事件的事件流1128和/或事件更新1130。生成连续数据的应用的示例可以包括例如传感器数据应用、金融价格收报机、网络性能测量工具(例如,网络监视和流量管理应用)、点击流分析工具、汽车流量监视等。

[0208] 通信子系统1124也可以被配置为向一个或多个数据库输出结构化和/或非结构化

的数据馈送1126、事件流1128、事件更新1130等，该一个或多个数据库可以与耦合到计算机系统1100的一个或多个流传输数据源计算机进行通信。

[0209] 计算机系统1100可以是各种类型中的一种类型，包括手持便携式设备(例如，iPhone®蜂窝电话、iPad®计算平板电脑、PDA)、可穿戴设备(例如，Google Glass®头戴式显示器)、PC、工作站、大型机、信息站(kiosk)、服务器机架、或任何其他数据处理系统。

[0210] 由于计算机和网络的不断变化的本质，在图中绘出的计算机系统1100的描述仅旨在作为具体的示例。具有比图中绘出的系统更多或更少组件的许多其他配置是可能的。例如，也可以使用定制的硬件和/或可以在硬件、固件、软件(包括小程序应用(applet))或其组合中实现特定的元素。另外，可以采用到诸如网络输入/输出设备的其他计算设备的连接。基于本文提供的公开内容和教导，本领域普通技术人员将认识到实现各种实施例的其他方式和/或方法。

[0211] 在前述说明书中，参考本发明的各方面的具体实施例对本发明的各方面进行了描述，但是本领域技术人员将认识到的是，本发明不限于此。上述发明的各个特征和方面可以被单独使用或联合使用。另外，在不脱离本说明书的更广泛精神和范围的情况下，可以在除本文所述的那些环境和应用之外的任何数量的环境和应用中利用实施例。因而，本说明书和附图应当被认为是说明性的而不是限制性的。

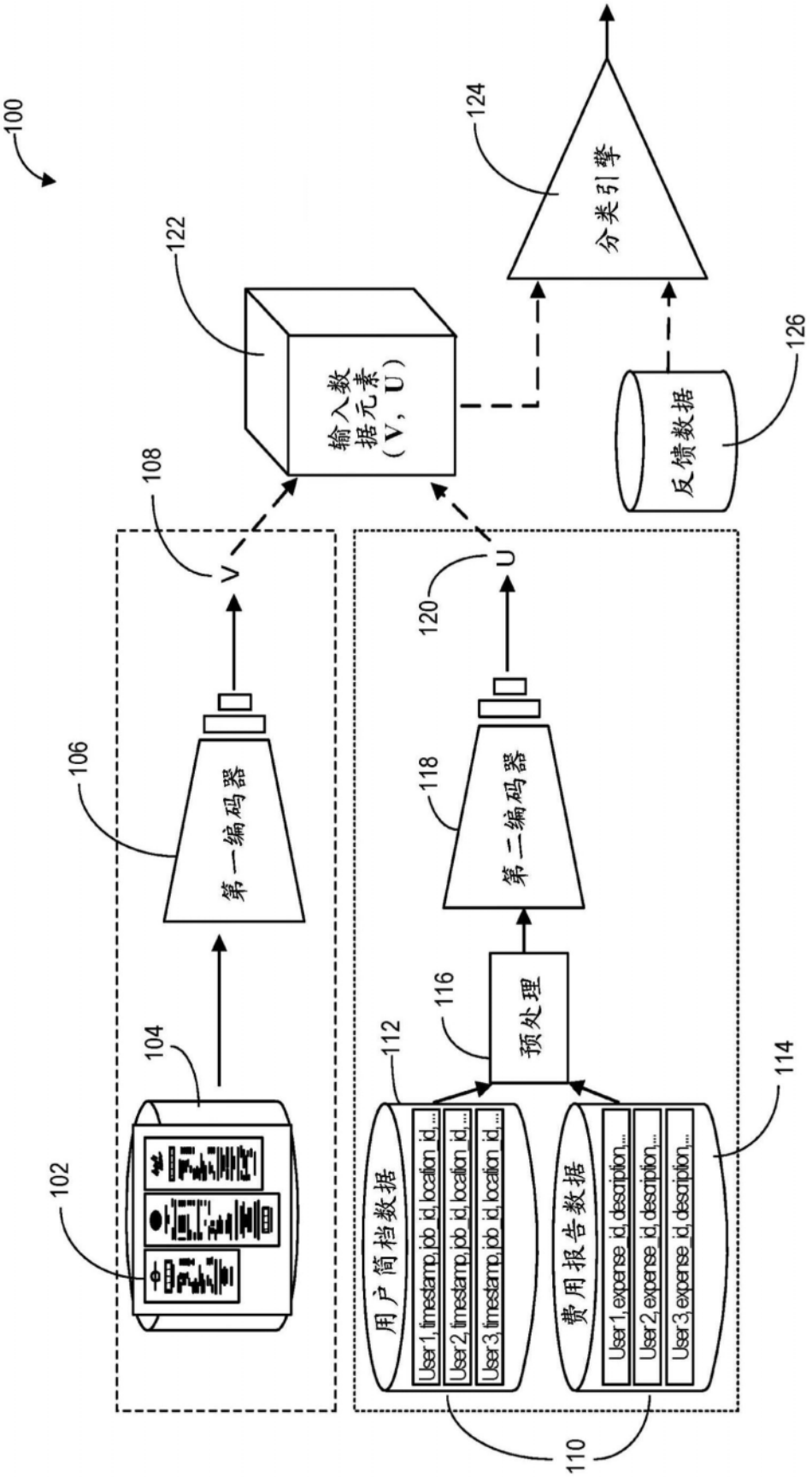

100
102
104
106
第一编码器
108
V
122
输入数据元素 (V, U)
124
分类引擎
126
反馈数据
110
112
用户简档数据
User1,timestamp,job id,location id,...
User2,timestamp,job id,location id,...
User3,timestamp,job id,location id,...
114
费用报告数据
User1,expense id,description,...
User2,expense id,description,...
User3,expense id,description,...
116
预处理
118
第二编码器
120
U

图1

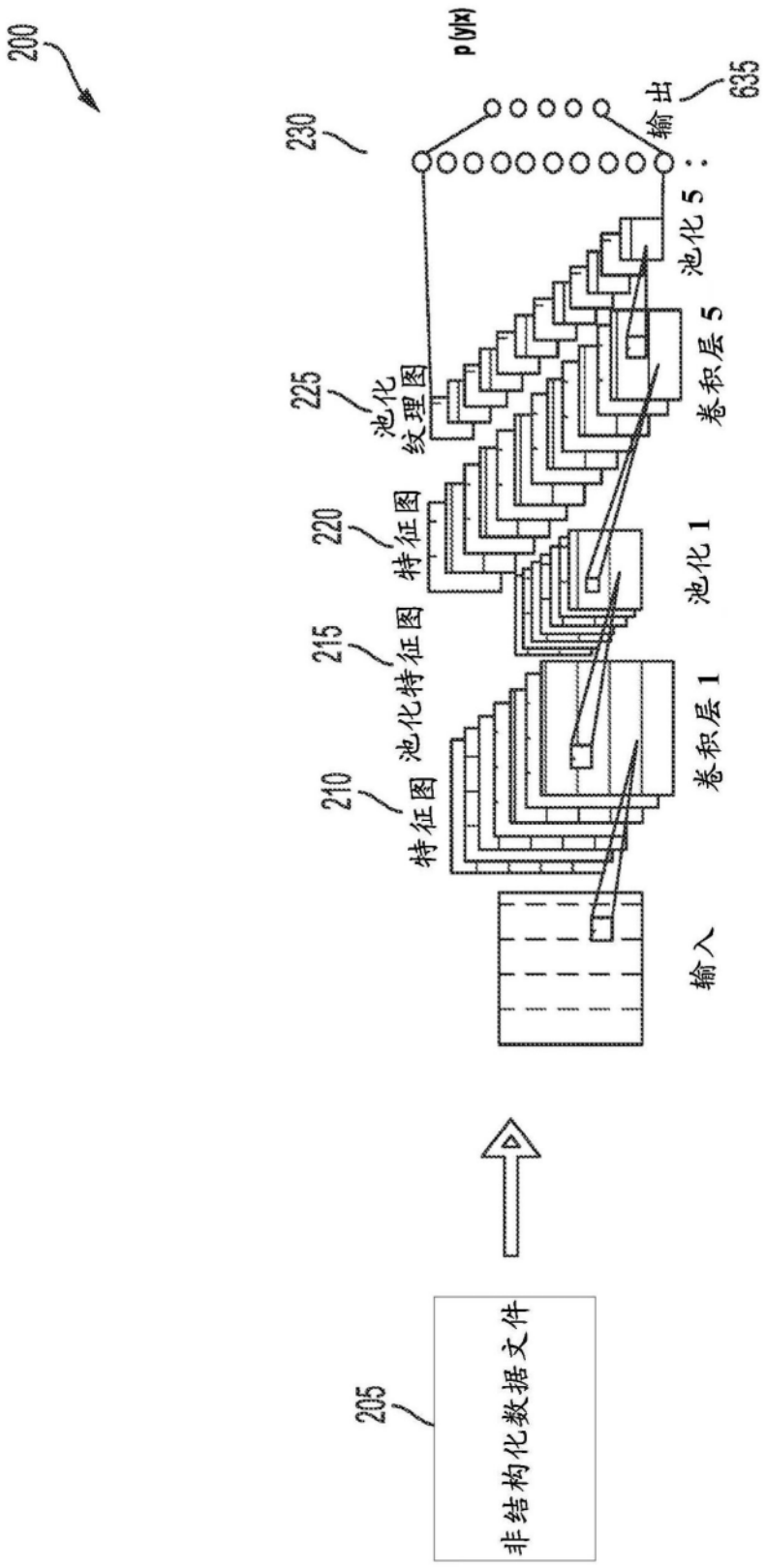

200
205
非结构化数据文件
210
特征图
215
池化特征图
220
特征图
225
池化纹理图
230
p(y|x)
输出
6335
输入
卷积层 1
池化 1
卷积层 5
池化 5

图2

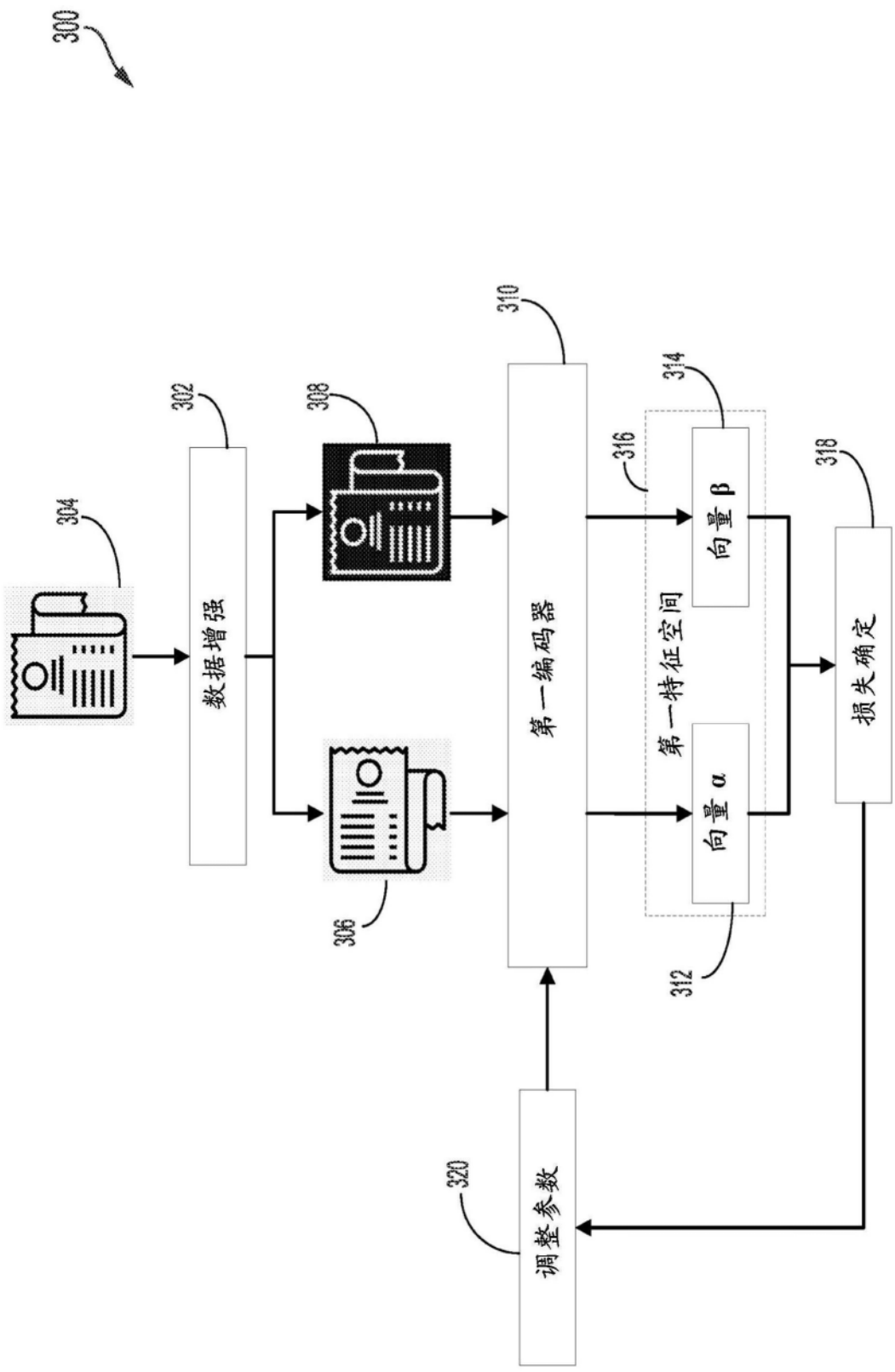

300
304
数据增强
302
308
306
第一编码器
310
第一特征空间
316
向量 β
向量 α
314
312
损失确定
318
调整参数
320

图3

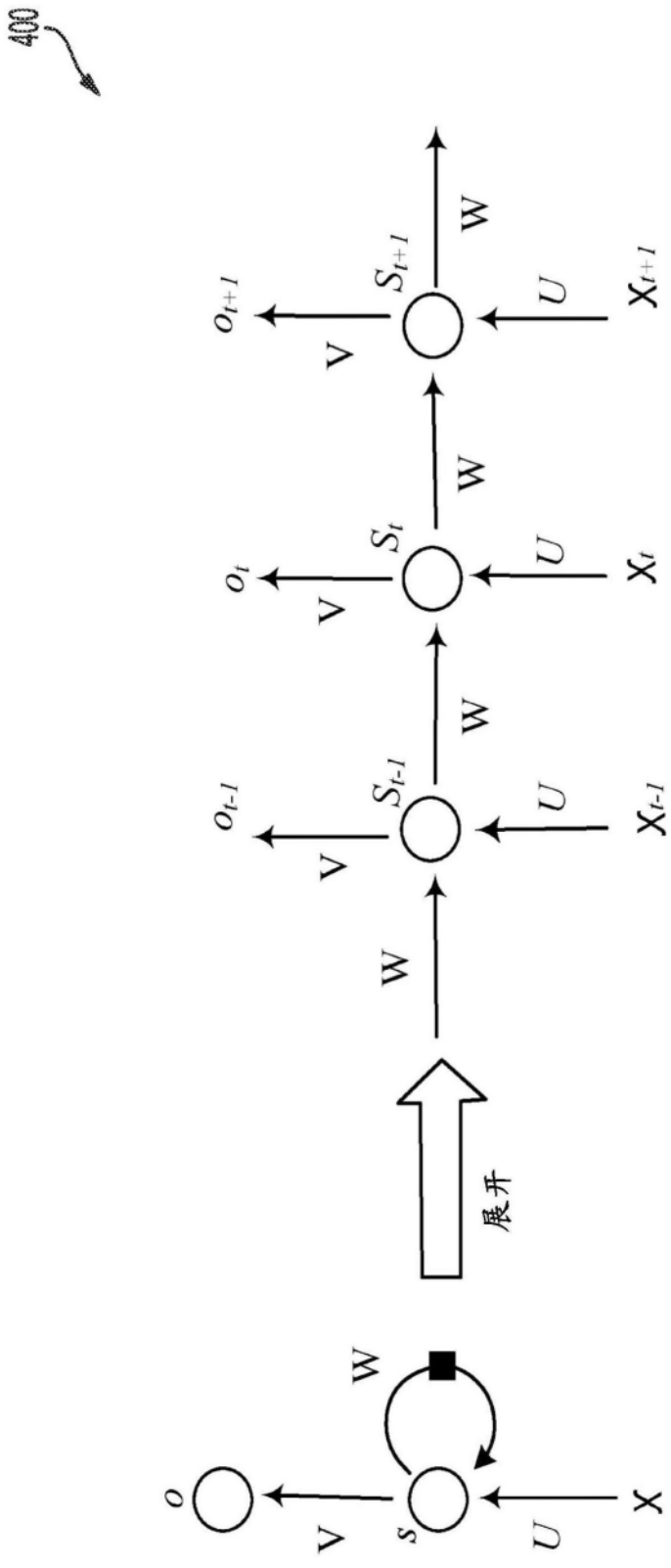


图4

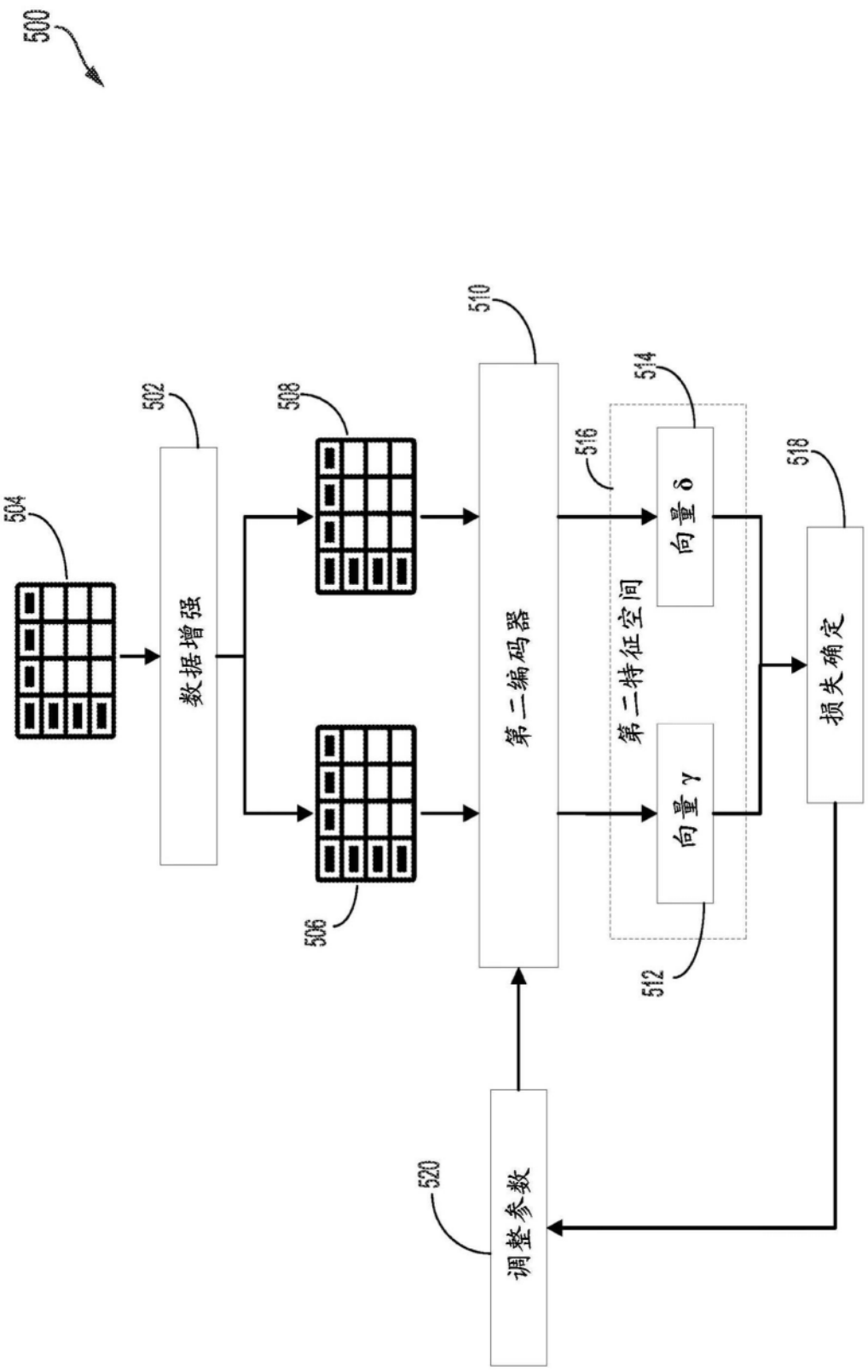

500
504
502
数据增强
508
506
510
第二编码器
514
516
第二特征空间
向量δ
向量γ
512
518
损失确定
520
调整参数

图5

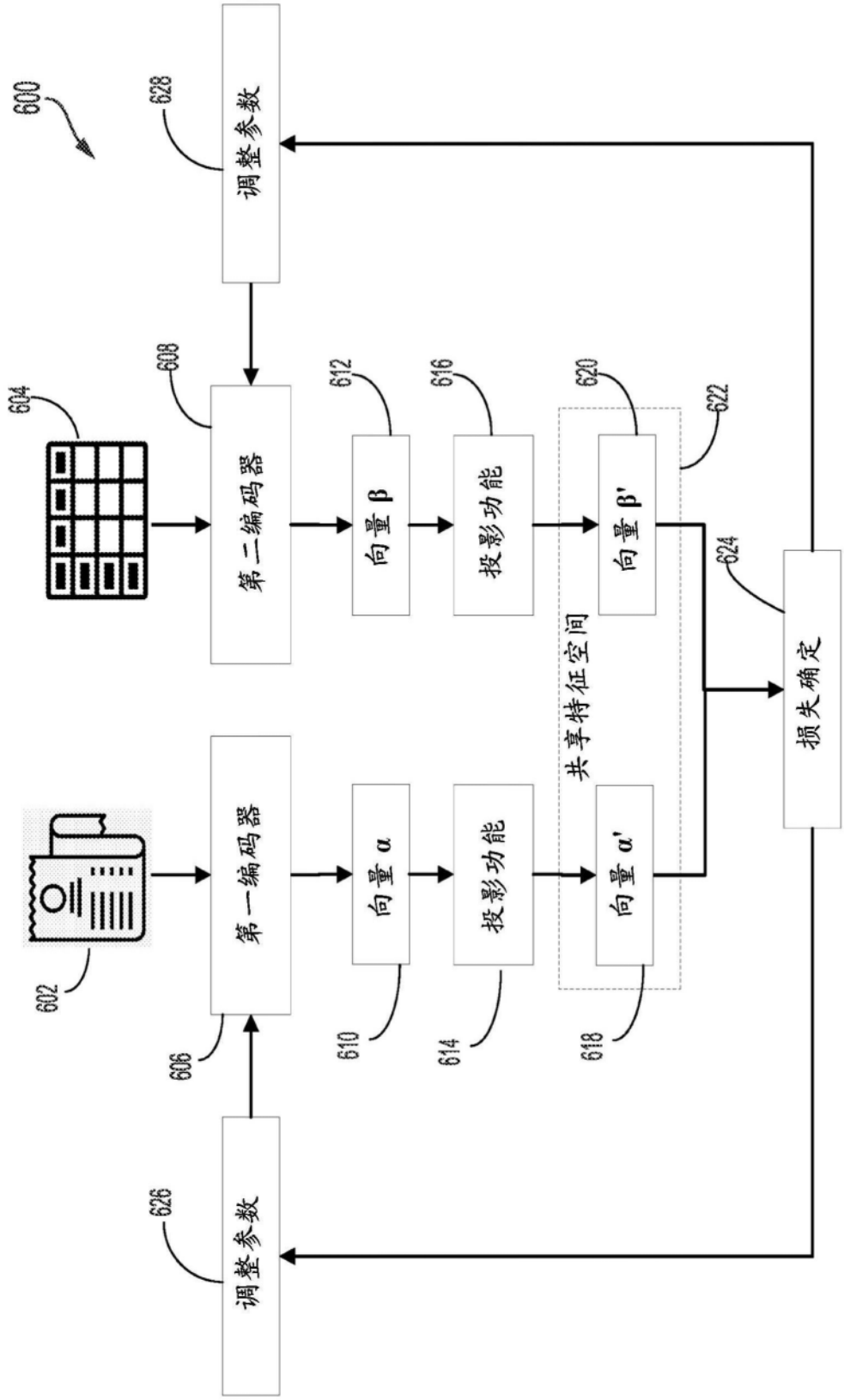

600
628
调整参数
608
604
第二编码器
612
向量β
616
投影功能
620
向量β'
622
624
损失确定
共享特征空间
602
第一编码器
610
向量α
614
投影功能
618
向量α'
606
626
调整参数

图6

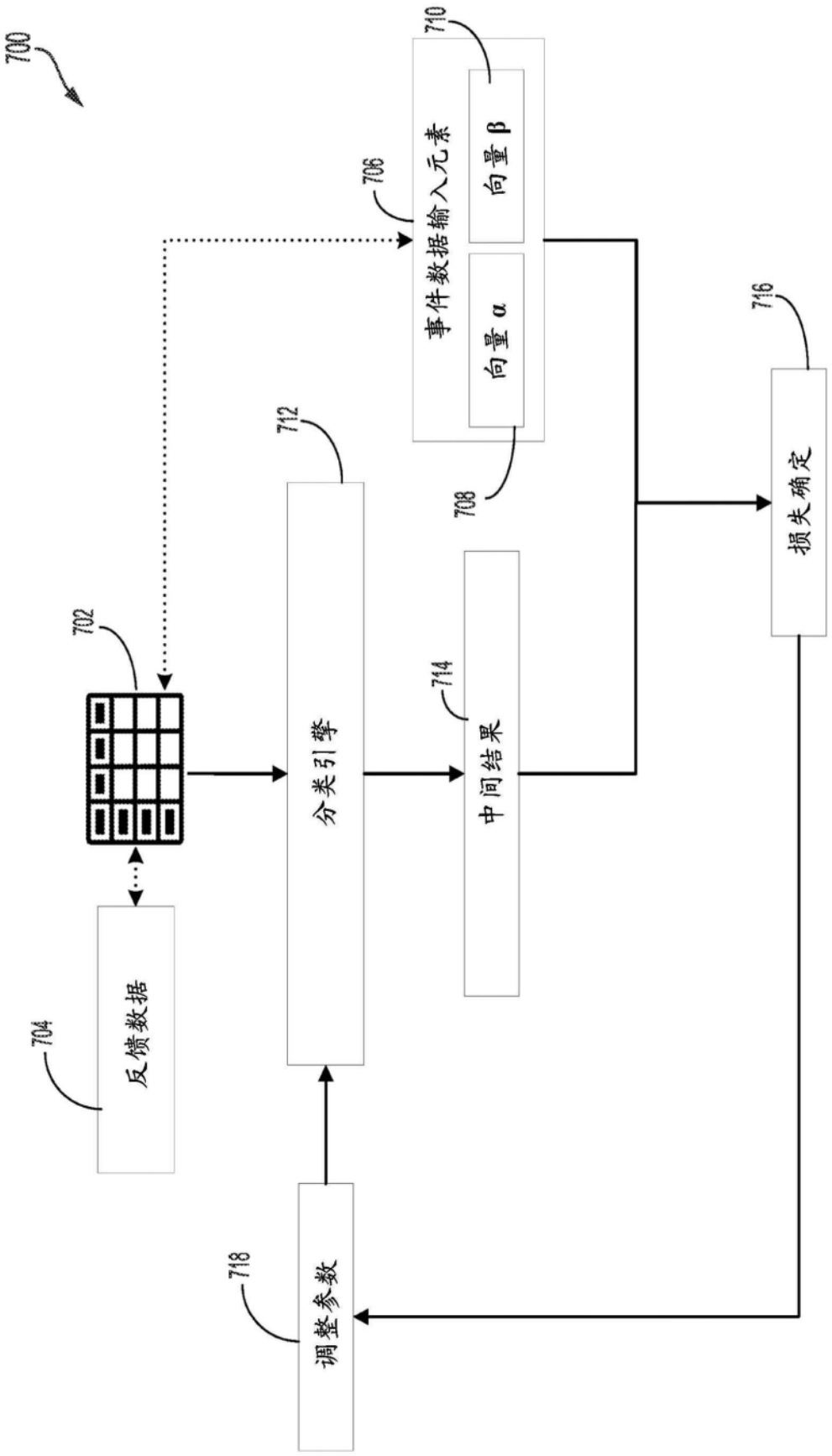

700
710
事件数据输入元素
向量 β
向量 α
706
708
712
分类引擎
714
中间结果
716
损失确定
702
704
反馈数据
718
调整参数

图7

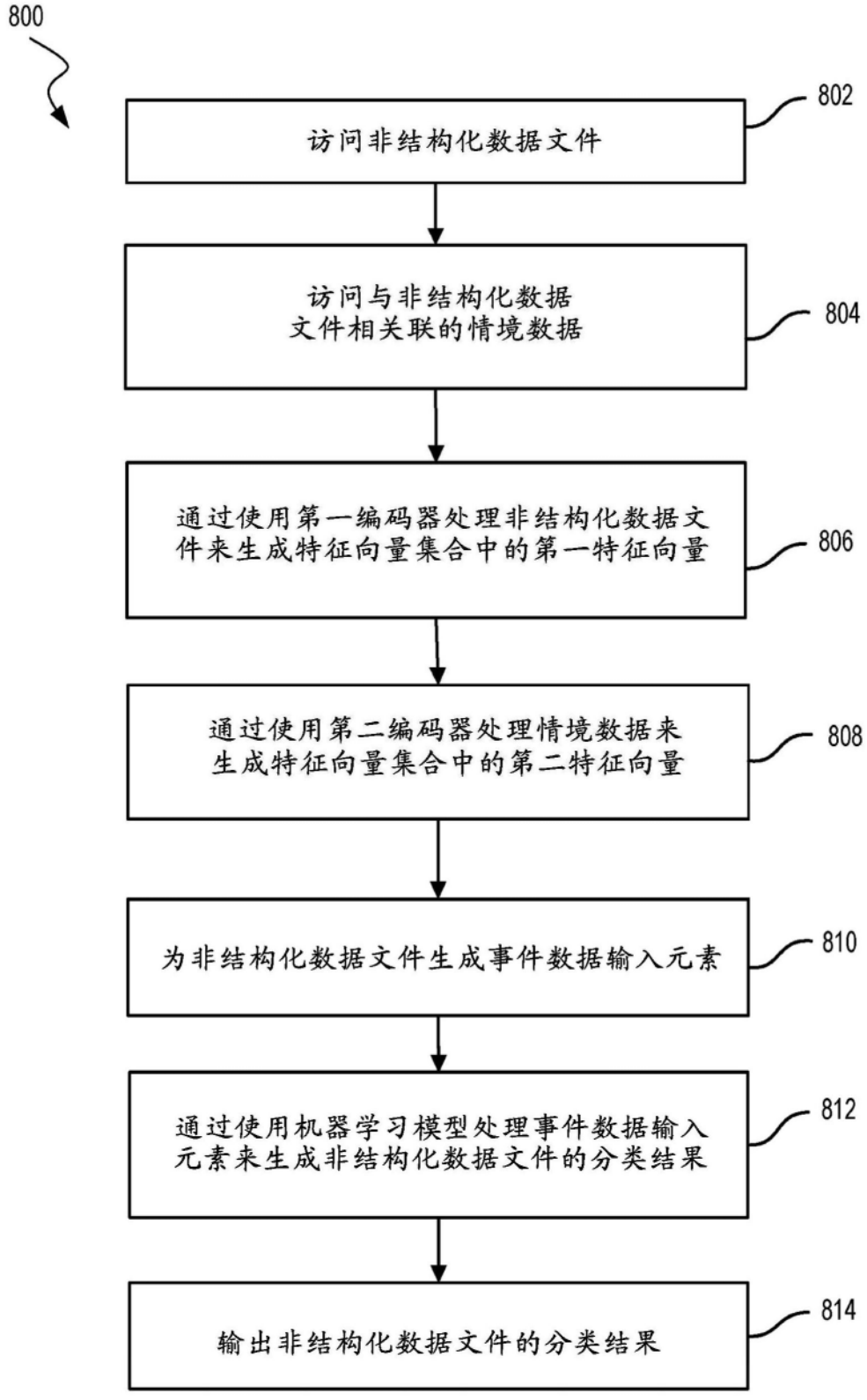

800
访问非结构化数据文件
802
访问与非结构化数据
文件相关联的情境数据
804
通过使用第一编码器处理非结构化数据文
件来生成特征向量集合中的第一特征向量
806
通过使用第二编码器处理情境数据来
生成特征向量集合中的第二特征向量
808
为非结构化数据文件生成事件数据输入元素
810
通过使用机器学习模型处理事件数据输入
元素来生成非结构化数据文件的分类结果
812
输出非结构化数据文件的分类结果
814

图8

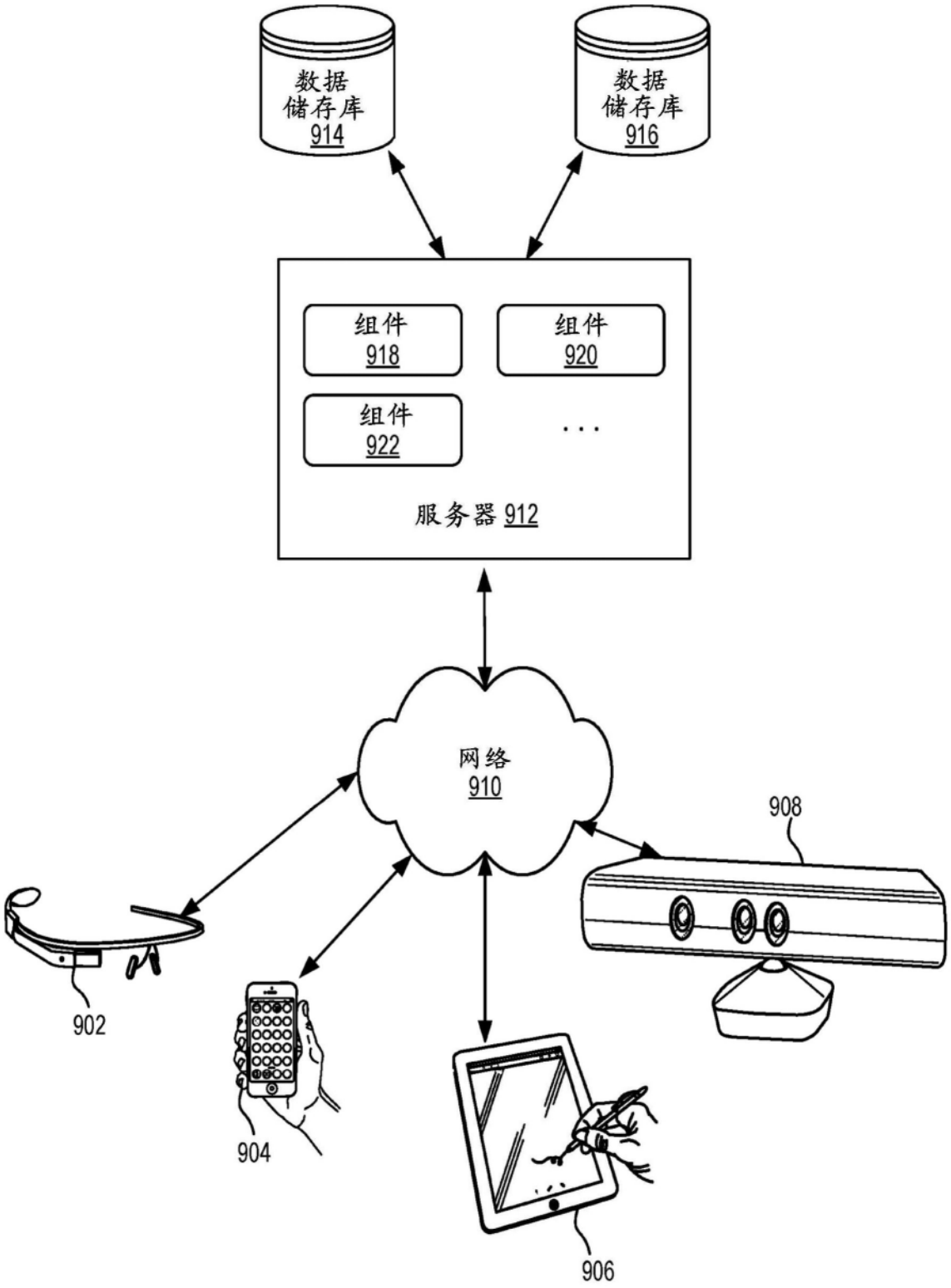

数据
储存库
914
数据
储存库
916
组件
918
组件
920
组件
922
...
服务器 912
网络
910
908
902
904
906

图9

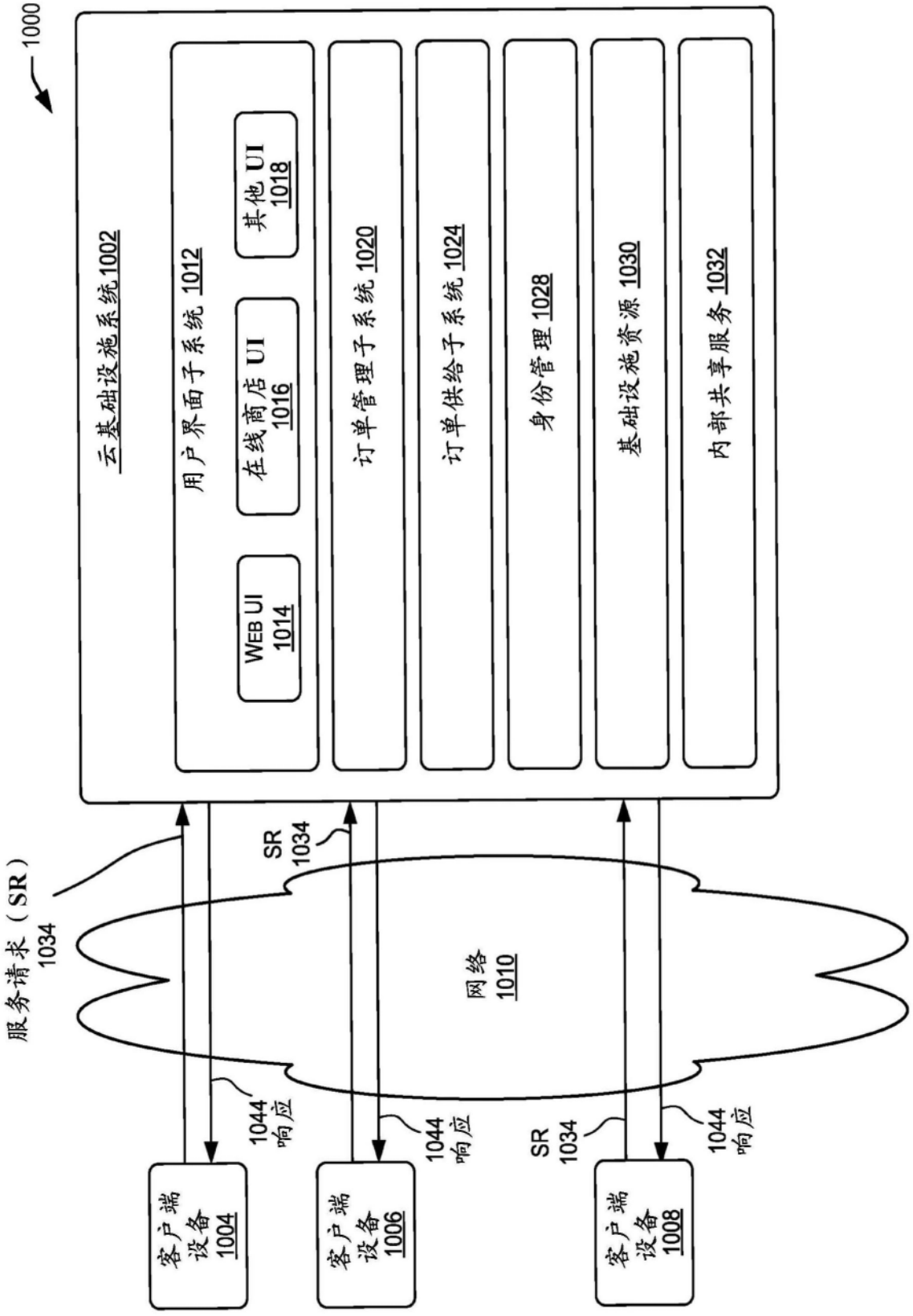

1000
云基础设施系统1002
用户界面子系统 1012
WEB UI 1014
在线商店 UI 1016
其他 UI 1018
订单管理子系统 1020
订单供给子系统 1024
身份管理 1028
基础设施资源 1030
内部共享服务 1032
服务请求（SR）1034
SR 1034
网络 1010
1044 响应
1044 响应
SR 1034
1044 响应
客户端设备 1004
客户端设备 1006
客户端设备 1008

图10

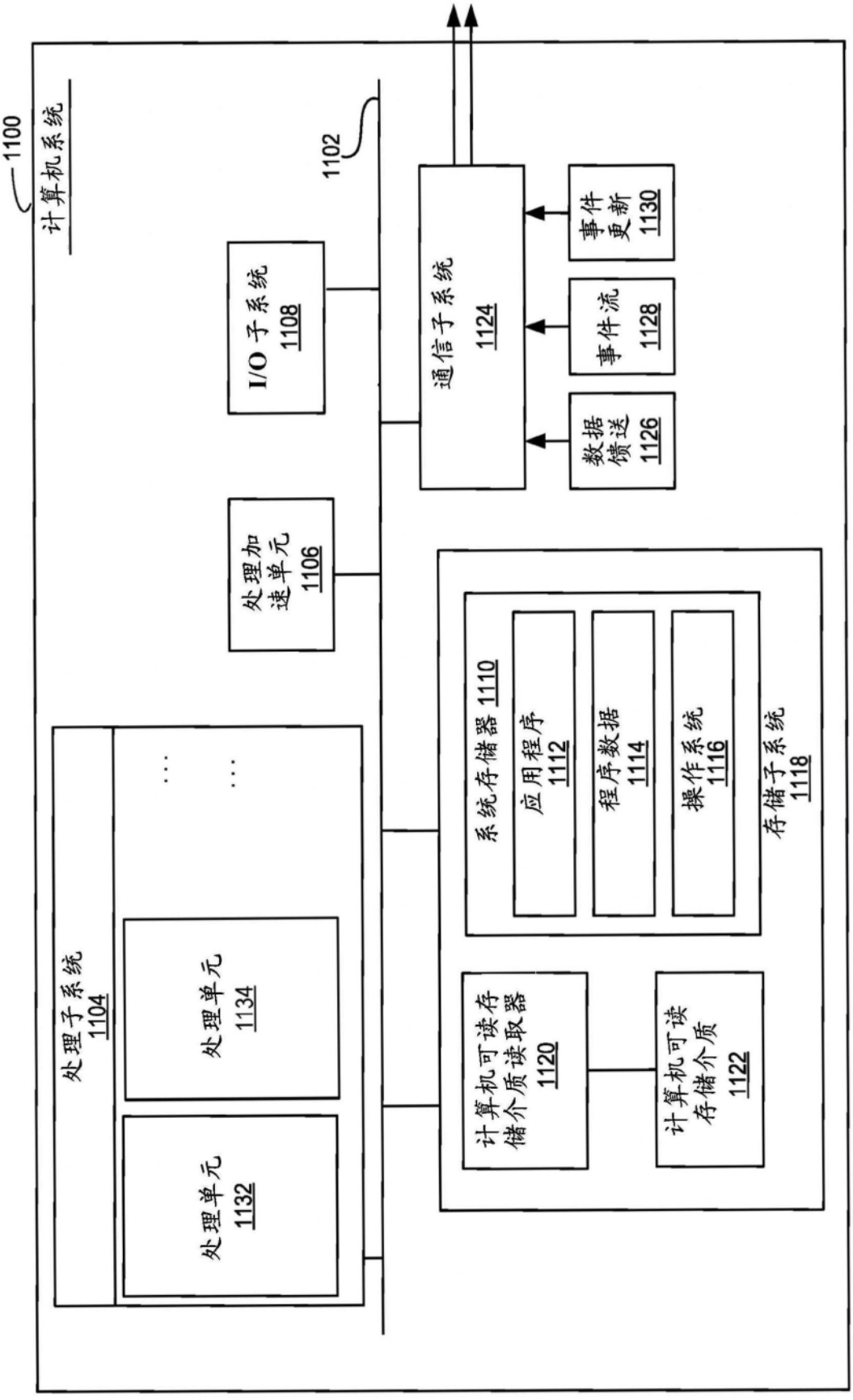

1100
计算机系统
1102
I/O 子系统
1108
通信子系统
1124
事件更新
1130
事件流
1128
数据馈送
1126
处理加速单元
1106
处理子系统
1104
处理单元
1132
处理单元
1134
...
...
系统存储器 1110
应用程序
1112
程序数据
1114
操作系统
1116
存储子系统
1118
计算机可读存储介质读取器
1120
计算机可读存储介质
1122

图11