(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 112334917 A
(43)申请公布日 2021.02.05

(21)申请号 201980042541.1
(22)申请日 2019.04.23
(30)优先权数据
62/786941 2018.12.31 US
(85)PCT国际申请进入国家阶段日
2020.12.24
(86)PCT国际申请的申请数据
PCT/US2019/028687 2019.04.23
(87)PCT国际申请的公布数据
WO2020/142110 EN 2020.07.09
(71)申请人 英特尔公司
地址 美国加利福尼亚州
(72)发明人 O·波哥列克 A·纳什图特 O·本-沙洛姆 D·克里莫夫 R·凯勒曼 G·巴恩哈特-马根 V·苏霍姆利诺夫
(74)专利代理机构 中国专利代理(香港)有限公司 72001
代理人 浩路 周学斌
(51)Int.Cl.
*G06N 3/08*(2006.01)
*G06N 3/04*(2006.01)
*G06N 20/00*(2006.01)

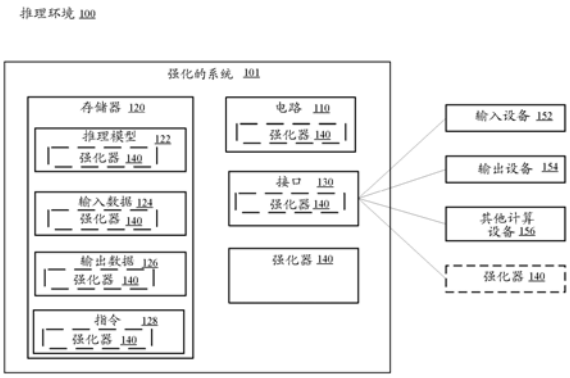
权利要求书3页 说明书91页 附图71页

(54)发明名称

对采用人工智能的系统进行防护

(57)摘要

提供了针对各种攻击来强化AI系统的技术和装置。在不同的技术和装置当中，提供了将推理模型的域扩展成包括可见类以及隐藏类两者的技术和装置。隐藏类可以用于检测针对所述模型的可能探测攻击。




CN 112334917 A

1.一种装置,包括:

电路;以及

耦合到所述电路的存储器,存储器存储指令,所述指令在由所述电路执行时使所述电路:

从输入设备接收输入数据;

部分地基于利用输入数据执行推理模型来生成输出数据,输出数据包括可见类或隐藏类之一的指示;

确定输出数据是否包括隐藏类的指示;以及

基于输出数据不包括隐藏类的指示的确定而将所生成的输出提供给输出消费者。

2.根据权利要求1所述的装置,存储器进一步存储指令,所述指令在由所述电路执行时使所述电路:基于输出数据包括隐藏类的指示的确定而向监测系统发送警报。

3.根据权利要求2所述的装置,存储器进一步存储指令,所述指令在由所述电路执行时使所述电路:

响应于输出数据包括隐藏类的指示的确定而生成混淆的输出,混淆的输出包括可见类的指示;以及

响应于输出数据包括隐藏类的指示的确定而将混淆的输出提供给输出消费者。

4.根据权利要求3所述的装置,存储器进一步存储指令,所述指令在由所述电路执行时使所述电路:

响应于输出数据包括隐藏类的指示的确定而收集取证数据,所述取证数据包括输入数据和输入设备中的至少一个的指示;以及

响应于输出数据包括隐藏类的指示的确定而向监测系统提供所述取证数据。

5.根据权利要求3所述的装置,可见类包括多个可见类,输出数据包括隐藏类或者所述多个可见类之一的指示。

6.根据权利要求1至5中任一项所述的装置,隐藏类包括多个隐藏类,输出数据包括所述多个隐藏类之一或所述多个可见类之一的指示。

7.根据权利要求6所述的装置,其中所述多个隐藏类中的至少一个对应于列入黑名单的输入。

8.根据权利要求6所述的装置,存储器进一步存储指令,所述指令在由所述电路执行时使所述电路:

响应于输出数据包括隐藏类的指示的确定而标识与输出数据所指示的所述多个隐藏类之一相关联的所述多个可见类之一;以及

生成混淆的输出以包括所述多个可见类中的所标识的可见类的指示。

9.根据权利要求1至5中任一项所述的装置,包括可信执行环境(TEE),所述TEE包括所述电路和存储器。

10.一种非暂时性计算机可读存储介质,包括指令,所述指令在由电路执行时使所述电路:

从输入设备接收输入数据;

部分地基于利用输入数据执行推理模型来生成输出数据,输出数据包括可见类或隐藏类之一的指示;

确定输出数据是否包括隐藏类的指示;以及

基于输出数据不包括隐藏类的指示的确定而将所生成的输出提供给输出消费者。

11.根据权利要求10所述的非暂时性计算机可读存储介质,存储器进一步存储指令,所述指令在由所述电路执行时使所述电路:基于输出数据包括隐藏类的指示的确定而向监测系统发送警报。

12.根据权利要求11所述的非暂时性计算机可读存储介质,进一步存储指令,所述指令在由所述电路执行时使所述电路:

响应于输出数据包括隐藏类的指示的确定而生成混淆的输出,混淆的输出包括可见类的指示;以及

响应于输出数据包括隐藏类的指示的确定而将混淆的输出提供给输出消费者。

13.根据权利要求12所述的非暂时性计算机可读存储介质,进一步存储指令,所述指令在由所述电路执行时使所述电路:

响应于输出数据包括隐藏类的指示的确定而收集取证数据,所述取证数据包括输入数据和输入设备中的至少一个的指示;以及

响应于输出数据包括隐藏类的指示的确定而向监测系统提供所述取证数据。

14.根据权利要求12所述的非暂时性计算机可读存储介质,可见类包括多个可见类,输出数据包括隐藏类或者所述多个可见类之一的指示。

15.根据权利要求10至14中任一项所述的非暂时性计算机可读存储介质,隐藏类包括多个隐藏类,输出数据包括所述多个隐藏类之一或所述多个可见类之一的指示。

16.根据权利要求15所述的非暂时性计算机可读存储介质,其中所述多个隐藏类中的至少一个对应于列入黑名单的输入。

17.根据权利要求15所述的非暂时性计算机可读存储介质,进一步存储指令,所述指令在由所述电路执行时使所述电路:

响应于输出数据包括隐藏类的指示的确定而标识与输出数据所指示的所述多个隐藏类之一相关联的所述多个可见类之一;以及

生成混淆的输出以包括所述多个可见类中的所标识的可见类的指示。

18.根据权利要求10至14中任一项所述的非暂时性计算机可读存储介质,其中所述电路是可信执行环境(TEE)的一部分。

19.一种计算机实现方法,包括:

从输入设备接收输入数据;

部分地基于利用输入数据执行推理模型来生成输出数据,输出数据包括可见类或隐藏类之一的指示;

确定输出数据是否包括隐藏类的指示;以及

基于输出数据不包括隐藏类的指示的确定而将所生成的输出提供给输出消费者。

20.根据权利要求19所述的计算机实现方法,包括:基于输出数据包括隐藏类的指示的确定而向监测系统发送警报。

21.根据权利要求20所述的计算机实现方法,包括:

响应于输出数据包括隐藏类的指示的确定而生成混淆的输出,混淆的输出包括可见类的指示;以及

响应于输出数据包括隐藏类的指示的确定而将混淆的输出提供给输出消费者。

22.根据权利要求21所述的计算机实现方法,包括:

响应于输出数据包括隐藏类的指示的确定而收集取证数据,所述取证数据包括输入数据和输入设备中的至少一个的指示;以及

响应于输出数据包括隐藏类的指示的确定而向监测系统提供所述取证数据。

23.根据权利要求21所述的计算机实现方法,可见类包括多个可见类,输出数据包括隐藏类或者所述多个可见类之一的指示。

24.根据权利要求23所述的计算机实现方法,隐藏类包括多个隐藏类,输出数据包括所述多个隐藏类之一或所述多个可见类之一的指示。

25.根据权利要求24所述的计算机实现方法,包括:

响应于输出数据包括隐藏类的指示的确定而标识与输出数据所指示的所述多个隐藏类之一相关联的所述多个可见类之一;以及

生成混淆的输出以包括所述多个可见类中的所标识的可见类的指示。

# 对采用人工智能的系统进行防护

[0001] 相关申请的交叉引用

本申请要求于2018年12月31日提交的题为“Securing Systems Employing Artificial Intelligence”的序列号为62/786,941的美国临时申请的权益，该申请的全部内容通过引用并入本文中。

## 背景技术

[0002] 也被称为机器智能的人工智能(AI)指代感知其环境并且采取动作以使成功地实现各种目标的机会最大化的设备、系统或过程。AI在众多领域中被采用，诸如例如汽车、工业、航空航天、军事、制造、旅行、物流、人类交互等。此外，AI依赖于众多科学和学术学科，诸如例如数学搜索和优化、人工神经网络、统计学、概率、经济学等。

[0003] 通常并且在极其高的级别处，AI系统可以被描述为黑箱，其中输入被提供给该箱，该黑箱基于输入进行一些推理，并且然后提供与该推理相对应的(一个或多个)输出。(一个或多个)输出用于做出决策或确定某些动作过程。在某些应用中，AI系统可以指导或控制动作或行为。例如，AI系统可以耦合到设备控制系统，其中可以基于来自AI系统的输出来致动控件。作为特定示例，AI系统可以耦合到传感器和/或相机以及汽车中的制动系统。这种AI系统可以基于从传感器和/或相机接收到的输入、以及根据这些输入做出的推理来使制动器致动。

[0004] 由于AI系统的复杂性以及它们在众多领域中增加的普及率和使用，因此对AI系统进行防护或针对攻击来强化(harden)AI系统是重要的。

## 附图说明

[0005] 图1图示了第一示例推理环境。

[0006] 图2图示了第一示例技术。

[0007] 图3图示了第二示例技术。

[0008] 图4图示了强化的系统的第一示例。

[0009] 图5图示了逻辑流程的第一示例。

[0010] 图6图示了强化的系统的第二示例。

[0011] 图7图示了逻辑流程的第二示例。

[0012] 图8图示了强化的系统的第三示例。

[0013] 图9图示了强化的系统的第四示例。

[0014] 图10A图示了第五示例强化的系统。

[0015] 图10B图示了第一示例训练/测试数据集。

[0016] 图11图示了强化的系统的第六示例。

[0017] 图12图示了逻辑流程的第三示例。

[0018] 图13图示了逻辑流程的第四示例。

[0019] 图14图示了强化的系统的第七示例。

[0020] 图15图示了输入设备。

[0021] 图16图示了第三示例技术。

[0022] 图17图示了逻辑流程的第五示例。

[0023] 图18图示了第一示例推理环境。

[0024] 图19图示了第二示例推理环境。

[0025] 图20图示了第三示例推理环境。

[0026] 图21图示了第四示例技术。

[0027] 图22图示了描绘了用于推理模型的示例域的图形。

[0028] 图23A图示了第四示例推理环境。

[0029] 图23B图示了第二示例训练/测试数据集。

[0030] 图24图示了第五示例推理环境。

[0031] 图25图示了第六示例逻辑流程。

[0032] 图26图示了强化的系统的第八示例。

[0033] 图27图示了第六示例推理环境。

[0034] 图28图示了第七示例逻辑流程。

[0035] 图29图示了第八示例推理环境。

[0036] 图30图示了第八示例逻辑流程。

[0037] 图31图示了强化的系统的第九示例。

[0038] 图32图示了第九示例逻辑流程。

[0039] 图33图示了第五示例技术。

[0040] 图34图示了强化的系统的第十示例。

[0041] 图35图示了逻辑流程的第十示例。

[0042] 图36图示了强化的系统的第十一示例。

[0043] 图37图示了逻辑流程的第十一示例。

[0044] 图38图示了强化的系统的第十二示例。

[0045] 图39图示了逻辑流程的第十二示例。

[0046] 图40图示了强化的系统的第十三示例。

[0047] 图41图示了逻辑流程的第十三示例。

[0048] 图42图示了第十四示例强化的系统。

[0049] 图43图示了第五示例技术。

[0050] 图44图示了包括错误的第一示例训练/测试数据集。

[0051] 图45图示了第一示例恢复训练/测试数据。

[0052] 图46图示了第十四示例逻辑流程。

[0053] 图47图示了第十五示例强化的系统。

[0054] 图48图示了第一示例推理模型。

[0055] 图49图示了第六示例技术。

[0056] 图50图示了强化的系统的第十六示例。

[0057] 图51图示了逻辑流程的第十五示例。

[0058] 图52图示了第一示例对抗性模型训练连接。

[0059] 图53图示了第九示例推理环境。

[0060] 图54图示了第七示例技术。

[0061] 图55图示了第十示例推理环境。

[0062] 图56图示了第十一示例推理环境。

[0063] 图57图示了第八示例技术。

[0064] 图58图示了第十七示例强化的系统。

[0065] 图59图示了逻辑流程的第十六示例。

[0066] 图60图示了逻辑流程的第十七示例。

[0067] 图61图示了逻辑流程的第十八示例。

[0068] 图62图示了第十八示例强化的系统。

[0069] 图63图示了逻辑流程的第十九示例。

[0070] 图64图示了逻辑流程的第二十示例。

[0071] 图65图示了第十九示例强化的系统。

[0072] 图66图示了强化的系统的第二十示例。

[0073] 图67图示了逻辑流程的第二十一示例。

[0074] 图68A图示了第二示例对抗性模型训练连接。

[0075] 图68B图示了第三示例对抗性模型训练连接。

[0076] 图68C图示了第四示例对抗性模型训练连接。

[0077] 图69图示了示例计算可读介质。

[0078] 图70图示了示例计算系统。

[0079] 图71图示了第一示例系统。

[0080] 图72图示了第二示例系统。

[0081] 图73图示了第三示例系统。

## 具体实施方式

[0082] 本文中提供的是针对各种攻击来强化AI系统的技术和装置。如本文中所使用的，术语“强化”意指暗示用于以某种方式来减少用于攻击AI系统的区域或减少AI系统对攻击的易受性的方法、系统、装置、技术或途径。术语“强化的”并不意图暗示与攻击向量相关联的所有风险被减少、或所有攻击将被阻止、预防、检测或以其他方式被缓解。取而代之，如本文中所使用的，针对某些攻击或攻击向量而“强化的”系统意指该系统包括被设计成缓解(一个或多个)攻击的硬件、软件、或硬件和软件的组合，而不是在任何情况或情形下都缓解(一个或多个)攻击。

[0083] 如本申请中使用的术语“人工智能(AI)”意图意指被布置成基于输入进行推理的计算机实现的系统。应当领会的是，术语“AI”在不同圈子(例如，学术界、工业、消费者等)中可能意指不同的事物。本文中对术语“AI”的使用不意图是限制性的，并且意图涵盖推理引擎、专家系统、机器学习模型、神经网络、深度学习系统、模糊逻辑引擎、或可以被分类为AI的其他这种系统、模型和引擎。通常，在伞状(umbrella)AI下被分组的术语(例如，神经网络，推理引擎等)在本文中可互换地使用。

[0084] 通常，AI可以用于通过数据来进行自动学习和发现。通常，AI用于向现有产品添加

“智能”。例如,图像识别和/或分类、模式识别、语音识别、预测分析等可以与大量数据组合以通过AI来改进众多技术。

[0085] AI可以逐步从数据中进行学习。换句话说,推理引擎或“模型”被训练,并且在训练期间找到数据中的结构和规律性,使得该模型可以从未见过或未来的数据中推理出某些事物(例如,分类、预测、排名等)。存在许多类型的模型以及用以训练这些模型的技术,例如前馈神经网络(FFNN)、递归神经网络(RNN)、卷积神经网络(CNN)、长/短期记忆(LSTM)网络、贝叶斯信念网络、玻尔兹曼机器等。为了方便和清楚起见,这里不呈现对这种推理模型及其训练技术的详尽讨论。然而,可以应用本公开以实现如下设备:该设备独立于构成这种AI系统的基础的模型来强化这种AI系统。

[0086] 如所指出的,可以在许多领域中实现AI。例如,在医疗保健领域中,可以实现AI以提供个性化的诊断、X射线读数、癌症筛查等。作为另一个示例,在零售领域中,可以实现AI以提供个性化的购物推荐或者与消费者讨论购买选项。作为进一步的示例,在制造领域中,AI可以预测关于工厂的预期负载或需求,对用于取放系统(pick-and-place system)的部件进行分类等。这些只是其中可以实现AI的许多行业中的几个。

[0087] 如上所介绍,AI系统可能容易受到各种攻击。例如,可以重新创建或克隆出AI系统(例如,黑箱)。AI系统容易受到探测型攻击。去往AI系统的输入并不总是受控制的,并且因此可能是对抗性的或损害性的(compromised)。由于AI系统高度不透明,因此AI系统的操作者和用户可能由于不具有对AI系统如何运行、它是如何被训练的、训练数据集等的可见性而经受风险。这些只是攻击向量或AI系统的攻击漏洞的几个示例。

[0088] 本公开提供了用以针对一些或所有这些攻击向量来强化AI系统的多个系统和方法。要注意的是,本公开并不尝试描述所有可能的AI攻击向量,本公开也不尝试提供在本文中讨论的系统和方法与相应系统或技术可以缓解的AI攻击向量的详尽列表之间的映射。然而,提供了这种攻击向量、以及用以缓解与那些攻击向量相关联的风险的系统和方法的众多示例,以呈现适合于本领域普通技术人员理解的描述。

[0089] 本公开被松散地组织成各小节,其中这些小节可以讨论用以缓解AI系统对一种类型的攻击或相关攻击向量的漏洞的多个系统和方法。然而,要注意的是,可以应用一个小节中描述的概念来缓解沿着与示例中使用的攻击向量不同的攻击向量的攻击。具体地,在一个小节中或关于一种类型的攻击而描述的系统和方法可以用于缓解另一种类型的攻击,诸如在不同小节中描述的攻击。

[0090] 在对本文中所构想的系统和方法进行详细描述之前,提供了针对各种AI攻击向量而强化的系统的概述,如可以根据本公开实现的那样。

[0091] 利用对本文中使用的符号和术语的一般参考,可以根据在计算机或计算机网络上执行的程序过程来呈现随后的详细描述的一个或多个部分。本领域技术人员使用这些过程性描述和表示来将其工作的实质最有效地传达给本领域其他技术人员。过程通常被确信是导致期望结果的自洽的操作序列。这些操作是需要对物理量进行物理操纵的那些操作。通常,但不一定,这些量采取能够被存储、传送、组合、比较和以其他方式操纵的电气、磁性或光学信号的形式。主要出于常见用途的原因,有时将这些信号称为位、值、元素、符号、字符、项、数字等证明是方便的。然而,应当注意的是,这些和类似的术语都与适当的物理量相关联,并且仅仅是应用于那些量的方便标签。此外,描述了被布置成执行这些操纵的各种系

统、计算环境、装置、专门设计的计算设备或电路。

[0092] 现在对附图进行参考，其中相似的附图标号通篇用于指代相似的元件。在以下描述中，出于解释的目的，阐述了许多具体细节以便提供对其的透彻理解。然而，可能显而易见的是，可以在没有这些具体细节的情况下实践新颖的实施例。在其他实例中，以框图形式示出了公知的结构和设备以便于其描述。意图是覆盖权利要求范围内的所有修改、等同物和替代方案。

[0093] 图1图示了包括可以根据本公开来实现的强化的系统101的推理环境100。强化的系统101可以由任何数量或类型的计算系统(诸如，服务器、工作站、膝上型计算机、虚拟化计算系统、边缘计算设备等)来体现。附加地，强化的系统101可以是嵌入式系统，诸如深度学习加速器卡、具有深度学习加速的处理器、神经计算棒等。在一些实现方式中，强化的系统101包括片上系统(SoC)，而在其他实现方式中，强化的系统101包括具有两个或更多个分立组件的印刷电路板或芯片封装。此外，强化的系统101可以采用各种类型的“模型”中的任一种，这些“模型”被布置成基于输入来推理出某些结果、分类或特性。

[0094] 强化的系统101包括电路110和存储器120。存储器120存储推理模型122、输入数据124、输出数据126和指令128。在操作期间，电路110可以执行指令128和/或推理模型122，以从输入数据124生成输出数据126。如本文中所描述，有时可以从输入数据124来生成输出数据126。例如，输入数据124可以包括训练数据，并且电路110可以执行指令128以基于输入数据124来生成“经训练的”推理模型122作为输出数据126。例如，输入数据可以包括被标记为包括猫或不包括猫的多个图片。在这种示例中，可以使用该多个图片来生成可以推理出一图片是否包括猫的推理模型，并且可以提供该经训练的推理模型作为输出数据122。在许多这种示例中，电路110可以执行指令128和推理模型122以对输入数据124进行分类，并且提供输入数据124的分类作为输出数据126。例如，输入数据124可以包括图片，并且输出数据126可以包括该图片的分类的指示(例如，猫、狗等)。在各种这种示例中，输入数据124可以包括测试数据集(例如，图片及其分类)，并且电路110可以执行指令128以利用测试数据集来评估推理模型122的性能，并且提供该评估的指示作为输出数据126。将诸如结合下面的示例来更详细地讨论执行指令128和/或推理模型122以从输入数据124来生成输出数据126的这些和各种其他方面。

[0095] 强化的系统101还可以包括接口130。接口130可以耦合到推理环境100中的一个或多个设备，诸如强化的系统101外部的设备。例如，示出了输入设备152、输出设备154、其他计算设备156。通常，接口130可以包括被布置成耦合到互连(例如，有线、无线等)以将强化的系统101耦合到另一设备或系统的硬件接口或控制器。例如，接口130可以包括如下电路：该电路被布置成经由互连来传输和/或接收信息元素，并且在也耦合到该互连的其他设备(例如，输入设备152、输出设备154、其他计算设备156等)之间传送和/或接收信息元素(例如，包括数据、控制信号等)。在一些示例中，接口130可以被布置成耦合到符合多种标准中的任一个的接口。在一些示例中，接口130可以被布置成耦合到以太网互连、蜂窝互连、通用串行总线(USB)互连、外围组件互连(PCI)等。在一些示例中，推理系统100可以包括多个接口，例如以通过不同的互连来耦合到不同的设备。

[0096] 通常，输入设备152可以是被布置成将信号作为输入提供给强化的系统101的任何设备。在一些示例的情况下，输入设备152可以是任何数量和类型的设备或传感器。在操作

期间，电路110可以执行指令128以经由接口130从这些传感器接收信号。电路110在执行指令128时可以将接收到的信号存储为输入数据124。可替代地，电路110在执行指令128时可以基于信号来生成输入数据124(例如，通过将一些处理应用于经由接口130从传感器接收到的原始信号)。作为另一个示例，电路110可以执行指令128以从其他计算设备156接收信息元素，该信息元素包括输入数据124的指示。在一些示例的情况下，输入设备152、输出设备154、和/或其他计算设备156中的任何一个或多个可以与强化的系统101封装在一起。示例在该情境中不受限制。

[0097] 强化的系统101进一步包括强化器140。通常，强化器140可以包括被布置成缓解可用于以强化的系统101为目标的一个或多个攻击向量或者针对这些攻击向量进行“强化”的硬件、软件和/或其组合。在下面的示例中给出了强化器140的众多示例。然而，这里要注意的是，强化器140被描绘成体现为强化的系统101的独立硬件、软件、和/或硬件软件元件组合。可替代地，强化器140可以体现在强化的系统101或推理环境100的任何一个或多个元件内。例如，电路110、存储器120、接口130、推理模型122、输入数据124、输出数据126和/或指令128都可以包括强化器140。作为进一步的示例，强化器140可以涵盖强化的系统101，或者可以与强化的系统101分离并且不同。例如，强化器可以耦合到接口140，或者耦合到另一设备(诸如，输入设备152)的一部分。

[0098] 如上所介绍，本公开提供了被布置成缓解或减少针对采用AI的系统(诸如，包括推理模型122的系统101)的攻击向量的装置和方法。强化器140因此被布置并且定位在推理环境100内以缓解或减少这些攻击向量。例如，强化器140可以被布置成缓解通常被称为模型盗窃(model theft)的攻击向量。模型盗窃是对推理模型122本身的盗窃(例如，提取、克隆、逆向工程、模型反演等)。强化器140可以被布置成缓解模型克隆、所提取模型的操作、模型反演或逆向工程。例如，强化器140可以被布置成缓解试图检索关于训练数据的信息的探测攻击。作为另一个示例，强化器140可以被布置成缓解通常被称为对抗性输入的攻击向量。对抗性输入是被设计成“欺骗”推理系统或基于输入而触发特定推理的输入。强化器140可以被布置成缓解输入探测攻击、输入失真、输入编辑等。在一些示例的情况下，强化器140可以被布置成增加输入数据信任度、有效性或相关性(例如，基于一个或多个输入设备特性)。强化器140可以被布置成净化(sanitize)输入数据或对输入设备进行认证。作为另一个示例，强化器140可以被布置成缓解隐私泄露或暴露。例如，应当领会的是，对抗方可以使用所有输出(例如，正确的推理、错误的推理、高概率的推理、低概率的推理、分类器的数量等)来收集关于该模型和/或训练数据的有用信息。下面详细描述了针对这种攻击以及其他类型的攻击而强化的系统的众多示例。

[0099] 通常，电路110表示硬件，诸如常规的中央处理单元(CPU)、专用集成电路(ASIC)、现场可编程门阵列(FPGA)或其他逻辑。例如，电路110可以实现图形处理单元(GPU)或加速器逻辑。作为特定示例，电路110可以是谷歌®的Tensor处理单元®、英特尔®的Nervana NNP®(神经网络处理器)、Movidius的Myriad 2®处理器、Mobileeye®的EyeQ®处理器。在一些示例中，电路110可以是具有多个核的处理器，其中一个或多个核被布置成处理AI指令。例如，高通®的骁龙845®处理器或苹果®的A11/A12 Bionic SoC®。提供这些示例是出于清楚和方便的目的而不是为了进行限制。

[0100] 电路110可以包括指令集(未示出)，或者可以符合任何数量的指令集架构，诸如例

如x86架构或ARM架构。该指令集可以是32位指令集、64位指令集等。附加地，指令集可以使用低精度算术，诸如半精度、bflaot16浮点格式等。示例在该情境中不受限制。

[0101] 存储器120可以基于各种各样的信息存储技术中的任一个。例如，存储器120可以基于需要不间断地提供电力的易失性技术、或不需要不间断地提供电力的非易失性技术，并且可能包括需要使用可能是或可能不是可移除的机器可读存储介质的技术。因此，这些存储装置中的每一个可以包括各种各样类型(或类型组合)的存储设备中的任一个，包括但不限于只读存储器(ROM)、随机存取存储器(RAM)、动态RAM(DRAM))、双数据速率DRAM(DDR-DRAM)、同步DRAM(SDRAM)、静态RAM(SRAM)、可编程ROM(PROM)、可擦除可编程ROM(EPROM)、电可擦除可编程ROM(EEPROM)、闪速存储器、聚合物存储器(例如，铁电聚合物存储器)、双向存储器(ovonic memory)、相变或铁电存储器、硅氧化氮氧化硅(SONOS)存储器、磁卡或光卡、一个或多个个体铁磁磁盘驱动器、或被组织成一个或多个阵列的多个存储设备(例如，被组织成独立磁盘阵列冗余阵列或RAID阵列的多个铁磁磁盘驱动器)。

[0102] 图2图示了可以在推理环境中实现的示例技术200，该技术诸如由上述推理环境100的强化的系统101来实现。要注意的是，为了方便和清楚起见，参考图1的推理环境100描述了技术200。然而，这并不意图是限制性的。

[0103] 技术200可以在圆圈2.1处开始。在圆圈2.1处，输入设备152和/或其他计算设备156可以将输入数据124传送到强化的系统101。在圆圈2.2处，强化的系统101可以接收输入数据124。例如，可以经由接口130来传送输入数据124。在一些示例的情况下，强化器140可以耦合到输入设备152、其他计算设备156或推理模型122，并且被布置成针对攻击向量来强化推理模型122。例如，强化器140可以被布置成净化输入数据124。作为另一个示例，强化器140可以被布置成验证输入设备152。作为进一步的示例，强化器140可以被布置成对输入数据124进行变换。下面将更详细地描述这些和其他示例。

[0104] 技术200可以继续到圆圈2.3。在圆圈2.3处，推理模型122可以在输入数据124下执行或操作，以做出一些推理或生成输出数据126。在一些示例的情况下，强化器140可以耦合到推理模型122。例如，推理模型122可以包括被布置成标识探测攻击的多个隐藏类。作为另一个示例，推理模型122可以被加密，或者可以包括被布置成在加密数据下进行操作的激活函数。作为另一个示例，推理模型122可以被分割或分区，以使得不同的计算实体可以执行推理模型122的不同部分。仍然作为进一步的示例，推理模型122可以被布置有多个隐藏类，以提供用于数字权利管理目的的模型标识，提供探测攻击的指示等。下面将更详细地描述这些和其他示例。

[0105] 技术200可以继续到圆圈2.4。在圆圈2.4处，强化的系统101可以输出数据126，例如强化的系统100可以将输出数据126传送到输出设备154(和/或其他计算设备156)。在一些示例中，输出数据126可以包括关于输入数据124的推理，诸如输入数据124的分类。在圆圈2.5处，输出设备154可以接收输出数据126。例如，可以经由接口130来传送输出数据126。在一些示例的情况下，强化器140可以耦合到强化的系统101的输出或者耦合到输出设备154，以针对攻击向量进行强化。例如，强化器140可以被布置成修剪(prune)输出数据124。下面将更详细地描述这些和其他示例。

[0106] 图3图示了可以在推理环境中实现的示例训练/测试技术300，该技术例如由推理环境100的强化的系统101来实现以训练推理模型(例如，推理模型122等)。要注意的是，为

了方便和清楚起见，参考图1的推理环境100描述了技术300。然而，这并不意图是限制性的。通常，推理模型是通过重复过程来训练的。本文中给出了推理模型训练的一些示例。然而，要注意的是，本文中提供的许多示例可以被实现为独立于用于训练推理模型的算法来针对攻击强化推理环境。

[0107] 技术300可以在圆圈3.1处开始。在圆圈3.1处，输入设备152和/或其他计算设备156可以将训练数据集362传送到强化的系统101。在圆圈3.2处，强化的系统101可以接收训练数据集362。例如，可以经由接口130来传送训练数据集362。在一些示例的情况下，强化器140可以耦合到输入设备152、其他计算设备156或推理模型122，并且被布置成针对攻击向量进行强化。例如，强化器140可以被布置成添加对抗性攻击输入和相关联的输出，使得可以将推理模型122训练成在操作期间标识对抗性攻击。下面将更详细地描述该示例和其他示例。

[0108] 技术300可以继续到圆圈3.3。在圆圈3.3处，推理模型122可以在来自训练数据集362的输入数据124下执行或操作，以做出一些推理或生成输出数据。在圆圈3.4处，可以基于在圆圈3.3处从训练数据集362的输入数据所生成的输出以及训练数据集362的对应期望输出来调整推理模型122。例如，可以在圆圈3.4处调整推理模型的权重、连接、层等。通常，交互式地重复圆圈3.3和3.4，直到推理模型122在可接受的成功率下收敛(通常被称为达到最小错误条件)为止。可以基于与阈值的比较(例如，大于阈值、大于或等于阈值等)来确定推理模型122是否关于可接受的成功率而收敛。在一些示例中，可以提供经训练的推理模型作为输出数据126。在一些示例的情况下，强化器140可以耦合到推理模型122。例如，推理模型122可以包括回滚(roll-back)功能，以基于将训练数据集362感染的恶意输入数据来撤销在圆圈3.4处做出的调整。下面将更详细地描述该示例和其他示例。

[0109] 一旦推理模型122被训练，就可以测试推理模型122。在圆圈3.5处，输入设备152和/或其他计算设备156可以将测试数据集364传送到强化的系统101。在圆圈3.6处，强化的系统101可以接收测试数据集364。例如，可以经由接口130来传送测试数据集364。在圆圈3.7处，推理模型122可以在测试数据集364的输入数据下执行或操作，以做出一些推理或生成输出数据126。在圆圈3.8处，强化的系统101可以将输出数据126例如传送到输出设备154(和/或其他计算设备156)。在圆圈3.9处，输出设备154可以接收输出数据126，并且可以将输出数据126与来自测试数据集324的期望输出数据进行比较，以确定推理模型122推理或生成正确输出的良好程度。在一些示例中，可以提供测试的结果作为输出数据126。要注意的是，在推理模型的训练和测试期间，输入设备152和输出设备154通常可以是相同的。在一些示例的情况下，强化器140可以耦合到推理模型122。例如，推理模型122可以用可信的方式与训练数据集362相关联，以便稍后证实利用训练数据集362对推理模型122的训练。下面将更详细地描述该示例和其他示例。

[0110] 现在的描述转向提供强化的推理系统的众多示例、以及可以被这种系统缓解的攻击向量的具体示例。基于相关的攻击向量，示例被松散地组织成多个小节。在小节I中，将可以被实现以缓解与模型盗窃(或IP盗窃)有关的攻击向量的示例分组在一起。在小节II中，将可以被实现以缓解与对抗性输入有关的攻击向量的示例分组在一起。在小节III中，将可以被实现以提供数据信任度的示例分组在一起。在小节IV中，将可以被实现以提供增强的推理模型训练的示例分组在一起。在小节V中，将可以被实现以针对推理模型提供增强的隐

私特征的示例分组在一起。最后，小节VI提供了可以被实现以缓解一般的攻击向量的示例。然而，要注意的是，尽管可以在与攻击向量的类型或分组有关的一个小节中描述示例，但是示例可以被实现以缓解在其他小节中讨论的其他攻击向量。此外，可以将多个示例组合以提供强化的系统，以缓解多个类型的攻击向量。示例在该情境中不受限制。

[0111] *小节I-推理模型盗窃。*

[0112] *用以缓解经由探测进行的模型重新创建的输出修整(Output Trimming)*

图4图示了可以根据本公开来实现的示例强化的系统401。强化的系统401包括电路410、存储器420和输入设备452。存储器420存储推理模型422、输入数据424、输出数据426、指令428、推理概率423、阈值425和空输出数据427。在操作期间，电路410可以执行指令428和/或推理模型422，以从输入数据424生成输出数据426。

[0113] 针对推理系统的常见攻击是对推理系统的重复查询，通常被称为“探测”。在探测期间，攻击者运行针对该系统的多个查询，以尝试学习系统行为、诸如例如分类类别。作为特定示例，可以重复地查询被布置成对图像进行分类的推理系统，以标识该推理系统被训练成利用其对图像进行分类的类别的类型。本公开提供了强化的系统401，其被布置成在推理概率低于阈值水平的情况下修整或掩盖(mask)输出。

[0114] 例如，在操作期间，电路410可以执行指令428以接收输入数据424。电路410可以执行推理模型422，以从输入数据424生成输出数据426。电路410可以执行指令428，以确定输出426是否具有高于阈值425的推理概率423。换句话说，电路410可以执行指令428，以确定输出426正确的概率是否大于阈值水平。通常，阈值425可以是足够大以阻止探测的任何值，例如80％正确、85％正确、90％正确、95％正确、99％正确等。在一些示例的情况下，可以动态地计算阈值425(例如，基于查询来源、攻击概率、查询频率、来自相同来源或相似来源的查询频率等)。在其他示例中，可以例如在模型创建时、在由模型所有者或授权管理员进行操作期间或之前静态地设置阈值425。

[0115] 电路410可以执行指令428，以在所确定的推理概率423大于阈值425的情况下提供输出数据426，并且在所确定的推理概率423大于阈值425的情况下提供空输出数据427。在一些示例的情况下，空输出数据427可以是“无”(例如，在分类系统的情况下)。

[0116] 图5图示了逻辑流程500。逻辑流程500可以表示由本文中描述的一个或多个设备执行的一些或全部操作。例如，强化的系统401(或其组件)可以执行逻辑流程500中的操作以修整来自推理系统的输出以阻止探测。

[0117] 逻辑流程500可以在框510“接收输入数据”处开始，系统401可以接收输入数据424。例如，强化的系统401可以从输入设备452接收输入数据424。继续到框520“经由推理模型、基于接收到的输入来生成输出”，系统401可以经由推理模型422从输入数据424生成输出数据426。例如，电路410可以执行推理模型422，以从输入数据424生成输出数据426。继续到框530“确定所生成的输出的推理概率”，系统401可以确定所生成的输出的推理概率。例如，电路410可以执行指令428，以确定输出数据426的推理概率423。

[0118] 继续到判定框540“推理概率是否大于阈值”，系统401可以确定在框530处确定的推理概率是否大于阈值水平。例如，电路410可以执行指令428，以确定推理概率423是否大于阈值425。从框540，逻辑流程500可以继续到框550或560。逻辑流程500可以基于推理概率423大于阈值425的确定而从框540继续到框550，而逻辑流程500可以基于推理概率423不大

于阈值425的确定而从框540继续到框560。

[0119] 在框550“提供所生成的输出”处，系统401可以提供在框520处生成的输出。例如，电路410可以执行指令428，以提供输出数据426(例如，提供给输出设备、另一计算设备等)。在框560“提供空输出”处，系统401可以提供空输出数据。例如，电路410可以执行指令428，以提供空输出数据427(例如，提供给输出设备、另一计算设备等)。

[0120] *用于模型数字权利管理的安全标记*

图6图示了可以根据本公开来实现的示例强化的系统601。强化的系统601包括电路610、存储器620、接口630、输入设备652和可信执行环境(TEE)670。存储器620存储推理模型622、输入数据624、输出数据626和指令628。TEE 670存储标记672和指令674。通常，TEE 670可以是系统601内的任何可信执行飞地(enclave)或分区。例如，TEE 670可以是电路610内的可信执行分区。作为另一个示例，TEE 670可以包括以可信的方式从电路610和存储器620中分割的其自己的电路和存储器(未示出)。通常，TEE 670可以提供安全性特征，诸如隔离执行(例如，对指令674的隔离执行)以及标记672的保密性(confidentiality)。

[0121] 例如，TEE 670可以包括被安全地布置成执行指令以实现本文中描述的功能的电路和/或存储装置。如本文中使用的TEE可以包括任何数量的安全执行分区(例如，软件防护扩展(SGX)、英特尔®可管理性引擎(IME)、创新引擎(IE)、苹果®的安全飞地、信任区等)。要注意的是，TEE可以执行存储在存储器(作为TEE的一部分的安全存储器、或其他存储器)内的指令。这些指令可以包括可在TEE的电路或(一个或多个)处理组件上操作以使得TEE执行本文中描述的技术的计算机可执行指令。重要的是要注意，TEE可以被合并到更大的电路(例如，处理器、SoC等)中，并且作为本文中描述的设备的一部分而提供。

[0122] 应当领会的是，经训练的推理模型(例如，推理模型622)是高度有价值的知识产权资产，其可能会被对抗性第三方所窃取，并且以不可控制的方式用于其他产品(诸如，竞争者的产品)中。因此，检测对推理模型的不适当使用的能力是合期望的。本文中详细描述的一些方法可以使用所谓的“后门类(backdoor class)”来支持对模型使用的检测。例如，推理模型是利用隐藏类来训练的，当适当的查询被发送到该系统时，该隐藏类可以被推理或输出。推理模型的所有者可以使用该查询和对应的隐藏类输出来检测未经授权的使用。然而，使用像这样的隐藏类仍然易于产生原生的不准确性(native inaccuracy)，使得在某些情况下，隐藏类可能会被错误分类。此外，将隐藏类添加到推理系统(例如，利用隐藏类来训练推理系统)可能会导致性能退化以及训练过程的更高成本。附加地，该方法是不可扩展的，这是因为添加新的隐藏类将减少模型容量。

[0123] 本公开提供了具有标记672的系统601，该标记672可以用于与传入的查询匹配以确定未经授权的使用。要注意的是，标记672例如经由TEE 670被加密并且以安全的方式被存储。因此，标记672可以用作水印或数字权利管理令牌。在操作期间，TEE 670可以执行指令674，以将输入数据624与标记672进行匹配。在一些示例的情况下，TEE 670内的指令674可以由逐位比较器来实现。由此，在操作期间，TEE 670可以应用标记672与输入数据624的逐位比较，以确定输入数据624是否与标记672匹配。在一些示例中，TEE 670可以执行指令674，以(例如，经由逐位比较等)确定输入数据624的部分是否与标记672匹配。作为特定示例，TEE 670在执行指令674时可以将输入数据624的所选位(例如，位x至y，其中x和y是整数，并且x大于或等于1且y大于x)与标记672进行比较。也就是说，可以将“标记”插入输入数

据624内的任何地方。因此，在操作期间，TEE 670可以被配置成寻找输入数据624内的该特定位置(例如，位的特定范围等)，并且确定该位置是否与标记672匹配。

[0124] 在未确定输入数据624与标记672之间的匹配的情况下，TEE 670在执行指令674时可以(例如，经由输出数据626等)输出未经授权的推理模型622使用的指示。在找到了匹配的情况下，可以照常使用推理模型622。重要的是要注意，系统601是鲁棒的且可扩展的。此外，可以在未来修改系统601(例如，通过更新标记674等)。在一些示例中，标记674可以被预加载(例如，制造等)。标记674可以例如存储在TEE 670中的可擦除只读存储器(e-prom)中，或者通过一次性熔丝(fuse)来编码。

[0125] 图7图示了逻辑流程700。逻辑流程700可以表示由本文中描述的一个或多个设备执行的一些或全部操作。例如，强化的系统601(或其组件)可以执行逻辑流程700中的操作，以检测对推理模型的未经授权使用。

[0126] 逻辑流程700可以在框710“接收输入数据”处开始，系统601可以接收输入数据624。例如，强化的系统601可以从输入设备652接收输入数据624。继续到框720“将输入数据与标记进行匹配”，系统601可以将接收到的输入数据与标记进行匹配。例如，TEE 670可以执行指令674，以将输入数据624与标记672进行匹配。

[0127] 继续到判定框730“标记是否与输入匹配”，系统601可以确定标记是否与输入匹配。例如，电路610可以执行指令628，以确定TEE 670是否指示标记674与输入数据624匹配。从框730，逻辑流程700可以继续到框740或750。逻辑流程700可以基于标记674与输入数据624不匹配的确定而从框730继续到框740，而逻辑流程700可以基于标记674与输入数据624匹配的确定而从框730继续到框750。

[0128] 在框740“经由推理模型、基于接收到的输入来生成输出”处，系统601可以经由推理模型622从输入数据624生成输出数据626。例如，电路610可以执行推理模型622，以从输入数据624生成输出数据626。在框750“输出未经授权使用的指示”处，系统601可以输出对推理模型622的未经授权使用的指示。换句话说，系统601可以确定接收到的输入数据624的提供者是否对推理模型622具有适当的所有权权利。

[0129] 在一些示例的情况下，可以将输入数据与多个标记进行比较，以确定针对一个或多个推理模型或推理模型特征的所有权、有效许可、或经授权的使用权利。图8图示了可以根据本公开来实现的示例强化的系统801。强化的系统801包括电路810、存储器820、接口830、输入设备852和可信执行环境(TEE)870。存储器820存储推理模型822-n、输入数据824、输出数据826和指令828。TEE 870存储标记872-n和指令874。

[0130] 通常，系统801可以被布置成基于与标记872-n之一匹配的输入数据824来启用推理模型822-n之一。例如，该图描绘了包括推理模型822-1、822-2和822-3的存储器820，而TEE 870存储标记872-1、872-2和872-3。在操作期间，TEE 870可以执行指令874，以将输入数据824与标记872-1、872-2或872-3之一进行匹配。在一些示例的情况下，TEE 870内的指令874可以由逐位比较器来实现。由此，在操作期间，TEE 870可以应用标记872-1、872-2或872-3与输入数据824的的逐位比较，以确定输入数据824是否与标记872-1、872-2或872-3之一匹配。

[0131] 在确定输入数据824与标记872-1、872-2或872-3之一之间的匹配时，可以假定输入数据提供者具有对于对应推理模型的权利(例如，所有权权利、许可权利等)。因此，TEE

870在执行指令874时可以基于输入数据824与标记872-1、872-2或872-3之一匹配的确定来启用模型822-1、822-2或822-3中的对应的一个。

[0132] 在一些示例中，输入数据可以被加密，以缓解输入数据被尝试伪造标记的第三方所窃取和/或重新创建。图9图示了可以根据本公开来实现的示例强化的系统901。强化的系统901包括电路910、存储器920、接口930、输入设备952和可信执行环境(TEE)970。存储器920存储推理模型922、加密的输入数据924、输出数据926和指令928。TEE 970存储标记972、指令974、解密的输入数据975和密钥976。

[0133] 通常，系统901可以被布置成类似于系统401，用以基于与标记972匹配的输入数据924来检查对使用推理模型922的所有权或授权。然而，强化的系统901进一步包括：对存储在存储器920中的输入进行加密，以防止输入数据被敌对方用于重新创建或伪造标记972。

[0134] 在操作期间，TEE 970可以执行指令974，以使用密钥976来对加密的输入数据924进行解密，并且将解密的输入数据975与标记972进行匹配。

[0135] *经由对抗性水印的窃取模型检测*

如所指出的，经训练的推理模型是高度有价值的知识产权资产，其可能会被对抗性第三方所窃取，并且以不可控制的方式用于其他产品(诸如，竞争者的产品)中。由此，检测对推理模型的不适当使用的能力是合期望的。在本公开的一些示例中，可以提供被训练成对一些对抗性输入进行分类的推理模型，以便于未经授权的模型使用。

[0136] 应当领会的是，对抗性输入是极其接近于校正后的分类但是仍不正确的输入。例如，去往被设计成对刀进行分类的分类系统的对抗性输入可能会将伪装成钢笔的刀分类为钢笔而不是刀。本公开提供了被训练成对多个对抗性输入进行分类的推理模型。因此，推理模型可以在针对其来训练该模型的对抗性输入集上执行，从而导致对抗性输入的正确分类。因此，对抗性输入集可以便于确定模型使用(例如，以标识未经授权的使用)。将对抗性输入集提供给不同的模型(例如，未在对抗性输入集上训练的模型)应当会导致错误的分类(例如，将刀分类为钢笔等)，从而指示该模型是不同的。由于对抗性输入的可传递性的特性，因此对该模型的分类的任何随机更改都应当足以在普通输入集上创建不可否认性(non-repudiation)，而不会更改多于几个百分比的对抗性输入的结果。因此，即使该模型将被对抗方修改，也将难以移除本文中详细描述的对抗性输入分类检测特征。由此，将难以隐藏该模型已经被复制的事实。

[0137] 图10A图示了用以训练推理模型的示例系统1001。系统1001包括电路1010和存储器1020。存储器1020存储推理模型1022、指令1028、测试/训练数据集1060和输出数据1026。测试/训练数据集1060包括测试数据1062以及训练数据1064两者。测试数据1062和训练数据1064可以用于例如使用图3的技术300来“训练”推理模型1022。

[0138] 测试/训练数据集1060的示例在图10B中被描绘。简要地转向该图，测试/训练数据集1060包括输入1061和对抗性输入1063。针对每个输入1061和对抗性输入1063，测试/训练数据集1060还包括预期输出1065。通常，输入1061可以是要利用其来训练(或测试)该模型的普通输入，而对抗性输入可以是要针对其来训练(或测试)该模型以提供模型盗窃检测特征的对抗性输入，如本文中所描述的那样。输入1061可以包括(一个或多个)输入值1061-n，其中n是正整数，而对抗性输入1063可以包括(一个或多个)对抗性输入值1063-m，其中m是正整数。要注意的是，在实践期间，输入1061和对抗性输入1063的数量可以很大。然而，该图

出于呈现清楚的目的描绘了多个输入。例如，针对输入1061，示出了(一个或多个)输入值1061-1、(一个或多个)输入值1061-2、(一个或多个)输入值1061-3、(一个或多个)输入值1061-4、(一个或多个)输入值1061-5至(一个或多个)输入值1061-N。同样地，针对对抗性输入1063，示出了(一个或多个)对抗性输入值1063-1、(一个或多个)对抗性输入值1063-2、(一个或多个)对抗性输入值1063-3至(一个或多个)对抗性输入值1063-M。

[0139] 针对每个输入值1061-n和对抗性输入值1063-m，示出了相关联的预期输出1065。每个预期输出可以对应于推理模型1022要被训练成向其指派输入的推理(例如，分类等)。例如，预期输出1065包括输入将被推理模型1022分类到的类1067的指示。例如，描绘了类1067-1、类1067-2和类1067-3。要注意的是，推理模型可以被训练成将输入分类到任何数量的类中。然而，出于呈现清楚的目的，选择此处描绘的数量。

[0140] 电路1010可以执行指令1028，以例如使用测试/训练数据集1060来训练推理模型1022，如参考图3所描述的那样。因此，在操作期间，经训练的推理模型1022可以用于：通过将对抗性输入1063提供给推理模型并且将输出1026与对应的预期输出1065进行比较来确定未经授权的使用(例如，指示模型盗窃等)。针对对抗性输入1063的正确分类可以指示该推理模型与推理模型1022相同。例如，将对抗性输入1063-M分类到类1067-2可以指示向其提供了对抗性输入1063-M的推理模型与推理模型1022相同。

[0141] 图11图示了可以根据本公开来实现的示例强化的系统1101。强化的系统1101包括电路1110、存储器1120和输入设备1152。存储器1120存储推理模型1122、输入数据1124、输出数据1126和指令1128。在一些示例中，存储器1120还可以包括分类阈值1123。在操作期间，电路1110可以执行指令1128和/或推理模型1122，以从输入数据1124生成输出数据1126。推理模型1122可以类似于推理模型1022，推理模型1022是利用多个对抗性输入(例如，对抗性输入1063-m等)来训练的。由此，将对抗性输入1063-m提供给推理模型1122并且将输出数据1126与对抗性输入1063-m的预期输出1065进行比较可以便于对模型盗窃的标识。

[0142] 例如，在操作期间，电路1110可以执行指令1128以接收输入数据1152。电路1110可以执行推理模型1122，以从输入数据1124生成输出数据1126。在输入数据1152对应于对抗性输入1063-m的情况下，电路1110可以执行指令1128以确定输出1126是否对应于对抗性输入1063-m的预期输出1065。换句话说，电路1110可以执行指令1128，以确定推理模型1122是否将输入数据1124(例如，对抗性输入1063-m)分类到正确的类1067中。在推理模型1122将对抗性输入正确地分类的情况下，可以假定模型1122是模型1022。

[0143] 在一些示例的情况下，在推理模型1122将对抗性输入正确地分类超过分类阈值1123的次数的情况下，可以假定推理模型1122被非法地使用。在一些示例的情况下，可以对推理模型1122将对抗性输入分类的次数进行计数，并且在滚动基础(rolling basis)上(例如，基于时间的、所呈现的对抗性输入的数量、输入的总数等)将该次数与分类阈值1123进行比较。

[0144] 图12图示了逻辑流程1200。逻辑流程1200可以表示由本文中描述的一个或多个设备执行的一些或全部操作。例如，强化的系统1101(或其组件)可以执行逻辑流程1200中的操作，以检测对模型的未经授权使用(例如，模型盗窃)。

[0145] 逻辑流程1200可以在框1210“接收输入数据”处开始，系统1101可以接收输入数据

1124。例如，强化的系统1101可以从输入设备1152接收输入数据1124。在一些情况下，输入数据可以对应于图10A和10B的测试/训练数据集1060的对抗性输入1063-m中的任一个。

[0146] 继续到框1220“经由推理模型、基于接收到的输入来生成输出”，系统1101可以经由推理模型1122从输入数据1124生成输出数据1126。例如，电路1110可以执行推理模型1122，以从输入数据1124生成输出数据1126。继续到判定框1230“输入数据是否是对抗性的”，系统1101可以确定输入数据1124是否是对抗性的(例如，输入是否是输入1063-m等)。例如，电路1110可以执行指令1128以确定输入1124是否是对抗性的。从判定框1230，逻辑流程1200可以继续到判定框1240或框1250。逻辑流程1200可以基于该输入不是对抗性的确定而从判定框1230继续到判定框1240，而逻辑流程1200可以基于该输入是对抗性的确定而从判定框1230继续到判定框1240。

[0147] 在判定框1240“输入数据是否被正确地分类”处，系统1101可以确定输入1124是否被正确地分类。例如，电路1110可以执行指令1128，以确定由推理模型1122从输入数据1124生成的输出数据1126是否将输入数据1124与其相对应的对抗性输入正确地分类。作为特定示例，如果对抗性输入是伪装成钢笔的刀，则在判定框1240处，可以确定推理模型1122是否将刀正确地分类为刀，而不是错误地将刀分类为钢笔。从判定框1240，逻辑流程1200可以继续到框1250或框1260。逻辑流程1200可以基于输入数据未被正确地分类的确定而从判定框1240继续到框1250，而逻辑流程1200可以基于输入数据被正确地分类的确定而从判定框1240继续到框1260。

[0148] 在框1250“提供所生成的输出”处，系统1101可以提供从推理模型1122和输入数据1124生成的输出。例如，电路1110可以执行推理模型1122，以提供输出数据1126(例如，提供给主机系统、提供给经由互连而耦合的系统等)。

[0149] 在框1260“提供模型使用未经授权的指示”处，系统1101可以提供模型使用未经授权(例如，该模型被窃取等)的指示。例如，电路1110在执行指令1128时可以将信息元素发送到报告机构(或类似机构)，该信息元素包括模型1122可能与模型1022相同的指示(例如，指示模型1122被窃取等)。从框1260，逻辑流程1200可以继续到框1250。

[0150] 在一些示例中，可以在分类阈值1123被超过的情况下提供未经授权的模型使用的指示。也就是说，框1260本身可以对应于确定分类阈值1123是否已经被超过的逻辑流程。图13图示了逻辑流程1300。逻辑流程1300可以表示由本文中描述的一个或多个设备执行的一些或全部操作。例如，强化的系统1101(或其组件)可以执行逻辑流程1300中的操作，以确定是否输出对模型的未经授权使用(例如，模型盗窃)的指示，诸如在图12的框1260处那样。

[0151] 逻辑流程1300可以在框1310“使计数器递增”处开始，系统1101可以使计数器递增。继续到判定框1320“分类阈值是否被超过”，系统1101可以(例如，通过计数器等)确定分类阈值1123是否已经被超过。从判定框1320，逻辑流程1300可以继续到判定框1330或判定框1340。逻辑流程1300可以基于分类阈值1123已经被超过的确定而从判定框1320继续到框1330，而逻辑流程1300可以基于分类阈值1123尚未被超过的确定而从判定框1320继续到框1340。

[0152] 在框1330“提供模型使用未经授权的指示”处，系统1101可以提供模型使用未经授权(例如，该模型被窃取等)的指示。例如，电路1110在执行指令1128时可以将信息元素发送到报告机构(或类似机构)，该信息元素包括模型1122可能与模型1022相同的指示(例如，指

示模型1122被窃取等)。

[0153] 在判定框1340“是否需要重置计数器”处,系统1101可以确定是否需要重置计数器(例如,从框1310开始)。例如,在一些实现方式中,可以在一时间段之后重置计数器。在其他示例中,计数器可以在一时间段之后递减。在其他示例中,计数器可以在处理了多个输入之后递减。

[0154] 从判定框1340,逻辑流程1300可以继续到框1350或结束。逻辑流程1300可以基于需要重置计数器的确定而从判定框1340继续到框1350,而逻辑流程1300可以基于不需要重置计数器的确定而从判定框1340继续到结束。

[0155] 在框1350“重置计数器”处,系统1101可以重置计数器。例如,计数器可以递减、重置到零等。

[0156] *用以防止模型盗窃的硬件相关输入变换*

图14图示了可以根据本公开来实现的强化的系统1401的示例。强化的系统1401包括电路1410、存储器1420、接口1430、输入设备1452和可信执行环境(TEE)1470。存储器1420存储经变换的输入数据1425、输出数据1426、平台配置1427和指令1428。TEE 1470存储推理模型1422、预加载的平台配置1472、恢复的输入数据1424和TEE指令1478。通常,TEE 1470可以是系统1401内的任何可信执行飞地(enclave)或分区。例如,TEE 1470可以是电路1410内的可信执行分区。作为另一个示例,TEE 1470可以包括以可信的方式从电路1410和存储器1420中分割的其自己的电路和存储器(未示出)。通常,TEE 1470可以提供安全性特征,诸如隔离执行(例如,对TEE指令1478、对推理模型1422的隔离执行)、以及预加载的平台配置1472的保密性。

[0157] 如所指出的,经训练的推理模型(例如,推理模型1422)是高度有价值的知识产权资产,其可能会被对抗性第三方所窃取并且以不可控制的方式用于其他产品(诸如,竞争者的产品)中。由此,检测对推理模型的不适当使用或未经授权使用的能力是合期望的。

[0158] 本公开提供了系统1401,其被布置成基于计算平台(例如,在其上收集输入数据的平台、在其上执行推理模型的平台等)的特性来对输入数据进行变换(例如,编码)。在将输入提供给推理模型之前,系统1401被布置成基于预先配置的平台特性来对输入数据进行恢复(例如,解码)。要注意的是,输入恢复基于预加载的平台配置1472,该配置以保密的方式被维持在TEE 1470中。由此,在其中该模型在平台上执行、而该平台并没有被授权这么做的实例中,预先配置的平台特性将无法正确地将输入数据解码,并且输入将无法用作针对推理模型的输入。也就是说,输出将是没有意义的。因此,可以阻止对模型的未经授权使用。

[0159] 在操作期间,系统1401的电路1410可以执行指令1428以接收输入数据1424。例如,电路1410可以从输入设备1452接收输入数据1424。作为另一个示例,电路1410可以经由接口1430从另一计算设备接收输入数据1424。电路1410在执行指令1428时可以生成平台配置1427。在一些示例中,平台配置1427可以是系统1401的唯一标识特征。例如,平台配置1427可以包括与系统1401相关联的通用唯一标识符(UUID)的指示。作为另一个示例,平台配置1427可以包括针对电路1410、存储器1420、接口1430、输入设备1452等的唯一标识符的指示。在一些示例中,平台配置1427可以包括标识符组合的指示(例如,用于电路1410和用于输入设备1452的UUID等等)。

[0160] 电路1410在执行指令1428时可以从输入数据1424和平台配置1427生成经变换的

输入数据1425。例如，电路1410在执行指令1428时可以使用平台配置1427对输入数据1424进行编码、加密、混淆等。作为简单的示例，电路1410可以使用在平台配置1427中所指示的UUID作为加密密钥来对输入数据1425进行加密。可以实现多种编码方案、加密方案、特别是无损方案中的任一个。

[0161] TEE 1470在执行TEE指令1478时可以从经变换的输入数据1425和预加载的平台配置1472中生成恢复的输入数据1474。在一些示例中，预加载的平台配置1472可以包括系统(例如，系统1401等)的唯一标识特征的指示。TEE 1470可以(例如，在提供1401的时间处，等等)被配置有预加载的平台配置1472。因此，在操作期间，TEE 170可以从预加载的平台配置1472和经变换的输入数据1425中生成恢复的输入数据1474。作为简单的示例，TEE 170可以使用预加载的平台配置1472作为解密密钥来对经变换的输入数据1425进行解密。假定该系统正在其上操作的平台与预加载的平台配置数据匹配，则恢复的输入数据1474应当与输入数据1424匹配。换句话说，在系统1401具有与预加载的平台配置1472匹配的配置信息(例如，UUID等)的情况下，对输入数据1424的变换(例如，加密)和恢复(例如，解密)将导致使输入数据1424返回到其原始状态。

[0162] TEE 1470在执行指令1478时可以执行推理模型1422，以从恢复的输入数据1474生成输出数据1426。如上所指出，在平台配置1427与预加载的平台配置1472不匹配的情况下，恢复的输入数据1474将不对应于输入数据1424，并且输出数据1426将是没有意义的。因此，将阻止对推理模型1422的未经授权使用。

[0163] 在一些示例的情况下，可以直接从输入设备输出经变换的输入数据。图15图示了输入设备1552，其可以被实现为来自图14的系统1401的输入设备1452。输入设备1552可以包括输入捕获电路1591、输入变换电路1592和输入数据缓冲器1593。输入捕获电路1591可以是被布置成捕获输入数据1524的电路。更特别地，输入捕获电路1591可以捕获输入的指示，诸如例如麦克风、图像传感器、全球定位传感器、热传感器等。输入变换电路1592可以包括被布置成对输入捕获电路1591生成的所捕获输入的指示进行变换的电路。在一些示例的情况下，输入变换电路1592可以被布置成基于输入设备1552的特性(诸如例如，与输入设备1552相关联的UUID)来对所捕获的输入进行编码。在一些示例的情况下，输入变换电路1592可以被布置成基于硬编码到输入设备1552中的数据(诸如例如，存储在一次性可编程熔丝1594中的信息)的特性来对所捕获的输入进行编码。输入数据缓冲器1593可以包括被布置成存储数据的电路。例如，输入数据缓冲器1593可以存储经变换的输入数据1525。

[0164] 图16图示了示例技术1600，该技术可以由强化的系统(例如，强化的系统1401等)来实现，以经由如本文中讨论的输入数据的硬件变换来保护模型。要注意的是，为了方便和清楚起见，参考图14的强化的系统1401描述了技术1600。然而，这并不意图是限制性的。通过维持预加载的平台配置1472，便于保护推理模型1422免受盗窃(例如，在未经授权或未经许可的平台上执行)，该平台配置1472用于以保密的方式(例如，经由TEE 1470等等)对经变换的输入数据1425进行恢复。

[0165] 技术1600可以在圆圈16.1处开始。在圆圈16.1处，强化的系统1401可以接收输入数据1424。例如，电路1410可以从输入设备1452接收输入数据1424。在圆圈16.2处，电路1410可以从输入数据1424和平台配置1427来生成经变换的输入数据1425。更特别地，可以基于平台配置1427来对输入数据1424进行变换以生成经变换的输入数据1425。

[0166] 在圆圈16.3处，TEE 1470可以基于预加载的平台配置1472从经变换的输入数据1425来恢复输入数据。更特别地，可以使用预加载的平台配置1472对经变换的输入数据1425进行解码，以生成恢复的输入数据1474。在圆圈16.4处，可以使用推理模型1422基于恢复的输入数据1474来生成输出数据。

[0167] 图17图示了逻辑流程1700。逻辑流程1700可以表示由本文中描述的一个或多个设备执行的一些或全部操作。例如，强化的系统1401(或其组件)可以执行逻辑流程1700中的操作，以通过基于平台的硬件特性来对输入数据进行变换和恢复，从而阻止对推理模型的未经授权使用。

[0168] 逻辑流程1700可以在框1710“接收输入数据”处开始，系统1401可以接收输入数据1424。例如，强化的系统1401可以经由接口1430等从输入设备1452接收输入数据1424。继续到框1720“确定平台配置”，系统1401可以确定包括系统1401的硬件的指示的平台配置1427。例如，电路1410在执行指令1428时可以确定系统1401的组件(例如，电路1410、存储器1420、接口1430、输入设备1452等)的UUID，并且将所确定的(一个或多个)UUID设置为平台配置1427。

[0169] 继续到框1730“基于平台配置来对输入数据进行变换”，系统1401可以基于配置数据1427来对输入数据1424进行变换。例如，电路1410在执行指令1428时可以通过使用平台配置1427作为对该变换的密钥来对输入数据1424进行变换(例如，编码、加密等)从而生成经变换的输入数据1425。继续到框1740“基于预加载的平台配置和经变换的输入数据来恢复输入数据”，系统1401可以从经变换的输入数据1425和预加载的平台配置1472来恢复输入数据1424。要注意的是，系统1401可以以安全的方式来恢复输入数据1424。也就是说，可以安全地(例如，在TEE 1470等中)存储预加载的平台配置1472，并且可以经由安全处理(例如，在TEE 1470等中)生成恢复的输入数据1474。例如，TEE 1470在执行指令1478时可以通过使用预加载的平台配置1472作为对该恢复的密钥来对经变换的输入数据1425进行恢复(例如，解密、解码等)从而生成恢复的输入数据1474。

[0170] 继续到框1750“经由推理模型、基于恢复的输入数据来生成输出”，系统1401可以经由推理模型1422从恢复的输入数据1474生成输出数据1426。例如，TEE 1470可以执行推理模型1422以从恢复的输入数据1474生成输出数据1426。如上所指出，在预加载的平台配置1472(例如，预加载的UUID等)与在框1720处确定的平台配置(例如，系统1401的UUID)匹配的情况下，则恢复的输入数据1474将与原始输入数据1424匹配。然而，在预加载的平台配置1472(例如，预加载的UUID等)与在框1720处确定的平台配置(例如，系统1401的UUID)不匹配的情况下，恢复的输入数据1474将与原始输入数据1424不匹配。因此，输出数据1426可能是不可用的或无意义的，并且将阻止以未经授权的方式或由未经授权的硬件来使用推理模型1422。

[0171] *小节II-对抗性输入*

*输入净化的客户端侧卸载(offloading)*

图18图示了包括服务器1801和客户端系统1803的示例推理环境1800。通常，服务器1801和客户端系统1803可以经由任何合适的互连(诸如，广域网、局域网或互联网)而耦合。服务器1801包括电路1810-1、存储器1820-1和接口1830-1。存储器1820-1存储推理模型1822、输出数据1826、指令1828-1、净化有效负载1823和净化后的输入数据1825。客户端系

统1803包括电路1810-3、存储器1820-3、接口1830-3和输入设备1852。存储器1820-3存储原始输入数据1824、指令1828-3、净化有效负载1823和净化后的输入数据1825。

[0172] 通常，可以通过提供对抗性输入或者提供被布置或设计成引发特定推理的输入来攻击推理系统。本公开提供了对输入进行净化例如以减少来自对抗性输入的影响。在一些示例的情况下，可以将输入净化卸载到客户端或边缘设备，这与在推理服务器上实现的情况相反。作为特定示例，客户端设备可以被布置成在将净化后的输入发送到服务器以用于输入到推理模型之前接收输入并且净化输入。

[0173] 为此，客户端系统1803可以接收净化有效负载1823，该净化有效负载包括针对原始输入数据1824的输入净化过程的指示。客户端系统1803可以基于净化有效负载1823和原始输入数据1824来生成净化后的输入数据1825。例如，净化有效负载1823可以包括可用于净化原始输入数据1824以尝试缓解对抗性输入的解码、过滤和编码方案的指示。客户端系统1803可以将净化后的输入数据1825提供给服务器1801，以便与推理模型1822一起使用。

[0174] 在一些示例中，净化有效负载可以包括净化协议的指示，诸如用于净化原始输入数据1824的解码技术、过滤技术和编码技术。在一些示例中，净化有效负载1823可以包括关于对原始输入数据1824的处理和/或对净化后的输入数据1825和净化有效负载1823的处置要实现的信任度或加密级别的指示。

[0175] 作为特定示例，电路1810-1可以执行指令1828-1以将净化有效负载1823发送到客户端系统1803。电路1810-3可以执行指令1828-3以从服务器1801接收净化有效负载1823。例如，可以经由接口1830-1和1830-3来传送净化有效负载1823。

[0176] 电路1810-3可以执行指令1828-3以例如从输入设备1852接收原始输入数据1824。电路1810-3可以执行指令1828-3，以基于净化有效负载1823来净化原始输入数据1824。例如，电路1810-3可以执行指令1828-3，以对原始输入数据1824进行解码、过滤和编码，以生成净化后的输入数据1825。作为另一个示例，电路1810-3可以执行指令1828-3以对原始输入数据1824进行转码以生成净化后的输入数据1825(例如，解码并且编码成另一种格式，解码并且重新编码成相同的格式等等)。电路1810-3可以执行指令1828-3以将净化后的输入数据1825发送到服务器1801。电路1810-1可以执行指令1828-1以接收净化后的输入数据1825。例如，可以经由接口1830-1和1830-3来传送净化后的输入数据1825。电路1810-1可以执行推理模型1822，以从净化后的输入数据1825生成输出数据1826。

[0177] 在一些示例的情况下，可以向客户端系统1803提供被布置成净化输入数据的可信执行环境(TEE)。例如，图19图示了包括服务器1801和客户端系统1903的示例推理环境1900。客户端系统1903类似于图18的客户端系统1801，除了客户端系统1903还包括TEE 1970。通常，TEE 1970可以是客户端系统1903内的任何可信执行飞地或分区。例如，TEE 1970可以是电路1810-3内的可信执行分区。作为另一个示例，TEE 1970可以包括以可信的方式从电路1810-3和存储器1820-3中分割的其自己的电路和存储器(未示出)。通常，TEE 1970可以提供安全性特征，诸如隔离执行(例如，生成净化后的输入数据1825)、以及净化有效负载1823的保密性。

[0178] 图20图示了包括服务器1801、客户端系统1803和边缘设备2005的示例推理环境2000。通常，边缘设备2005可以是提供进入服务器1801中、或去往耦合到服务器1801的网络的入口点的任何设备。例如，边缘设备2005可以是路由器、路由交换机、集成接入设备

(IAD)、多路复用器、小区节点、基站等。类似于推理环境1800和1900，推理环境2000从服务器1801卸载对原始输入的净化。然而，在该示例中，边缘设备2005生成净化后的输入数据1825。

[0179] 边缘设备2005包括电路1810-5、存储器1820-5和接口1830-5。存储器1820-5存储原始输入数据1824、指令1828-5、净化有效负载1823和净化后的输入数据1825。在操作期间，边缘设备2005可以从客户端系统1803接收原始输入数据1824，可以净化原始输入数据1824，并且可以将净化后的输入数据1825提供给服务器1801。

[0180] 电路1810-1可以执行指令1828-1，以将净化有效负载1823发送到边缘设备2005。电路1810-5可以执行指令1828-3以从服务器1801接收净化有效负载1823。例如，可以经由接口1830-1和1830-3来传送净化有效负载1823。

[0181] 电路1810-3可以执行指令1828-3以例如从输入设备1852接收原始输入数据1824。电路1810-3可以执行指令1828-3以将原始输入数据1824发送到边缘设备2005。电路1810-5可以执行指令1828-5以接收原始输入数据1824。例如，可以经由接口1830-3和1830-5来传送原始输入数据1824。

[0182] 电路1810-5可以执行指令1828-5，以基于净化有效负载1823来净化原始输入数据1824。例如，电路1810-5可以执行指令1828-5，以对原始输入数据1824进行解码、过滤和编码，以生成净化后的输入数据1825。电路1810-5可以执行指令1828-5以将净化后的输入数据1825发送到服务器1801。电路1810-1可以执行指令1828-1以接收净化后的输入数据1825。例如，可以经由接口1830-1和1830-5来传送净化后的输入数据1825。电路1810-1可以执行推理模型1822，以从净化后的输入数据1825生成输出数据1826。

[0183] 重要的是要注意，尽管未描绘，但是边缘设备2005可以包括TEE，可以在TEE内生成净化后的输入数据1825，例如类似于图19的边缘设备1903。

[0184] 图21图示了示例技术2100，该技术可以在推理环境(例如，推理环境1800、推理环境1900或推理环境2000等)中实现以净化输入，如本文中详细描述的那样。要注意的是，为了方便和清楚起见，参考图20的推理环境2000描述了技术2100。然而，这并不意图是限制性的。

[0185] 技术2100可以在圆圈21.1处开始。在圆圈21.1处，强化的系统2001可以接收原始输入数据1824。例如，客户端系统1803可以从输入设备1852接收原始输入数据1824。要注意的是，如本文中所描述，可以接收和净化原始输入数据1824的多个不同实例。例如，客户端系统1803可以在圆21.1处重复地接收原始输入数据1824，并且如本文中所描述的那样净化原始输入数据1824。

[0186] 在圆圈21.2处，服务器1801可以查询客户端系统1803以确定客户端系统1803的计算能力。在圆圈21.3处，客户端系统1803可以将计算能力的指示发送到服务器1801。通常，服务器1801可以查询客户端系统1803以确定客户端系统1803的净化能力。在一些示例中，服务器1801可以查询客户端系统1803以确定客户端系统1803是否满足净化有效负载1823的最小计算要求(例如，处理、信任度级别、存储器等)。在圆圈21.4处，服务器1801和客户端系统1803可以建立可信连接。例如，服务器1801和客户端系统1803可以交换证书(例如，加密、验证等)以建立可信和/或加密的通信信道。

[0187] 在圆圈21.5处，服务器1801可以将净化有效负载发送到客户端系统1803。在一些

示例中，净化有效负载可以包括净化协议的指示，诸如用于净化原始输入数据1824的解码技术、过滤技术和编码技术。在圆圈21.6处，客户端系统1803可以基于净化有效负载1823从原始输入数据1824生成净化后的输入数据1825。例如，客户端系统1803可以对原始输入数据1824进行解码、过滤和编码。

[0188] 在圆圈21.7处，客户端系统1803可以将净化后的输入数据1825发送到服务器1801。附加地，在圆圈21.7处，服务器1801可以监测客户端系统1803的状态(例如，安全性策略、计算资源等)。此外，在圆圈21.7处，服务器1801可以从客户端系统1803停用和/或移除净化有效负载1823。

[0189] *用以发现对抗性攻击的隐藏类*

图22图示了图形2200，其描绘了针对示例推理模型的可能输入而绘制的潜在输出。描绘了包括各种类的潜在输出。具体地，潜在输出包括可见类2210-1、2210-2、2210-3和2210-4、以及隐藏类2220-1、2220-2和2220-3。此外，潜在输出是针对可能的输入[2201、2203]而绘制的。要注意的是，出于呈现清楚的目的，可能的输入2201和2203被描绘为二维的(2D)。然而，在实践中，推理模型的域可能是许多维的。

[0190] 通常，本公开提供了推理模型的域可以包括可见类(例如，类2210)以及隐藏类(2220)两者。隐藏类用于检测针对模型的可能探测攻击，这将在下面详细解释。对推理系统的一个安全性风险是探测。可以利用输入扰动对推理模型进行探测或攻击，以试图标识可以“欺骗”该系统的输入。更具体地，攻击者可以对图像输入中使用的已知对象进行装饰，注入噪声，或者以其他方式对去往推理模型的输入进行变换。这些经修改或变换的输入可以用于探测推理模型，以生成导致错误分类的输入。例如，使枪和水龙头的图像变形可能导致将枪的图像错误分类为水龙头的推理引擎。

[0191] 作为另一个示例，探测可以发生在其中对抗方不知道模型支持的功能的实例中。在这种情况下，对抗方可以应用人工合成的输入或噪声输入，以试图得到对该模型的功能的理解。这种输入通常会激活或激发该模型中的通常不会被常规或预期输入所激活的区域。

[0192] 一些推理系统尝试对输入进行净化，例如通过对输入进行预处理以过滤掉对抗性扰动。然而，这种技术并不总是可靠的。此外，针对这种预处理的计算要求增加了推理系统的所需功率和计算资源要求。由此，预处理并不总是可扩展的。

[0193] 图23A图示了用以训练推理模型的示例系统2300。系统2300包括电路2310和存储器2320。存储器2320存储推理模型2322、指令2328、测试/训练数据集2360和输出数据2326。测试/训练数据集2360包括测试数据2362和训练数据2364两者。测试数据2362和训练数据2364可以用于例如使用图3的技术300来“训练”推理模型2322。

[0194] 测试/训练数据集2360的示例在图23B中被描绘。简要地转向该图，测试/训练数据集2360包括输入2361以及预期输出2363。通常，预期输入可以包括(一个或多个)输入值2361-n，其中n是正整数。例如，如针对输入2361示出了(一个或多个)输入值2361-1、(一个或多个)输入值2361-2、(一个或多个)输入值2361-3、(一个或多个)输入值2361-4、(一个或多个)输入值2361-5至(一个或多个)输入值2361-N。针对每个输入值2361-n，示出了相关联的预期输出2363。预期输出2363中的每一个对应于可见类2210或隐藏类2220，推理模型2322被训练成将输入分类到该可见类2210或隐藏类2220。作为示例，可见类2210可以对应

于将由推理系统2322分类的对象的类别，诸如钢笔、刀、剑、直尺、剪刀和铅笔。隐藏类2220可以对应于可以指示潜在探测的类别，诸如例如伪装成钢笔的刀、伪装成直尺的剑等。在一些示例的情况下，隐藏类2220可以对应于不被允许作为输入的列入黑名单的对象。与隐藏类相关联的(一个或多个)输入值2361-n可以对应于已经被变换以尝试触发由推理模型2322做出错误分类的输入。例如，可以通过利用水龙头的图像使枪的图像变形来生成(一个或多个)示例输入值2361-n。

[0195] 在操作期间，当输入触发了隐藏类之一的归类或推理时，经训练的推理模型2322可以用于标识针对该网络的潜在探测攻击。其中实现了推理模型2322的推理系统可以被布置成采取一些缓解动作，这将在下面更详细地描述。

[0196] 图24图示了包括强化的系统2401和服务器2403的示例推理环境2400。通常，强化的系统2401和服务器2403可以经由任何适当的互连(诸如、广域网、局域网或互联网)而耦合。强化的系统2401包括电路2410-1、存储器2420-1、接口2430-1、输入设备2252和可信执行环境(TEE)2470。存储器2420-1存储指令2428-1、输入数据2424和混淆的输出数据2427。TEE 2470包括可信电路2411和可信存储器2421。可信存储器2421存储TEE指令2429、推理模型2322、输出数据2426、混淆的输出数据2427、取证(forensic)数据2482和探测警报2484。

[0197] 服务器2403包括电路2410-3、存储器2420-3和接口2430-3。存储器2420-3存储指令2428-3、取证数据2482、探测警报2484和取证数据库2486。

[0198] 通常，强化的系统2401进行操作以检测潜在探测攻击。电路2410-1可以执行指令2428-1以例如从输入设备2452接收输入数据2424。TEE电路2411可以执行TEE指令2429，以通过使用输入数据DD324对推理模型2322的执行来生成输出数据2426。TEE电路2411可以执行TEE指令2429，以确定输入数据2424是否可能是潜在探测攻击，或者也就是说输入数据DD324是否与潜在探测攻击相关联。例如，TEE电路2411在执行TEE指令2429时可以确定输出数据2426是否对应于隐藏类2220(例如，隐藏类2220之一等等)。换句话说，如果推理模型2322将输入数据2424分类为隐藏类2220之一，则TEE电路2411可以确定输入数据2424与探测攻击相关联。

[0199] 响应于确定2424与潜在探测攻击相关联，TEE电路2411在执行TEE指令2429时可以收集关于潜在探测攻击的信息，并且将收集到的信息保存为取证数据2482。例如，TEE电路2411在执行TEE指令2429时可以保存包括输入设备2452、输入设备2452操作者、输入数据2424、关于推理模型2322的数据(例如，情境、历史、状态等)的指示的信息作为取证数据2482。此外，TEE电路2411在执行TEE指令2429时可以生成探测警报2484，并且可以向服务器2403(例如，经由接口2430-1等)发送该探测警报的指示。探测警报2484可以包括针对系统2401发起了潜在探测攻击并且保存了取证数据2482的指示。附加地，TEE电路2411在执行TEE指令2429时可以向服务器2403提供取证数据2482。

[0200] 在一些实现方式中，强化的系统2401可以将输出数据2426混淆，以隐藏可能已经检测到探测的事实。例如，TEE电路2411在执行TEE指令2429时可以从输出数据2426生成混淆的输出数据2427，其中混淆的输出数据2427可以包括可见类2210。例如，TEE电路2411在执行TEE指令2429时可以生成如下混淆的输出数据2427:其包括与输出数据2426的隐藏类2220相关联的可见类2210的指示。使用来自上面的枪和水龙头示例;输出数据2426可以包括与输入数据2424相对应的隐藏分类2220的指示，该输入数据2424是枪的图像，该图像被

修改以便将枪欺骗或错误分类为水龙头。然而，混淆的输出数据2427可以指示输入数据2424被分类为水龙头。例如，可以生成混淆的输出数据2427，从而不会向攻击者提供检测到探测的通知。

[0201] 电路2410-3可以执行指令2428-3以接收探测警报2482。响应于接收到探测警报2482，电路2410-3在执行指令2428-3时可以从强化的系统2401接收取证数据2484，并且可以将取证数据2484添加到取证数据库2486。取证数据库2486可以用于例如防止进一步的错误分类。

[0202] 图25图示了逻辑流程2500。逻辑流程2500可以表示由本文中描述的一个或多个设备执行的一些或全部操作。例如，强化的系统2401(或其组件)可以执行逻辑流程2500中的操作，以使用隐藏类来检测潜在的探测。

[0203] 逻辑流程2500可以在框2510“接收输入数据”处开始，系统2401可以接收输入数据2424。例如，电路2410-1在执行指令2428-1时可以(例如，从输入设备2452等)接收输入数据2424。继续到框2520“经由推理模型、基于接收到的输入数据来生成输出数据”，强化的系统2401可以从推理模型2322和输入数据2424生成输出数据2426。例如，TEE电路2411在执行指令2429时可以通过在输入数据2424上执行推理模型2322来生成输出数据2426。

[0204] 继续到判定框2530“输出数据是否与隐藏类相关联”，强化的系统2401可以确定输出数据2426是否与隐藏类2220相关联，或者换句话说，推理系统2322是否将输入数据2424分类为隐藏类2220。从判定框2530，逻辑流程2500可以继续到框2540或框2550。逻辑流程2500可以基于输出数据2426不与隐藏类2220相关联的确定而从判定框2530继续到框2540，而逻辑流程2500可以基于输出数据2426与隐藏类2220相关联的确定而从判定框2530继续到框2550。

[0205] 在框2540“提供所生成的输出”处，强化的系统2401可以将输出数据2426提供给输出数据消费者(例如，输出设备、另一计算设备等)。例如，电路2410-1在执行指令2428-1时可以将输出数据2426提供给输出数据消费者。从框2540，逻辑流程2500可以结束，或者可以基于接收到新的输入而重复。

[0206] 在框2550“向监测系统警告潜在探测”处，强化的系统2401可以向监测系统(例如，服务器2403)警告潜在探测攻击。例如，电路2410-1在执行指令2428-1时可以向服务器2403提供探测警报2482。继续到框2560“收集取证数据”，强化的系统2401可以收集与潜在探测攻击相关联的取证数据。例如，TEE电路2411在执行TEE指令2429时可以收集与潜在探测攻击相关联的信息，并且将收集到的信息的指示保存为取证数据2484。继续到框2570“将取证数据提供给监测系统”，强化的系统2401可以将取证数据2484提供给监测系统(例如，服务器2403)。例如，TEE电路2411在执行TEE指令2429时可以将取证数据2484提供给服务器2403。

[0207] 继续到框2580“将输出数据混淆”，强化的系统2401可以将输出数据2426混淆。例如，TEE电路2411在执行TEE指令2429时可以从输出数据2426生成混淆的输出数据2427。继续到框2590“将混淆的输出数据提供给输出数据消费者”(例如，输出设备、另一计算设备等)。例如，电路2410-1在执行指令2428-1时可以将混淆的输出数据2427提供给输出数据消费者。从框2590，逻辑流程2500可以结束，或者可以基于接收到新的输入而重复。

[0208] *用以针对对抗性攻击进行保护的输入失真*

如所指出的，推理模型容易受到沿着多个攻击向量的对抗性攻击。例如，对推理模型的盗窃和/或未经授权使用、以及对探测的易受性就是两个这种攻击向量。由此，检测对推理模型的不适当或未经授权使用以及防止或阻止探测的能力是合期望的。

[0209] 通常，本公开提供了利用混淆的输入数据来训练的推理模型。换句话说，本公开提供了被训练成对被混淆或包括一些失真的输入数据正确地进行分类的推理模型。可以经由处理(图26至28)或作为输入设备的物理性质的结果(图29至30)来提供该混淆或失真。因此，基于对抗性扰动的攻击可能不太有效或无效，这是因为该模型被训练成对混淆的输入数据进行分类，而不是对常规输入数据进行分类。

[0210] 在一些示例中，可以经由使用预先选择的密钥等对输入数据进行编码来将输入数据混淆(例如，图26和27)。在其他示例中，可以经由过滤器、输入失真设备等在输入设备处将输入数据混淆(例如，图28)。要注意的是，本文中经常使用图像捕获设备和图像混淆的示例。然而，本公开可以被实现为将任何类型的输入数据混淆，该输入数据诸如例如音频、视频、其他数据类型或结构等。

[0211] 图26图示了可以根据本公开来实现的示例强化的系统2601。强化的系统2601包括电路2610、存储器2620、接口2630、输入设备2652和可信执行环境(TEE)2670。存储器2620存储推理模型2622、输入数据2624、混淆的输入数据2625、输出数据2626和指令2628。TEE 2670存储混淆密钥2672和TEE指令2678。通常，TEE 2670可以是系统2601内的任何可信执行飞地(enclave)或分区。例如，TEE 2670可以是电路2610内的可信执行分区。作为另一个示例，TEE 2670可以包括以可信的方式从电路2610和存储器2620中分割的其自己的电路和存储器(未示出)。通常，TEE 2670可以提供安全性特征，诸如隔离执行(例如，对TEE指令2678的隔离执行)、以及混淆密钥2672的保密性。通常，系统2601可以以安全的方式(例如，经由TEE 2670等)生成混淆的输入数据。可以将这种混淆的输入数据用作针对推理模型(例如，推理模型2622)的输入数据，该推理模型被训练成对这种混淆的输入数据进行分类。因此，在没有正确混淆的输入数据的情况下，推理模型可能是无法使用的。也就是说，推理模型可能无法对输入数据本身正确地进行分类，这是由于它是在混淆的输入数据上训练的。这可以阻止模型盗窃、对模型的未经授权使用或者探测。

[0212] 在操作期间，系统2601的电路2610可以执行指令2628以接收输入数据2624。例如，电路2610可以从输入设备2652接收输入数据2624。作为另一个示例，电路2610可以经由接口2630从另一计算设备接收输入数据2624。

[0213] TEE 2670在执行指令2678时可以从输入数据2624和混淆密钥2672生成混淆的输入数据2625。例如，TEE 2670在执行指令2678时可以使用混淆密钥2672对输入数据2624进行编码、失真或以其他方式进行混淆。

[0214] 电路2610在执行指令2628时可以执行推理模型2622以从混淆的输入数据2625生成输出数据2626。如上所指出，推理模型2622是在混淆的输入数据(例如，如混淆的输入数据2625)上训练的，因此推理模型2622可以对被正确混淆的输入数据2624正确地进行分类，如本文中所描述的那样。

[0215] 在一些示例的情况下，输入设备可以被布置成输出混淆的输入数据。图27图示了包括系统2701和输入设备2752的推理环境2700。输入设备2752包括电路2710-1、存储器2720-1、接口2730-1、输入捕获电路2751和可信执行环境(TEE)2770。存储器2720-1存储输

入数据2724、混淆的输入数据2725和指令2728-1。TEE 2770存储混淆密钥2772和TEE指令2778。通常，TEE 2770可以是输入设备2752内的任何可信执行飞地或分区。例如，TEE 2770可以是电路2710-1内的可信执行分区。作为另一个示例，TEE 2770可以包括以可信的方式从电路2710-1和存储器2720-1中分割的其自己的电路和存储器(未示出)。通常，TEE 2770可以提供安全性特征，诸如隔离执行(例如，对TEE指令2778的隔离执行)、以及混淆密钥2772的保密性。

[0216] 输入设备2752可以被布置成捕获输入数据，将输入数据混淆，并且将混淆的输入数据提供给系统2701。通常，输入设备2752可以以安全的方式(例如，经由TEE 2770等)生成混淆的输入数据。输入捕获电路2751可以是被布置成捕获输入数据2724的任何硬件。例如，输入捕获电路2751可以是图像传感器(例如，CMOS传感器等)、麦克风、GPS传感器、热传感器，生物识别传感器等。在一些示例中，电路2710-1可以执行指令2728-1以接收由输入捕获电路2751生成的信号，并且将这些信号的指示存储为输入数据2724。

[0217] TEE 2770在执行指令2778时可以从输入数据2724和混淆密钥2772生成混淆的输入数据2725。例如，TEE 2770在执行指令2778时可以使用混淆密钥2772对输入数据2724进行编码、失真或以其他方式进行混淆。电路2710-1可以执行指令2728-1以将混淆的输入数据2725发送到系统2701。例如，电路2710-1可以将控制信号发送到接口2730-1，以使接口2730-1(例如，经由接口2730-2等)向系统2701发送包括混淆的输入数据2725的指示的信息元素。

[0218] 电路2710-2在执行指令2728-2时可以执行推理模型2722以从混淆的输入数据2725生成输出数据2726。如上所指出，推理模型2722是在混淆的输入数据(例如，如混淆的输入数据2725)上训练的，因此推理模型2722可以对被正确混淆的输入数据2724正确地进行分类，如本文中所描述的那样。

[0219] 图28图示了逻辑流程2800。逻辑流程2800可以表示由本文中描述的一个或多个设备执行的一些或全部操作。例如，强化的系统2601(或其组件)可以执行逻辑流程2800中的操作，以通过将输入数据混淆来阻止对抗性攻击。作为另一个示例，推理环境系统2700(或其组件)可以执行逻辑流程2800中的操作，以通过将输入数据混淆来阻止对抗性攻击。出于清楚的目的，参考强化的系统2601描述了逻辑流程2800。然而，示例在该情境中不受限制。

[0220] 逻辑流程2800可以在框2810“接收输入数据”处开始，系统2601可以接收输入数据2624。例如，强化的系统2601可以经由接口2630等从输入设备2652接收输入数据2624。继续到框2820“将输入数据混淆”，系统2601可以基于混淆密钥2672将输入数据2624混淆。要注意的是，系统2601可以以安全的方式将输入数据2624混淆。也就是说，可以安全地(例如，在TEE 2670等中)存储混淆密钥2672，并且可以经由安全处理(例如，在TEE 2670等中)生成混淆的输入数据2625。例如，TEE 2670在执行指令2678时可以使用混淆密钥2672将输入数据2624混淆以生成混淆的输入数据2625。

[0221] 继续到框2830“经由推理模型、基于混淆的输入数据来生成输出”，系统2601可以经由推理模型2622从混淆的输入数据2625生成输出数据2626。例如，电路2610可以执行推理模型2622以从混淆的输入数据2625生成输出数据2626。

[0222] 在一些实现方式中，输入可以由输入设备的硬件特征或物理特征来混淆。例如，相机通常用作被布置成对图像进行分类的推理系统的输入设备。在这种示例中，可以向相机

提供用以将所捕获的图像失真或混淆并且提供混淆的输入数据的镜头。

[0223] 图29图示了包括系统2901和输入设备2952的推理环境2900。输入设备2952包括电路2910-1、存储器2920-1、接口2930-1和输入捕获电路2951。存储器2920-1存储混淆的输入数据2925和指令2928-1。输入设备2952可以被布置成捕获混淆的输入数据，并且将混淆的输入数据提供给系统2901。通常，输入设备2952可以捕获输入，并且作为捕获该输入的一部分将所捕获的输入混淆。例如，在输入设备2952是相机的情况下，可以向相机提供混淆镜头2953，该混淆镜头2953被布置成使该相机所捕获的图像失真。电路2910-1可以执行指令2928-1以接收由输入捕获电路2951生成的信号，并且将这些信号的指示存储为混淆的输入数据2925。此外，电路2910-1可以执行指令2928-1以将混淆的输入数据2925发送到系统2901。例如，电路2910-1可以将控制信号发送到接口2930-1，以使接口2930-1(例如，经由接口2930-2等)向系统2901发送包括混淆的输入数据2925的指示的信息元素。

[0224] 电路2910-2在执行指令2928-2时可以执行推理模型2922以从混淆的输入数据2925生成输出数据2926。如上所指出，推理模型2922是在混淆的输入数据(例如，如混淆的输入数据2925)上训练的，因此推理模型2922可以对经由图像捕获电路2951和混淆镜头2953捕获的输入正确地进行分类。

[0225] 图30图示了逻辑流程3000。逻辑流程3000可以表示由本文中描述的一个或多个设备执行的一些或全部操作。例如，推理环境系统2900(或其组件)可以执行逻辑流程3000中的操作，以通过将输入数据混淆来阻止对抗性攻击。

[0226] 逻辑流程3000可以在框3010“从输入设备接收混淆的输入数据”处开始，可以从输入设备接收混淆的输入数据2925。输入数据2952可以被布置成捕获图像数据，并且(例如，经由混淆镜头2953等)将所捕获的图像数据混淆，作为捕获过程的一部分。

[0227] 继续到框3020“经由推理模型、基于混淆的输入数据来生成输出”，系统2901可以经由推理模型2922从混淆的输入数据2925生成输出数据2926。例如，电路2910-2可以执行推理模型2922以从混淆的输入数据2925生成输出数据2926。

[0228] *基于网络激活的推理概率*

如上所指出，针对推理系统的潜在攻击向量正在利用推理系统中的固有不准确性(例如，损失函数等)，以意图引起错误分类。这种攻击的常见示例是操纵图像中的选定像素，以使分类系统将该图像错误分类。作为特定示例，提供给加拿大高级研究学院(CIFAR-10)分类系统的单个像素图像使得以80％置信度将该单个像素图像分类为飞机。

[0229] 本公开提供了强化的系统，其被布置成利用与从推理系统生成输出有关的信息(例如，网络激活等)来生成推理可靠性得分。可以生成可靠性得分以增强(augment)或修改由推理模型本身生成的置信概率。换句话说，一些推理模型(例如，卷积神经网络等)生成包括推理以及置信概率的指示的输出。在这种示例中，可以生成可靠性得分以补充由推理模型生成的置信概率。

[0230] 图31图示了可以根据本公开来实现的示例强化的系统3101。强化的系统3101包括电路3110、存储器3120、接口3130和输入设备3152。存储器3120存储推理增强图3121、推理模型3122、输入区域影响图3123、输入数据3124、可靠性得分3125、输出数据3126、可靠性规则3127和指令3128。

[0231] 通常，强化的系统3101可以部地分基于输入数据3124的哪些部分或组分影响了输

出数据3126的指示来提供可靠性得分。在操作期间，系统3101的电路3110可以执行指令3128以接收输入数据3124。例如，电路3110可以从输入设备3152接收输入数据3124。作为另一个示例，电路3110可以经由接口3130从另一计算设备接收输入数据3124。

[0232] 电路3110在执行指令3128时可以执行推理模型3122，以便既从输入数据3124生成输出数据3126，并且又生成推理增强图3121。换句话说，推理模型3122基于输入数据3124可以输出：(1)输出数据3126和(2)推理增强图3121。通常，推理增强图3121可以包括在推理期间执行的算术运算的指示或信息、推理模型3122在执行期间(例如，当生成输出3126时)的状态等。在一些示例的情况下，推理增强图3121可以包括在生成输出数据3126期间的模型激活图或网络激活热图(例如，包括层激活、神经元激活等的指示)。在一些示例的情况下，推理增强图3121可以包括多个层(例如，与针对推理模型3122的相应层的神经元激活相对应的每个层等等)。

[0233] 电路3110在执行指令3128时可以生成输入区域影响图3123。例如，电路3110可以基于推理增强图3121来掩盖输入数据3124。作为特定示例，电路3110可以标识输入数据3124中的如下区域、部分或组份：这些区域、部分或组份并不主要负责激活推理模型3122以生成输出数据3126。作为特定示例，假定推理模型3122是图像分类器并且输入数据3124是图像。输入区域影响图3123可以包括输入数据3124的副本，该副本标识了影响输出数据3126的那些区域(基于推理增强图3121)。在一些示例的情况下，电路3110可以基于推理增强图3121和可靠性规则3127来生成输入区域影响图3123。在这种示例中，可靠性规则3127可以包括与神经元激活相对应的阈值激活水平的指示。例如，在神经元激活函数导致了在0与1之间的输出的情况下，可靠性规则3127可以包括神经元激活应当大于0.3以便贡献于或影响输出数据3126的指示。在一些示例的情况下，可靠性规则3127可以在域、任务、分类的类等当中变化。作为特定示例，可靠性规则3127可以包括用于分类的类“狗”的规则和用于分类的类“猫”的不同规则。

[0234] 电路3110在执行指令3128时可以基于输入区域影响图3123和可靠性规则3127来生成可靠性得分3125。使用上面讨论的图像分类器示例，电路3110可以基于对输出数据3126有贡献(或没有贡献)的输入数据3124(例如，图像)的百分比来生成可靠性得分3125(例如，基于输入区域影响图3123等)。作为另一个示例，电路3110可以基于对输出数据3126有贡献(或没有贡献)的图像的部分的图像数据变化性(例如，像素颜色、像素颜色之间的差异等)来生成可靠性得分3125(例如，基于输入区域影响图3123等)。在一些示例的情况下，电路3110可以基于对输出数据3126有贡献的图像的部分的图像数据变化性(例如，像素颜色、像素颜色之间的差异等)相比于对输出数据3126没有贡献的图像的部分的图像数据变化性的比率来生成可靠性得分3125。在一些示例的情况下，电路3110可以基于输入数据3124变化性、对输出数据3126有贡献(或没有贡献)的输入数据3124的百分比、和/或输入数据可变性的比率的组合来生成可靠性得分3125。在一些示例的情况下，电路3110可以基于对输出数据3126有贡献(或没有贡献)的输入数据3124的哪个部分(哪些部分)来生成可靠性得分3125。例如，在输入数据3124的所选部分(例如，图像的拐角等)对输出数据3126有贡献的情况下，可靠性得分可能较低。

[0235] 图32图示了示例技术3200，该技术可以由强化的系统(例如，强化的系统3100等)来实现以生成如本文中所描述的推理可靠性得分。为了方便和清楚起见，参考图31的强化

的系统3100描述了技术3200。然而，这并不意图是限制性的。技术3200可以在圆圈32.1处开始。在圆圈32.1处，输入设备3152可以捕获输入数据3124。

[0236] 在圆圈32.2处，推理模型3122可以生成输出数据3126。例如，电路3110可以执行推理模型3122以生成输出数据3126。在一些示例中，输出数据可以包括推理和推理概率两者。在圆圈32.3处，推理模型3122可以输出推理增强图3121。例如，推理模型3122可以生成推理增强图3121作为生成输出数据3126的一部分。

[0237] 在圆圈32.4处，电路3110在执行指令3128时可以生成输入区域影响图3123。例如，电路3110可以(例如，基于推理增强图3121)确定输入数据3124激活了推理模型3122的哪些部分。在圆圈32.5处，电路3110在执行指令3128时可以生成输入区域影响图3123。例如，电路3110可以部分地基于可靠性规则和输入区域影响图3123来生成可靠性得分3125。技术3200可以可选地包括圆圈32.6。在圆圈32.6处，电路3110可以更新、修改或以其他方式增强由推理模型3122作为输出数据3126的一部分所输出的概率。

[0238] 图33图示了逻辑流程3300。逻辑流程3300可以表示由本文中描述的一个或多个设备执行的一些或全部操作。例如，强化的系统3101(或其组件)可以执行逻辑流程3300中的操作，以生成可靠性得分，该可靠性得分包括由推理模型生成的输出的可靠性的指示。出于清楚的目的，参考强化的系统3101描述了逻辑流程3300。然而，示例在该情境中不受限制。

[0239] 逻辑流程330可以在框3310“接收输入数据”处开始，系统3101可以接收输入数据3124。例如，强化的系统3101可以经由接口3130等从输入设备3152接收输入数据3124。继续到框3320“经由推理模型、基于接收到的输入数据来生成输出数据”和框3330“经由推理模型来生成推理增强图”，系统3101可以经由推理模型3122从输入数据3124生成输出数据3126并且生成推理增强图3121。例如，电路3110可以执行推理模型3122，以从输入数据3124生成输出数据3126并且生成与推理模型3122的执行(或推理)相关联的推理增强图3121。如所指出的，推理增强图3121基于来自输出数据生成的推理模型激活。因此，推理增强图3121对应于生成输出数据3126期间的推理模型3122的“激活”。如所指出的，推理增强图3121可以包括与推理模型3122的激活相对应的网络激活热图。

[0240] 继续到框3340“基于输入数据和推理增强图来生成输入区域影响图”，系统3101可以生成输入区域影响图3123，该输入区域影响图3123包括影响输出数据3126(例如，对推理模型3122的激活有贡献)的输入数据3124的区域的指示。例如，电路3110在执行指令3128时可以基于推理增强图3121来标识对输出数据3126有贡献的部分输入数据3124，并且可以生成包括输入数据3124的这种部分的指示的输入区域影响图3123。

[0241] 继续到框3350“至少基于输入区域影响图来生成可靠性得分”，系统3101可以基于输入区域影响图3123来生成可靠性得分3125。例如，电路3110在执行指令3128时可以基于输入区域影响图3123来确定指示输出数据3126的可靠性的得分，并且将该得分的指示存储为可靠性得分3125。

[0242] 逻辑流程3300可以可选地包括框3360“将可靠性得分与由推理模型生成的概率融合”，系统3101可以更新、修改或以其他方式增强由推理模型3122作为输出数据3126的一部分所输出的概率。

[0243] *小节III-数据信任度*

*通过组合第三部分推理的输入集合净化*

图34图示了可以根据本公开来实现的示例强化的系统3401。强化的系统3401包括电路3410、存储器3420和输入设备3452。存储器3420存储一个或多个推理模型3422-1、3422-2、3422-*N*(或推理模型3422)、训练数据集3424、一个或多个分类数据3425-1、3425-2、3425-*N*(或分类数据3425)、不相关的输出数据3426、相关的输出数据3427和指令3428。在操作期间,电路3410可以执行指令3428和/或一个或多个推理模型3422,以将训练数据集3424分离成不相关的输出数据3426和/或相关的输出数据3427。

[0244] 针对推理系统的攻击可以包括在推理模型的训练期间创建后门(例如,隐藏类)。后门可以向恶意行为者提供非意图的访问或者控制推理系统的一个或多个方面。例如,恶意行为者(例如,流氓数据科学家或数据供应者)可能会将不适当的数据注入到训练集中,从而导致经训练的推理模型中的隐藏漏洞。这种漏洞可能导致推理模型对输入数据进行错误分类。作为另一个示例,注入到训练集中的不适当数据可能会产生标识出恐怖分子和知名人士两者的经训练的推理模型。

[0245] 通常,准备训练数据(例如,对输入进行标记等)是暴露攻击向量(诸如,恶意行为者对训练数据的修改)的人类劳动。复查、检测并且校正不正确地准备的训练数据也是一个耗时且劳动密集的过程。本公开提供了强化的系统3401,其被布置成利用两个或更多个推理模型3422来对训练数据集3424中的每个样本进行分类,以确定每一个样本可以被认为是不相关的输出数据3426还是相关的输出数据3427。

[0246] 例如,在操作期间,电路3410可以执行指令3428以接收训练数据集3424(例如,作为来自输入设备的输入数据)。通常,训练数据集3424可以包括多个样本。对于训练数据集3424中的每个样本,电路3410可以执行用以生成分类数据3425-1的推理模型3422-1、用以生成分类数据3425-2的推理模型3422-2、以及用以生成分类数据3425-*N*的推理模型CA122-*N*。有时,训练数据集3424中的一个或多个样本的分类数据可以由第三方服务来生成。电路3410可以执行指令3428以比较分类数据3425,并且确定训练数据集3424中的每个样本是否被推理模型3422中的每一个分类为相同的。

[0247] 此外,电路3410可以执行指令3428以将训练数据集3424中未被推理模型3422中的每一个分类为相同的样本添加到不相关的输出数据3426,并且可以将训练数据集3424中被推理模型3422中的每一个分类为相同的样本添加到相关的输出数据3427。在其他时候,可以使用不同分类的阈值数量来确定是应当将样本添加到不相关的输出数据3426还是相关的输出数据3427。在各种示例中,可以基于由不同系统在相同输入上提供的输出中的分类器分布的函数来建立用于推理模型3422的信任度参数。

[0248] 相关的输出数据3427可以用于训练被信任为不包括后门的推理模型。换句话说,相关的输出数据3427可以用作如下机器学习算法的训练集,该机器学习算法用于生成被信任为不包括后门的推理模型。在一些示例中,在相关的输出数据(相关的输出数据3427)上训练的推理模型可以被称为可信推理模型。附加地,不相关的输出数据3426可以被用在进一步可控制的对抗性或弹性训练流程中。有时,电路3410可以执行指令3428以将不相关的输出数据3426和/或相关的输出数据3427提供给输出设备、另一计算设备等。在一个或多个示例中,可以生成表征相关的输出数据3426和/或不相关的输出数据的一组测量结果。

[0249] 图35图示了逻辑流程3500。逻辑流程3500可以表示由本文中描述的一个或多个设备执行的一些或全部操作。例如,强化的系统3401(或其组件)可以执行逻辑流程3500中的

操作以将训练数据集分离成不相关的输出数据和相关的输出数据。

[0250] 逻辑流程3500可以在框3510“接收具有多个样本的训练数据集”处开始，其中强化的系统3401可以接收包括多个样本的训练数据集3424。例如，强化的系统3401可以从输入设备3452接收训练数据集3424。继续到框3520“基于训练数据集中的样本针对两个或更多个推理模型来生成分类数据”，强化的系统3401可以基于训练数据集3424中的样本针对两个或更多个推理模型3422中的每一个来生成分类数据3425。例如，强化的系统3401可以基于推理模型3422-1针对训练数据集3424中的样本来生成分类数据3425-1或其一部分，基于推理模型3422-2针对训练数据集3424中的样本来生成分类数据3425-2或其一部分，以及基于推理模型3422-*N*针对训练数据集3424中的样本来生成分类数据3425-*N*或其一部分。

[0251] 进行到判定框3530“基于样本针对每个推理模型而生成的分类数据是否满足相关性阈值”，强化的系统3401可以确定基于训练数据集3424中的样本针对两个或更多个推理模型3530中的每一个而生成的分类数据是否满足相关性阈值。在一些示例中，强化的系统3401可以确定训练数据集3424中的样本是否被两个或更多个推理模型3422指派了具有最小置信度水平的共同分类。例如，强化的系统3401可以确定与该样本相对应的分类数据3425-1的部分、与该样本相对应的分类数据3425-2的部分、以及与该样本相对应的分类数据3425-*N*的部分是否包括超过相关性阈值(例如80％)或者在最高“x”置信度得分内的样本分类。

[0252] 有时，没有样本可以满足相关性阈值。从判定框3530，逻辑流程3500可以继续到框3540或3550。在许多示例中，样本不满足相关性阈值。逻辑流程3500可以基于根据该样本针对每个推理模型而生成的分类数据相匹配的确定而从框3530继续到框3540，而逻辑流程3500可以基于根据该样本针对每个推理模型而生成的分类数据不匹配的确定而从框3530继续到框3550。

[0253] 在框3540“将训练数据集中的样本添加到相关的输出数据”处，强化的系统3401可以将来自训练数据集3424的样本添加到相关的输出数据3427。在框3550“将训练数据集中的样本添加到不相关的输出数据”处，强化的系统3401可以将来自训练数据集3424的样本添加到不相关的输出数据3427。继续到判定框3560“训练数据集中的每个样本是否被添加到不相关的输出数据或相关的输出数据”，强化的系统3401可以确定训练数据集3424中的每个样本是否已经被添加到不相关的输出数据3426或相关的输出数据3427。换句话说，强化的系统3401可以确定训练数据集3424中的所有样本是否已经被添加到不相关的输出数据3426或相关的输出数据3427。从判定框3560，逻辑流程3500可以结束，或者返回到3520并且针对训练数据集3424中的下一个样本而重复框3520-3560。

[0254] *针对输入设备的来源信任度适配*

当前推理系统的一个问题是各种输入设备可能生成具有不同特性(诸如，分辨率、格式、帧速率、动态范围、信任度等)的输入数据。各种输入设备的不同特性可能会影响来自这些各种输入设备的输入数据如何训练推理模型或者如何被推理模型分类。例如，在来自具有高分辨率的输入设备(例如，视频相机)的输入数据上训练的推理模型可能难以对来自具有低分辨率的输入设备的输入数据进行分类。在另一个示例中，输入设备可以位于具有一定范围可访问性的环境范围内(例如，相比于公园，位于私有建筑物中)。在诸如公园之类的高度可访问的环境中，恶意行为者可能具有更好的机会来操纵由输入设备生成的输入数

据，例如创建如本文中讨论的后门。因此，与难以访问的环境中的输入设备相比，在高度可访问的环境中的输入设备可能与较低的信任度级别相关联。

[0255] 在另一个示例中，输入设备与强化的系统之间的数据路径的安全性可能是输入设备的信任度级别中的因素(例如，加密的数据路径可以被认为比未加密的数据路径更值得信赖)。在一些实例中，信任度级别可以指代输入设备的一个或多个操作参数和/或特性。因此，本公开提供了一种推理系统或推理环境，其中部分地基于输入设备的特性来进行训练或推理。例如，本公开提供了一种系统，该系统被布置成：当选择要使用来自输入设备的输入数据来训练(多个推理模型中的)哪个推理模型时，使用输入设备特性(例如，参见图38-39)。作为另一个示例，本公开提供了一种系统，该系统被布置成：当选择要使用(多个推理模型中的)哪个推理模型以便使用来自输入设备的输入数据来生成推理时，使用输入设备特性(例如，参见图38-39)。

[0256] 要注意的是，尽管本文中提供的许多示例(例如，与图36-39相关联的描述)使用视频相机作为输入设备；但是，许多输入设备可以是合适的。例如，本公开可以用静物相机、麦克风、红外检测器、微波检测器、或被布置成发射指示输入的信号的任何其他设备来实现。

[0257] 图36图示了可以根据本公开来实现的示例强化的系统3601。强化的系统3601包括电路3610、存储器3620、接口3630和输入设备3652-1和3652-2。存储器3620存储一个或多个未经训练的推理模型3621-1、3621-2、3621-*N*(或未经训练的推理模型3621)、输入数据3624-1和3624-2、一个或多个输入设备特性3625-1和3625-2和指令3628。在许多示例中，输入数据3624可以包括训练数据集(例如，训练数据集362)，而未经训练的推理模型3621可以指代基于训练数据集(例如3624)来生成推理模型的算法。在操作期间，电路3610可以执行指令3628，以基于相应输入数据3624的输入设备特性3625来选择要利用输入数据3624来训练的未经训练的推理模型3621之一。

[0258] 例如，电路3610可以执行指令3628以分别从输入设备3652-1和3652-2接收输入数据3624-1和3624-2。电路3610可以执行指令3628以确定分别对应于输入设备3652-1和3652-2的(一个或多个)输入设备特性3625-1和3625-2。有时，一个或多个输入设备特性3625-1和3625-2可以通过诸如在启动时经由接口3630查询相应的输入设备(例如3652-1、3652-2等)来确定。例如，输入设备3652-1可以具有应用编程接口(API)，经由该应用编程接口，可以确定一个或多个输入设备特性3625-1。在其他时间，相应的输入设备(例如3652-1、3652-2等)可以与将输入数据3624-1和3624-2的一个或多个部分提供给存储器3620结合地提供一个或多个输入设备特性3625-1和3625-2。例如，输入设备3652-2可以与提供输入数据3624-2结合地提供输入设备特性3625-2。

[0259] 有时，电路3610可以执行指令3628，以基于存储器3620利用其耦合到输入设备3652的接口3630、或基于接口3630通过其与输入设备3652进行通信的通信方案来标识一个或多个输入设备特性3625-1。例如，可以至少部分地基于接口3630来确定输入设备3652-1与存储器3620之间的数据路径的安全性。在这种示例中，可以至少部分地使用数据路径的安全性来确定输入设备特性3625-1和输入设备3652-1的信任度级别。要注意的是，描绘了单个接口3630。然而，在一些示例中，可以针对不同的输入设备3652或具有不同特性(例如，不同的安全性要求等)的输入设备3652提供不同的接口3630。

[0260] 电路3610可以执行指令3628，以基于相应的输入设备特性3625-1和3625-2来选择

最适合于利用输入数据3624-1或3624-2来训练的未经训练的推理模型3621之一。如所指出的，在各种示例中，电路3610可以执行指令3628以从多个输入设备(例如，输入设备3652-1和3652-2等)接收输入数据和/或输入设备特性。在这种示例中，电路3610可以执行指令3628，以通过将来自一个或多个输入设备的数据与期望输入设备特性进行组合，来建立针对特定未经训练的推理模型的训练数据集。尽管描绘了两个输入设备3652-1和3652-2，但是实际上，大量的输入设备3652(例如，输入设备3652的阵列、集群等)可以耦合到强化的系统3601，并且用于训练未经训练的推理模型3621。

[0261] 图37图示了逻辑流程3700。逻辑流程3700可以表示由本文中描述的一个或多个设备执行的一些或全部操作。例如，强化的系统3601(或其组件)可以执行逻辑流程3700中的操作，以标识要在来自输入设备的输入数据上训练的未经训练的推理模型。

[0262] 逻辑流程3700可以在框3710“确定输入设备特性”处开始，其中强化的系统3601可以确定输入设备3652-1的输入设备特性3625-1、输入设备3652-2的输入设备特性3625-2等。例如，强化的系统3601可以经由API来查询输入设备3652-1以确定输入设备特性3625-1。作为另一个示例，强化的系统3601可以与接收输入数据3624-2结合地从输入设备3652-2接收输入设备特性3625-2。

[0263] 继续到框3720“基于输入设备特性来选择未经训练的推理模型”，强化的系统3601可以基于输入设备特性3625-1、3625-2等从一个或多个未经训练的推理模型3621中进行选择。例如，强化的系统3601可以基于输入设备特性3625-1来选择要利用输入数据3624-1来训练的未经训练的推理模型3621-2。作为另一个示例，强化的系统3601可以基于输入设备特性3625-2来选择要利用输入数据3624-2来训练的未经训练的推理模型3621-1。

[0264] 进行到框3730“利用来自输入设备的输入数据来训练基于输入设备特性而选择的推理模型”，强化的系统3601可以利用输入数据(例如，输入数据3624-1、输入数据3624-2等)来训练从未经训练的推理模型121中选择的推理模型。例如，强化的系统3601可以基于输入设备特性3625-1利用输入数据3624-1来训练未经训练的推理模型3621-2，并且可以基于输入设备特性3625-2利用输入数据3624-2来训练未经训练的推理模型3621-1。从框3730，逻辑流程3700可以结束，或者可以基于接收到新的输入而重复。

[0265] 图38图示了可以根据本公开来实现的示例强化的系统3801。强化的系统3801包括电路3810、存储器3820、接口3630和输入设备3652-1和3652-2。存储器3820存储一个或多个经训练的推理模型3623-1、3623-2、3623-*N*(或经训练的推理模型3623)、输入数据3824-1和3824-2、输入设备特性3625-1和3625-2和指令3828。在许多示例中，输入数据3824-1和3824-2可以包括用于由一个或多个经训练的推理模型3623进行分类的数据。在操作期间，电路3810可以执行指令3828，以分别基于输入设备特性3625-1和3625-2来选择用于对输入数据3824-1和3824-2进行分类的经训练的推理模型3623之一。

[0266] 例如，在操作期间，电路3810可以执行指令3828以确定输入设备3652-1的一个或多个输入设备特性3625-1、以及确定输入设备3652-2的一个或多个输入设备特性3625-2。有时，一个或多个输入设备特性3625可以通过诸如在启动时经由接口3630查询相应的输入设备3652来确定。例如，输入设备3652-1可以具有API，经由该API，可以确定一个或多个输入设备特性3625-1。在其他时间，输入设备3652可以与将输入数据3624的一个或多个部分提供给存储器3820结合地提供一个或多个输入设备特性3625。例如，输入设备3652-2可以

与提供输入数据3624-2结合地提供输入设备特性3625-2。

[0267] 有时，电路3810可以执行指令3828，以基于存储器3820利用其耦合到输入设备152的接口3630、或基于接口3630通过其与输入设备3652进行通信的通信方案来标识一个或多个输入设备特性3625。例如，可以至少部分地基于接口3630来确定输入设备3652-1与存储器3820之间的数据路径的安全性。在这种示例中，可以至少部分地使用数据路径的安全性来确定输入设备3652-1和输入设备特性3625-1的信任度级别。要注意的是，描绘了单个接口3630。然而，在一些示例中，可以为不同的输入设备3652或具有不同特性(例如，不同的安全性要求等)的输入设备3652提供不同的接口3630。在一些示例中，输入设备特性和/或信任度可以用于确定输入来源质量项或基于输入来源质量项。在许多示例中，输入来源质量项可以是在选择推理模型(诸如，被调整成以所标识的信任度和/或输入来源质量项来工作的推理模型)时的因素。

[0268] 电路3810可以执行指令3828，以基于输入设备特性3625-1和3625-2来选择最适合对输入数据3824-1和3824-2进行分类的经训练的推理模型3623之一。如所指出的，在各种示例中，电路3810可以执行指令3828以从多个输入设备(例如，输入设备452-1和452-2等)接收输入数据和/或输入设备特性。在这种示例中，电路3810可以执行指令3828，以基于从其接收到了输入数据3824的相应输入设备3652的输入设备特性3625来选择用于对输入数据3824进行分类的经训练的推理模型3623。

[0269] 图39图示了逻辑流程3900。逻辑流程3900可以表示由本文中描述的一个或多个设备执行的一些或全部操作。例如，强化的系统3801(或其组件)可以执行逻辑流程3900中的操作，以基于一个或多个输入设备特性来标识要对来自输入设备的输入数据进行分类的经训练的推理模型。

[0270] 逻辑流程3900可以在框3910“确定输入设备特性”处开始，其中强化的系统3801可以分别确定输入设备3652-1和3652-2的输入设备特性3625-1和3625-2。例如，强化的系统3801可以经由API来查询输入设备3652-1以确定输入设备特性3625-1。作为另一个示例，强化的系统3801可以与接收输入数据3824-2结合地接收输入设备特性3625-2。在一些示例中，输入设备特性3625-1、3625-2可以包括镜头特性和音频放大器中的一个或多个。

[0271] 继续到框3920“基于输入设备特性来选择经训练的推理模型”，强化的系统3801可以基于输入设备特性3625(例如3625-1、3625-2等)从一个或多个经训练的推理模型3623中进行选择。例如，强化的系统3801可以基于输入设备特性3625-2来选择用于对输入数据3824-2进行分类的经训练的推理模型3623-1。作为另一个示例，强化的系统3801可以基于输入设备特性3625-1来选择用于对输入数据3824-1进行分类的经训练的推理模型3623-2。

[0272] 进行到框3930“利用基于输入设备特性而选择的推理模型对来自输入设备的输入数据进行分类”，强化的系统3801可以利用从经训练的推理模型123中选择的推理模型对输入数据3824进行分类。例如，强化的系统3801可以利用基于输入设备特性3625-2的经训练的推理模型3623-1对输入数据3824-2进行分类。作为另一个示例，强化的系统3801可以利用基于输入设备特性3625-1的经训练的推理模型3623-2对输入数据3824-1进行分类。从框3930，逻辑流程3900可以结束，或者可以基于接收到新的输入而重复。

[0273] *训练集与模型的可信关联*

图40图示了可以根据本公开来实现的示例强化的系统4001。强化的系统4001包括电路

4010、存储器4020和可信执行环境(TEE)4070。存储器4020存储输入数据4024、指令4028和输出数据4026。在若干个示例中,输入数据4024可以包括训练数据集。在操作期间,电路4010可以执行指令4028以执行与TEE 4070相关联的各种输入/输出操作。TEE 4070包括可信电路4011和可信存储器4021。可信存储器存储TEE指令4072、经转码的输入数据4074、训练数据集完整性状态(TSIS)4076、未经训练的推理模型4078、经训练的推理模型4080和推理模型完整性签名(IMIS)4082。在许多示例中,经转码的输入数据4074可以包括经转码的训练数据集。在操作期间,可信电路4011可以执行TEE指令4072,以启用关于在指定训练数据集上训练了推理模型的验证。

[0274] 如先前所提及,针对推理系统的攻击可以包括在推理模型的训练期间创建后门(例如,隐藏类)。有时,恶意行为者(例如,流氓数据科学家或数据供应者)可能将不适当的数据注入到训练数据集中,其意图是在经训练的推理模型中产生隐藏漏洞。因此,本公开提供了强化的系统4001,该强化的系统4001被布置成启用关于所声称的训练数据集训练了推理模型的验证。

[0275] 例如,在操作期间,电路4010可以执行指令4028以使得输入数据4024可用于TEE 4070。通常,输入数据4024可以包括具有多个样本的训练数据集。可信电路4011可以执行TEE指令4072,将输入数据4024转换成经转码的输入数据4074。例如,输入数据4024中的每个样本可以被变换成数学表示。可信电路4011可以执行TEE指令4072,以基于经转码的输入数据4074来生成TSIS 4076。例如,可以利用对次序和重复不敏感的密码算法,以基于经转码的输入数据4074中的每个样本来生成TSIS 4076。在许多示例中,训练数据集的内容可以由第三方利用TSIS 4076来验证(例如,通过重新计算TSIS 4076)。

[0276] 此外,可信电路4011可以执行TEE指令4072,以在经转码的输入数据4074上训练未经训练的推理模型4078以生成经训练的推理模型4080。例如,未经训练的推理模型4078可以包括机器学习算法,用以基于经转码的输入数据4074来生成经训练的推理模型4080。在一些示例中,未经训练的推理模型4078可以包括训练数据。在各种示例中,在生成了经训练的推理模型4080之后,未经训练的推理模型4078可以不保留在存储器中。

[0277] 可信电路4011可以执行TEE指令4072以生成推理模型完整性签名(IMIS)4082。例如,IMIS 4082可以包括经训练的推理模型4080的密码哈希。在许多示例中,推理模型的内容可以由第三方利用IMIS 4082来验证(例如,通过重新计算IMIS 4082)。可以提供TSIS 4076、IMIS 4082和经训练的推理模型4080中的一个或多个作为输出(例如,存储为输出数据4026)。例如,可信电路4011可以执行TEE指令4072以将TSIS 4076、IMIS 4082和经训练的推理模型4080输出为元组。有时,电路4010可以执行指令4028以将输出数据4026提供给输出设备、另一计算设备等。

[0278] 图41图示了逻辑流程4100。逻辑流程4100可以表示由本文中描述的一个或多个设备执行的一些或全部操作。例如,强化的系统4001(或其组件)可以执行逻辑流程4100中的操作,以生成TSIS 4076、IMIS 4082和经训练的推理模型4080中的一个或多个,从而启用对经训练的推理模型4080和/或与经训练的推理模型4080相对应的所声称的训练集的验证。

[0279] 逻辑流程4100可以在框4110“接收输入数据”处开始,其中强化的系统4001可以接收输入数据4024。例如,强化的系统4001可以接收具有多个样本的训练数据集作为输入数据4024。继续到框4120“对输入数据进行转码”,强化的系统4001可以对输入数据4024进行

转码以生成经转码的输入数据4074。例如，经转码的输入数据4074可以包括输入数据4024中的每个样本的数字表示。接下来，逻辑流程4100可以进行到框4130和/或框4140。换句话说，可以使用经转码的输入数据4120并行地执行框4130和框4140。

[0280] 在框4130“基于经转码的输入数据来生成TSIS值”处，可以基于经转码的输入数据来生成TSIS值。例如，强化的系统4001可以使用密码算法基于经转码的输入数据4074来生成TSIS值4074。逻辑流程4100可以从框4130进行到框4160。参考回框4140“利用经转码的输入数据来训练推理模型”，可以基于经转码的输入数据来训练推理模型。例如，强化的系统4001可以利用机器学习算法，用以基于经转码的输入数据4074来生成经训练的推理模型4080。继续到框4150“基于推理模型来生成IMIS值”，可以基于经训练的推理模型4080来生成IMIS值。例如，密码散列可以用于基于经训练的推理模型4080来生成IMIS值4082。从框4160，逻辑流程4100可以结束，或者可以基于接收到新的输入而重复。

[0281] *小节IV-增强的推理模型训练*

*在恶意训练之后的推理模型恢复*

如本文中详细描述的，推理模型是使用数据集(例如，训练和测试数据集)来训练的。通常，可以使用不同的数据集来重复地训练推理模型。例如，推理模型可以在第一数据集上训练，并且然后随后在第二数据集上训练。在一些实现方式中，训练数据集源自于不受控制的来源(例如，研究机构、开源、在线数据集等)。作为结果，推理模型可能是利用包括使该推理模型的性能退化的数据的数据集来训练的。例如，推理模型可能是利用包括被错误分类的数据(例如，被有意地错误分类，被意外地错误分类等)的数据集来训练的，这会使该推理模型的性能退化。在一些实例中，恶意方可能会有意地修改训练中使用的数据集，其意图是将后门、隐藏类或已知的错误分类漏洞引入到推理模型。

[0282] 本公开提供了有意地训练“通配符”类，该“通配符”类与使推理模型的性能退化的训练数据集的发现相对应。可以基于训练数据集内的错误来训练(一个或多个)这种通配符类，以恢复退化的性能并且缓解潜在的对抗性攻击。

[0283] 图42图示了用以训练推理模型的示例系统4201。系统4201包括电路4210、存储器4220和接口4230。存储器4220存储推理模型4222、指令4228和测试/训练数据集4260。存储器4220还可以可选地包括经恢复的推理模型4223。测试/训练数据集4260包括训练数据4262以及测试数据4264两者。如上所指出，可以使用多个不同的训练和/或测试数据集来按阶段实现训练。例如，描绘了数据集4260，其包括训练数据集4262-1、训练数据集4262-2和训练数据集4262-3、以及测试数据集4264-1和测试数据集4264-1。要注意的是，数据集4260可以包括任何数量的训练数据集4262和任何数量的测试数据集4264。选择此处描绘的数量是出于清楚而非限制的目的。此外，存储器4220可以包括恢复测试/训练数据集4270，恢复测试/训练数据集4270包括恢复训练数据集4272和恢复测试数据集4274。恢复训练数据集4272和恢复测试数据集4274可以用于从使用包括错误(例如，恶意引入的错误、意外引入的错误等)的训练数据集4262之一来训练推理模型4222中“恢复”。更特别地，恢复训练数据集4272可以用于利用附加类(诸如例如，通配符类)来训练推理模型4222，以基于训练数据集4262中的错误来缓解潜在的对抗性攻击。

[0284] 图43图示了示例技术4300，该技术可以由诸如图42的系统4201之类的系统来实现，以从利用包括错误的数据集来训练推理模型中恢复。要注意的是，为了方便和清楚起

见，参考图42的系统4201描述了技术4300。然而，这并不意图是限制性的。

[0285] 技术4300可以在圆圈43.1处开始。在圆圈43.1处，系统4201使用训练数据集4262-1来训练推理模型4222。例如，电路4210可以执行指令4228以实现推理模型训练算法(例如，基于图3的技术300等)以使用训练数据集4262-1来训练推理模型4222。使用训练数据集4262-1来训练推理模型4222可以导致推理模型4222-1的迭代。在圆圈43.2处，系统4201可以使用测试数据集(例如，测试数据集4264-1等)来测试推理模型迭代1 4222-1。例如，电路4210可以执行指令4228以使用测试数据集4264-1来测试推理模型迭代1 4222-1的性能。

[0286] 可以利用附加训练数据集来进一步训练推理模型4222。在圆圈43.3处，系统4201使用训练数据集4262-2来训练推理模型4222。例如，电路4210可以执行指令4228以实现推理模型训练算法(例如，基于图3的技术300等)，以使用训练数据集4262-2来训练推理模型迭代1 4222-1。使用训练数据集4262-2来训练推理模型迭代1 4222-1可以导致推理模型4222的另一迭代(例如，推理模型迭代2 4222-2)。在圆圈43.4处，系统4201可以使用测试数据集(例如，测试数据集4264-1等)来测试推理模型迭代2 4222-2。例如，电路4210可以执行指令4228以使用测试数据集4264-1来测试推理模型迭代2 4222-2的性能。

[0287] 可以再次利用第三附加训练数据集来进一步训练推理模型4222。在圆圈43.5处，系统4201使用训练数据集4262-3来训练推理模型4222。例如，电路4210可以执行指令4228以实现推理模型训练算法(例如，基于图3的技术300等)，以使用训练数据集4262-2来训练推理模型迭代2 4222-2。使用训练数据集4262-2来训练推理模型迭代2 4222-2可以导致推理模型4222的另一迭代(例如，推理模型迭代3 4222-3)。在圆圈43.6处，系统4201可以使用测试数据集(例如，测试数据集4264-2等)来测试推理模型迭代3 4222-3。例如，电路4210可以执行指令4228，以使用测试数据集4264-2来测试推理模型迭代3 4222-3的性能。

[0288] 如所指出的，在某些真实世界用例的情况下，在训练之后可能发生对训练数据集内的错误的发现。在圆圈43.7处，可能发现训练数据集(例如，训练数据集4262-1、训练数据集4262-2和/或训练数据集4262-3)内的错误。在一些示例的情况下，在用于训练推理模型的过去迭代的训练数据集中发现了错误。作为特定示例，可能在训练数据集4262-2中发现错误。由此，由于训练数据集4262-2中的错误，推理模型4222内的错误(或漏洞)可能已经如在推理模型迭代2 4222-2中被引入。在圆圈43.8处，可以接收恢复测试/训练数据集4270(包括恢复训练数据集4272和恢复测试数据集4274)。电路4210在执行指令4228时可以经由接口4230等来接收恢复测试/训练数据集4270。在一些示例的情况下，可以将恢复训练数据集4272设计成利用附加的通配符类来训练推理模型4222(从当前迭代处开始)，该通配符类可以用于基于先前使用的训练数据集4262中的错误来缓解对抗性攻击。

[0289] 在圆圈43.9处，系统4201使用恢复训练数据集4272来训练推理模型4222。例如，电路4210可以执行指令4228以实现推理模型训练算法(例如，基于图3的技术300等)以使用恢复训练数据集4272来训练推理模型迭代3 4222-3。使用恢复训练数据集4272来训练推理模型迭代3 4222-3可以导致推理模型4222的另一迭代(例如，经恢复的推理模型4223等)。在圆圈43.10处，系统4201可以使用恢复测试数据集4274来测试经恢复的推理模型4223。例如，电路4210可以执行指令4228以使用恢复测试数据集4274来测试经恢复的推理模型4223的性能。

[0290] 系统4201和技术4300提供的优点在于：可以实现对对抗性攻击的缓解(例如，警报

触发等)。此外,系统4201和技术4300提供了如下优点:即不需要从开头或从所捕获的实例(例如,迭代等)来重新开始训练。

[0291] 如本文中所指出,测试/训练数据用于“训练”和评估推理模型的性能。然而,测试/训练数据集可能包括错误。例如,图44描绘了可以用于训练推理模型4222的示例测试/训练数据集4460。测试/训练数据集4460包括训练数据集4462和测试数据集4464,它们可以对应于图42的系统4201的训练数据集4262和测试数据集4264之一。测试/训练数据集4460包括输入4461以及预期输出4465。通常,输入4461可以包括(一个或多个)输入值4463-n,其中n是正整数。例如,针对输入4461,示出了(一个或多个)输入值4463-1、(一个或多个)输入值4463-2、(一个或多个)输入值4463-3、(一个或多个)输入值4463-4、(一个或多个)输入值4463-5至(一个或多个)输入值4463-N。针对每个输入值4463-n,示出了相关联的预期输出4465。预期输出4465中的每一个对应于类4467,推理模型4222要被训练成将输入(例如,输入4461)分类到类4467。通常,预期输出4465可以包括任何数量的类4467。作为特定示例,类4467可以对应于来自CIFER-10分类数据集的类别(例如,飞机、汽车、鸟、猫、鹿、狗、青蛙、马、船和卡车)。然而,示例在该情境中不受限制。在某些用例中,测试/训练数据集(例如4460)可能包括错误4469。如所描绘的,数据集4460包括多个错误4469。作为特定示例,错误4469之一示出了对应于类4467-2的(一个或多个)输入值4463-4。为了解释,假定推理系统4222是图像分类器,其被布置成将图像分类到类。在该示例下,错误4469可能是:与输入值4463-4相关联的图像被不正确地标记或加标签为对应于错误的类(例如,该图像是枪但是被加标签为发梳等)。如上所指出,这种不正确的标签可能是恶意的或无意的。然而,本公开所提供的是使用如上详细描述的恢复测试/训练数据集继续进行训练,尽管测试/训练数据集4460中存在错误。

[0292] 图45描绘了可以用于训练推理模型4222的示例恢复测试/训练数据集4570。恢复测试/训练数据集4570包括训练数据集4572和测试数据集4574,它们可以对应于图42的系统4201的恢复训练数据集4272和恢复测试数据集4274。参考图44的测试/训练数据集4460对恢复测试/训练数据集4570进行描述。然而,示例在该情境中不受限制。

[0293] 恢复测试/训练数据集4560包括输入4461以及预期输出4465。通常,输入4461可以包括(一个或多个)输入值4573-m,其中m是正整数。此外,输入4461可以包括与错误4469相关联的(一个或多个)任何输入值。例如,与来自测试/训练数据集4460的错误4469相关联的(一个或多个)输入值4463-4被包括在恢复测试/训练数据集4570的输入4461中。此外,针对输入4461,示出了(一个或多个)输入值4573-1、(一个或多个)输入值4573-2、(一个或多个)输入值4573-3、(一个或多个)输入值4573-4至(一个或多个)输入值4573-M。

[0294] 针对每个输入值4463或4573,示出了相关联的预期输出4465。如上所描述,恢复测试/训练数据集4570被设计或布置成训练推理模型4222,以将图像识别成特定的通配符类4577。该通配符类4577可以是推理模型4222被训练以分类到的新类。作为替代方案,通配符类4577可以是被重新运用(repurpose)作为通配符类4577的类中的一个(例如,图44的类4467等)。例如,使用上面提供的假设——被加标签为看起来像发梳的枪,可以将通配符类4577引入为“看起来像发梳的枪”,并且图像提供输入值4573-4来训练推理模型4222以将输入分类成该新的通配符类4577。

[0295] 图46图示了逻辑流程4600。逻辑流程4600可以表示由本文中描述的一个或多个设

备执行的一些或全部操作。例如，系统4201(或其组件)可以执行逻辑流程4600中的操作，以在利用包括错误的测试/训练数据集4260(例如，包括错误4469的数据集4460等)训练了推理模型4222之后进行恢复。

[0296] 逻辑流程4600可以在框4610“接收测试/训练数据集”处开始，其中系统4201可以接收测试/训练数据集4260。例如，电路4210在执行指令4228时可以接收训练数据集4262-1(和测试数据集4264-1等)。继续到框4620“基于测试/训练数据集来训练推理模型”，系统4201可以基于在框4610处接收到的测试/训练数据集4260来训练推理模型4222。例如，电路4210在执行指令4228时可以实现推理模型训练算法(例如，图3的技术300等)以使用在框4610处接收到的测试/训练数据集4260(例如，训练数据集4262-1、测试数据集4264-1)来训练推理模型4222等。

[0297] 继续到判定框4630“在测试/训练数据集中是否检测到错误”，可以确定在框4610处接收到的并且在框4620处用于训练推理模型4222的测试/训练数据集4260(例如，训练数据集4262-1等)中是否存在错误。从判定框4630，逻辑流程4600可以继续到判定框4640或框4650。逻辑流程4600可以基于在框4610处接收到的并且在框4620处用于训练推理模型4222的测试/训练数据集4260中不存在错误的确定而从判定框4630继续到判定框4640。可替代地，逻辑流程4600可以基于在框4610处接收到的并且在框4620处用于训练推理模型4222的测试/训练数据集4260中存在错误的确定而从判定框4630继续到框4650。

[0298] 在框4650“接收恢复测试/训练数据集”处，系统4201可以接收恢复测试/训练数据集4270。例如，电路4210在执行指令4228时可以接收恢复训练数据集4272(和恢复测试数据集4274等)。继续到框4660“基于恢复测试/训练数据集来训练推理模型”，系统4201可以基于在框4650处接收到的恢复测试/训练数据集4270来训练推理模型4222。例如，电路4210在执行指令4228时可以实现推理模型训练算法(例如，图3的技术300等)，以使用在框4610处接收到的恢复测试/训练数据集4270(例如，训练数据集4272、测试数据集4274等)来训练推理模型4222。具体地，可以在框4660处训练推理模型4222，以将输入分类成通配符类4577。

[0299] 从框4660，逻辑流程可以继续到判定框4640。在判定框4640“是否接收到附加的测试/训练数据集”处，可以确定是否已经接收到附加的测试/训练数据集。例如，电路4210在执行指令4228时可以确定是否已经接收到附加的测试/训练数据集(例如，训练数据集4262-2、训练数据集4262-3、测试数据集4274-2等)。从判定框4640，逻辑流程4600可以返回到框4620，例如以使用新接收到的测试/训练数据集4260来进一步训练推理模型4222；或者逻辑流程4600可以结束。逻辑流程4600可以基于接收到了附加的测试/训练数据集4260的确定而从判定框4640返回到框4620。可替代地，逻辑流程4600可以基于没有接收到附加的测试/训练数据集4260的确定而结束。

[0300] *用于安全模型执行的对抗性训练*

图47图示了可以根据本公开来实现的示例强化的系统4701。强化的系统4701包括电路4710、存储器4720、接口4730、输入设备4752和可信执行环境(TEE)4770。存储器4720存储公共推理模型分区4722-PUB、输入数据4724、中间输出数据4723、输出数据4726和指令4728。TEE 4770存储私有推理模型分区4722-PRIV和TEE指令4778。通常，TEE 4770可以是系统4701内的任何可信执行飞地(enclave)或分区。例如，TEE 4770可以是电路4710内的可信执行分区。作为另一个示例，TEE 4770可以包括以可信的方式从电路4710和存储器4720中分

割的其自己的电路和存储器(未示出)。通常,TEE 4770可以提供安全性特征,诸如隔离执行(例如,对TEE指令4778的隔离执行)、以及私有推理模型分区4722-PRIV的保密性。通常,系统4701可以以安全的方式(例如,在TEE 4770内)执行推理模型的一部分(例如,私有推理模型分区4722-PRIV),而推理模型的其余部分(例如,公共推理模型分区4722-PUB)在不安全的计算资源(例如,电路4710和存储器4720等)中执行。

[0301] 在一些示例的情况下,可以利用对抗性推理模型来训练公共推理模型分区4722-PUB,以使得公共推理模型分区4722-PUB解析到不准确或不太准确的局部最小值。换句话说,公共推理模型分区4722-PUB在由其自己来执行时解析到不准确或不太准确的局部最小值。然而,完整的推理模型(例如,公共推理模型分区4722-PUB+私有推理模型分区4722-PRIV)解析到正确的局部最小值。

[0302] 这提供的优点在于:推理模型是受保护的,这是由于该模型的一部分以安全的方式被执行并且以保密的方式被维持。要注意的是,这提供了相比于简单地在安全环境(TEE)中执行整个推理模型的优点。例如,许多安全环境具有有限的存储器和计算资源。由此,许多现代推理模型无法完全地在安全环境内执行。要注意的是,在一些示例中,可以使用除TEE 4770以外的系统或技术来保护私有推理模型分区4722-PRIV。例如,可以诸如使用同态加密、多方计算或其他这种加密执行技术将4722-PRIV移植(port)成加密形式。在这种示例中,代替于TEE 4770的有限存储器空间,加密的模型与未加密的执行相比可能需要更多的计算资源来执行并且花费更长的运行时间。应当领会的是,尽管本文中提供的示例使用TEE 4770,但是可以针对使用除TEE 4770以外的手段(例如,加密执行、多方执行等)而保护的私有推理模型分区(例如4722-PRIV)来实现所公开的示例。

[0303] 本公开提供了提供推理模型的安全性的系统和技术,其中推理模型的一部分被维持在安全环境中。此外,即使对抗方会获得公共推理模型分区4722-PUB,但是由于公共推理模型分区4722-PUB解析到不准确或不太准确的局部最小值,对抗方也将无法再现推理模型4722的整个功能。

[0304] 图48图示了分区推理系统4822的示例,该系统可以根据本公开的示例来实现(例如,实现为图1的强化的系统4701的推理系统4722)。图49图示了示例技术4900,该技术可以由强化的系统(例如,图1的强化的系统4701)来实现从而以安全的方式来执行推理模型。参考图48和49描述了图1的强化的系统4701的操作。

[0305] 更特别地转向图48,描绘了推理模型4822(或分区推理模型)。推理模型4822包括经由连接4893而连接的多个节点4891。推理模型4822可以是多种推理模型中的任一个,诸如例如神经网络等。通常,节点4891经由连接4893来接收(一个或多个)输入,并且基于激活函数来导出输出。可以使用多种激活函数中的任一个(例如,恒等、二元阶跃、正切、反正切、S形、逻辑、或软S形、高斯等)。此外,通常在每个连接4893处通过特定于连接的常数(被称为权重)来对值进行缩放。

[0306] 节点4891的群组可以被称为层4890。例如,描绘了层4890-1、层4890-2、层4890-3、层4890-4和层4890-5。要注意的是,推理模型可以包括任何数量的节点和层、以及每层节点的数量等。在实践中,推理模型通常将包括比此处描绘的更多的节点和层。此处描绘的网络架构仅仅是为了呈现的清楚而完成的,并且不是限制性的。

[0307] 推理模型4822被分区成公共推理模型分区4822-PUB和私有推理模型分区4822-

PRIV，其中所选的层被包括在分区之一中。例如，公共推理模型分区4822-PUB包括层4890-1、4890-2和4890-3，而私有推理模型分区4822-PRIV包括层4890-4和4890-5。要注意的是，私有分区或在TEE 4770中维持的分区可以包括推理模型4822的较早或初始的层，而不是所描绘的稍后的层。示例在该情境中不受限制。

[0308] 更特别地转向图49，技术4900可以在圆圈49.1处开始。在圆圈49.1处，强化的系统可以接收输入数据4724。例如，系统4701的电路4710可以执行指令4728以接收输入数据4724。例如，电路4710可以从输入设备4752接收输入数据4724。作为另一个示例，电路4710可以经由接口4730从另一计算设备接收输入数据4724。

[0309] 继续到圆圈49.2，强化的系统4701可以执行公共推理模型分区4722-PUB以从输入数据4724生成中间输出数据4723。例如，电路4710在执行指令4728时可以基于与公共推理模型分区4722-PUB的层4890相关联的计算、使用输入数据4724作为去往该分区的第一层4890的输入来导出中间输出数据4723。作为特定示例，中间输出数据4723可以对应于来自层4890-3(即，公共推理模型分区4722-PUB的最后一层)的节点的输出。继续到圆圈BD3.3，强化的系统4701的TEE 4770可以接收或访问中间输出数据4723。例如，TEE 4770在执行TEE指令4778时可以(例如，从存储器4720等)访问中间输出数据4723。

[0310] 继续到圆圈49.4，强化的系统4701可以以安全的方式来执行私有推理模型分区4722-PRIV，以从中间输出数据4723生成输出数据4726。例如，TEE 4770在执行TEE指令4778时可以基于与私有推理模型分区4722-PRIV的层4890相关联的计算、使用中间输入数据4723作为去往该分区的第一层4890的输入来导出输出数据4726。作为特定示例，输出数据4726可以对应于来自层4890-5(即，私有推理模型分区4722-PRIV的最后一层)的节点的输出。

[0311] 图50图示了可以被实现为训练分区推理模型的系统5001。图51图示了逻辑流程5100的示例。逻辑流程5100可以表示由本文中描述的一个或多个示例执行以训练分区推理模型(例如，推理模型4722、推理模型4822等)的一些或全部操作。例如，系统5001(或其组件)可以执行逻辑流程4600中的操作，以生成用于推理模型的私有和公共分区。系统5001和逻辑流程5100被彼此结合地描述并且还参考图48的分区推理模型4822来描述。然而，示例在该情境中不受限制。

[0312] 更特别地转向图50，系统5001可以包括电路5010、存储器5020和接口5030。存储器5020存储包括私有推理模型分区5022-PRIV和公共推理模型分区5022-PUB的推理模型5022。存储器5020进一步包括性能特性5027、指令4228、性能预算5029和对抗性训练模型5080。

[0313] 更特别地转向图51，逻辑流程5100可以在框5110处开始。在框5110“接收性能预算”处，系统5001可以接收性能预算5029。例如，电路5010在执行指令5028时可以基于用户选择来接收性能预算。在一些示例中，性能预算5029可以包括对性能退化(诸如例如，总体运行时间退化)的可接受水平的指示。作为特定示例，性能预算5029可以包括关于高达20％的性能惩罚是可接受的指示。

[0314] 继续到框5120“将层L*i*添加到私有分区/从公共分区中移除”，系统5001可以将层Li添加到私有模型分区5022-PRIV。在一些示例的情况下，将层分组到私有和公共分区中可以在推理模型的最后一层(例如，最接近输出的层)处开始。例如，针对具有5个层4890的推

理模型(例如,推理模型4822等),*i*可以从“5”开始。使用该示例,在框520处,电路5010可以执行指令5028以将层5090-5添加到私有推理模型分区5090-5,并且将层5090-5从公共推理模型分区5022-PUB中移除。

[0315] 继续到框5130“评估用于私有推理模型分区的性能成本”,系统5001可以评估用于在安全环境(例如,TEE 4770等)中执行私有推理模型分区DBV422-PRIV的性能成本。在一些示例中,TEE 4770可以包括暴露TEE的计算能力(例如,存储器空间、每(一个或多个)操作的性能影响、页面大小、高速缓存大小等)的应用编程接口(API)。因此,在一些示例中,在框5130处(或在框5120之前),逻辑流程5100可以确定用于在TEE 4770中执行推理模型的性能度量。在一些示例的情况下,在框5130处(或在框5120之前),逻辑流程5100还可以使用电路4710和存储器4720来确定用于执行推理模型的性能度量。例如,电路5010可以执行指令5028来查询TEE(例如,TEE 4770),以确定与TEE相关联的性能度量,并且可以将这种性能度量的指示保存为性能特性5027。在框BD530处,电路5010可以执行指令5028,以基于性能特性5027以及推理模型5022的层4890的架构来评估用以在TEE中执行私有推理模型分区5022-PRIV(或被添加到私有推理模型分区5022-PIRV的层)的计算性能中的成本。

[0316] 继续到判定框5140“是否超过性能预算”,可以通过在安全环境中(例如,在TEE 4770中)执行私有推理模型分区5022-PRIV来做出关于是否超过性能预算5029的确定。例如,假定私有推理模型分区5022-PRIV包括推理模型4822的层4890-5,并且在TEE 4770中执行层4890-5引入了10%的性能惩罚。进一步假定性能预算5029指示小于20%的可接受的性能退化。使用这种示例,包括层4890-5的私有推理模型分区5022-PRIV没有超过性能预算5029。在一些示例的情况下,电路5010在执行指令5028时可以确定基于性能特性5027对私有推理模型分区5022-PRIV所估计的执行运行时间是否超过性能预算5029。

[0317] 从判定框5140,逻辑流程5100可以继续到判定框5150或判定框5160。逻辑流程5100可以基于未超过性能预算5029的确定而从判定框5140继续到判定框5150,而逻辑流程5100可以基于超过性能预算5029的确定而从判定框5140继续到判定框5160。

[0318] 在判定框5150“*i*是否大于1”处,确定*i*是否大于1,也就是说确定是否仍然存在公共推理模型分区5022-PUB中的层。从判定框5150,逻辑流程5100可以继续到框5170或框5190。逻辑流程5100可以基于*i*大于1的确定而从判定框5150继续到框5170,而逻辑流程5100可以基于*i*不大于1的确定而从判定框5150继续到框5190。

[0319] 在判定框5160“私有推理模型分区中是否有多于一个层”处,做出私有推理模型分区5022-PRIV是否包括多于一个层的确定。例如,电路5010可以执行指令5028,以确定私有推理模型5022-PRIV是否包括多于一个层。从判定框5160,逻辑流程5100可以继续到框5180或框5190。逻辑流程5100可以基于私有推理模型分区包括多于一个层的确定而从判定框5160继续到框5180,而逻辑流程5100可以基于私有推理模型分区不包括多于一个层的确定而从判定框5160继续到框5190。

[0320] 在框5170“使*i*递减”处,系统5001可以使*i*递减。例如,电路5010在执行指令5028时可以使*i*递减。从框5170,逻辑流程5100可以返回到框5120以评估并且可能地添加另一层到私有推理模型分区。在框5180“将层L*i*从私有分区中移除/添加到公共分区”处,系统5001可以将层L*i*从私有模型分区5022-PRIV中移除,并且将层L*i*添加到公共推理模型分区5022-PUB。

[0321] 在框5190“利用对抗性模型来训练公共推理模型分区”处，系统5001可以利用对抗性模型5080来训练公共推理模型分区5022-PUB。通常，利用对抗性模型5080来训练公共推理模型分区5022-PUB将引导公共推理模型分区5022-PUB在利用替代性子模型而不是该私有模型执行时达到不准确或不太准确的局部最小值。也就是说，对抗性损失会“教导”整个模型对由该私有模型中的层所创建的中间表示进行加密或争夺，以使得该私有模型用作密钥。

[0322] 图52描绘了示例推理模型5222，其被拆分成私有推理模型分区5222-PRIV和两个公共推理模型分区5222-PUB 1和5222-PUB 2，这两个公共推理模型分区5222-PUB 1和5222-PUB 2耦合到对抗性模型5280和替代性子模型5282以用于所描绘的对抗性训练。如所陈述的，训练进行操作以引导公共推理模型分区5222-PUB 1和5222-PUB2在利用替代性子模型5282-1和5282-2执行时达到不准确或不太准确的局部最小值。在训练期间，推理模型5222具有对模型分区(例如，公共推理模型分区5222-PUB 1、私有推理模型分区5222-PRIV和公共推理模型分区5222-PUB 2)的读取和写入访问权。此外，对抗性模型5280具有对对应替代性子模型5282的读取和写入访问权(例如，对抗性模型5280-1具有对替代性子模型5282-1的读取和写入访问权，而对抗性模型5280-2具有对替代性子模型5282-2的读取和写入访问权)。其余的连接(例如，虚线连接)是只读的。

[0323] 在一些示例中，主模型5222可以被提供有：更多的训练迭代、对对抗性模型5280-1和5280-2的权重的访问权(例如，读取访问权)(例如，以确定梯度从而确定对抗方在训练期间移动的“方向”)、更多的训练数据等。这种提供可以通过私有推理模型分区5222-PRIV来提供更好的安全性或更好的加密。

[0324] *小节V-隐私*

*用于边缘设备的隐私感知特征平衡器*

当前推理系统的一个问题是对象检测器(例如，分类器模型)被布置成报告所有检测到的对象，而不管隐私暗示。例如，当前的推理系统将对图像中的所有对象进行分类，并且将报告所有已分类的对象，从而将该系统的任何隐私暗示或强化留给环境的上层。本公开提供了推理环境，该推理环境可以被布置成筛查处于限定范围之外的对象的报告，例如以保护某些检测到的对象的隐私。这一点的示例可以是被布置成检测由相机捕获的图像中的对象的智能监控相机。可以将智能监控相机配置成过滤或筛查在指定范围之外的检测的报告。

[0325] 应当领会的是，法规或政府条例越来越要求对隐私的管理。例如，欧盟(EU)的通用数据保护条例(GDPR)规范了个体个人数据的导出和使用，并且给予个体对其个人信息使用的控制权。由此，在一些法律下，可能需要维持如本文中讨论的检测到的一些对象的隐私。

[0326] 图53图示了包括服务器5301和边缘设备5303的示例推理环境5300。通常，边缘设备5303可以是提供进入服务器5303中、或去往耦合到服务器5303的网络的入口点的任何设备。例如，边缘设备5303可以是路由器、路由交换机、集成接入设备(IAD)、多路复用器、小区节点、基站等。作为另一个示例，边缘设备5303可以是输入捕获设备，诸如例如相机、麦克风等。通常，服务器5301和边缘设备5303可以经由任何合适的互连(诸如，广域网、局域网或互联网)而耦合。服务器5301包括电路5310-1、存储器5320-1和接口5330-1。存储器5320-1存储指令5328-1、筛查能力5329、隐私级别规范5323和经隐私筛查的输出数据5327。

[0327] 边缘设备5303包括电路5310-3、存储器5320-3、接口5330-3和输入设备5352。存储器5320-3存储指令5328-3、输入数据5324、推理模型5322、输出数据5326、筛查能力5329、隐私级别规范5323和隐私筛查的输出数据5327。

[0328] 电路5310-3可以执行指令5328-3以例如从输入设备5352接收输入数据5324。附加地，电路5310-3可以执行指令5328-3，以从推理模型5322和输入数据5324生成输出数据5326。例如，电路5310-3可以在输入数据5324上执行推理模型5322以生成输出数据5326。在一些示例中，输入设备5352可以是被布置成捕获图像帧(例如，静止帧、视频帧等)的图像传感器。推理模型5322可以被布置成接收输入数据5324，并且从输入数据5324中检测对象(例如，在图像帧中表示的对象等)并且将对象进行分类。

[0329] 电路5310-1可以执行指令5328-1以查询边缘设备5303从而收集和/或确定边缘设备129的筛查能力5329。通常，筛查能力5329可以包括可由推理模型5322分类的对象的指示。例如，筛查能力5329可以包括可以从输入数据5324中检测和分类的对象、动作、场景、行为者等的指示。此外，电路5310-1可以执行指令5328-1以查询边缘设备5303从而部分地基于筛查能力5329来生成隐私级别规范5323，并且可以将隐私级别规范5323传送到边缘设备5303。例如，假定推理引擎被布置成对图像中的以下对象进行检测和分类:[枪、刀、炸弹、球、钢笔、电话、钥匙、岩石、人类、狗、猫];筛查能力5329可以包括推理引擎5322被布置成对其进行检测和分类的对象的指示。在给定筛查能力的情况下，电路5310-1在执行指令5328-1时可以生成隐私级别规范，该隐私级别规范包括对筛查(或过滤)除了枪、刀和炸弹之外的所有检测到的对象的指示。

[0330] 电路5310-3可以执行指令5328-3，以从输出数据5326和隐私级别规范5323来生成经筛查的输出数据5327。例如，电路5310-3在执行指令5328-3时可以从输出数据5326中过滤在隐私级别规范5323中指示的对象列表之外的所分类的对象。此外，电路5310-3可以执行指令5328-3，以将经筛查的输出数据5327提供给服务器5301。例如，电路5310-3可以经由接口5330-3将经筛查的输出数据5327提供给服务器5301。

[0331] 图54图示了示例技术5400，该技术可以在推理环境(诸如，环境5300)中实现，以筛查边缘处的输出从而增加推理环境中的隐私性。要注意的是，为了方便和清楚起见，参考图53的推理环境5300描述了技术5400。然而，这并不意图是限制性的。

[0332] 技术5400可以在圆圈54.1处开始。在圆圈54.1处，服务器5301可以查询边缘设备5303以确定边缘设备5303的筛查能力。例如，服务器5301可以经由接口5330-1向边缘设备5303发送接收筛查能力5329的请求。在圆圈54.2处，边缘设备5303可以将包括筛查能力5329的指示的信息元素发送到服务器5301。例如，响应于在圆圈54.1处接收到请求，边缘设备5303可以经由接口5330-3将筛查能力5329发送到服务器5301。而且，在圆圈2.2处，服务器5301可以经由接口5330-1从边缘设备5303接收筛查能力5329。

[0333] 在圆圈54.3处，服务器5301可以部分地基于筛查能力5329来生成隐私级别规范5323。例如，在给定边缘设备5303的筛查能力5329的情况下，服务器5301可以生成要过滤的分类类别的列表以缓解或抑制隐私侵入。在圆圈54.4处，服务器5301可以将包括隐私级别规范5323的指示的信息元素发送到边缘设备5303。例如，服务器5301可以经由接口5330-1将隐私级别规范5323发送到边缘设备5303。而且，在圆圈2.4处，边缘设备5303可以经由接口5330-3从服务器5301接收隐私级别规范5323。

[0334] 在圆圈54.5处，边缘设备5303可以接收输入数据5324。例如，边缘设备5303可以从输入设备5352接收输入数据5324。要注意的是，如本文中所描述，可以接收和筛查(或过滤)输入数据5324的多个不同实例。例如，边缘设备5303可以在圆圈54.5处重复地接收输入数据5324，并且生成如本文中所描述的经隐私筛查的输出5327。在圆圈54.6处，边缘设备5303可以从推理模型5322和输入数据5324生成输出数据5326。例如，边缘设备5303可以在输入数据5324上执行推理模型5322以生成输出数据5326。在圆圈54.7处，边缘设备5303可以从输出数据5326和隐私级别规范5323来生成经隐私筛查的输出数据5327。例如，边缘设备5303可以从输出数据5326中筛查、过滤或以其他方式移除将不被报告的对象(例如，在隐私级别规范5323中列出的对象、从隐私级别规范5323中省略的对象等)，并且以经隐私筛查的输出数据5327来保存过滤后的输出。

[0335] 在圆圈54.8处，边缘设备5303可以将包括经隐私筛查的输出数据5327的指示的信息元素发送到服务器5301。例如，边缘设备5303可以经由接口5330-3将经隐私筛查的输出数据5327发送到服务器5301。而且，在圆圈2.8处，服务器5301可以经由接口5330-1从边缘设备5303接收经隐私筛查的输出数据5327。

[0336] 在一些示例中，推理模型本身可以被修改成对输出进行筛查，如本文中所讨论的那样。例如，推理模型可以被修改成丢弃特定的输出。作为特定示例，连接了推理模型中的最后一层的特定权重可以被修改(例如，设置为零等)以使该推理模型不输出与那些权重相关联的类。图55图示了包括服务器5501和边缘设备5503的示例推理环境5500。通常，边缘设备5503可以是提供进入服务器5501中、或去往耦合到服务器5501的网络的入口点的任何设备。例如，边缘设备5503可以是路由器、路由交换机、集成接入设备(IAD)、多路复用器、小区节点、基站等。作为另一个示例，边缘设备5503可以是输入捕获设备，诸如例如相机、麦克风等。通常，服务器5501和边缘设备5503可以经由任何合适的互连(诸如，广域网、局域网或互联网)而耦合。服务器5501包括电路5510-1、存储器5520-1和接口5530-1。存储器5520-1存储指令5528-1和经隐私筛查的输出数据5527。

[0337] 边缘设备5503包括电路5510-3、存储器5520-3、接口5530-3和输入设备5552。存储器5520-3存储指令5528-3、输入数据5524、隐私筛查推理模型5523和隐私筛查的输出数据5527。隐私筛查推理模型5523可以是被调整成自动生成经隐私筛查的输出数据5527的推理模型。例如，可以通过调整推理模型中的权重或连接以使该推理模型输出与非私有类(或类似物)相关联的数据，来生成隐私筛查推理模型5523。使用上面的示例，推理模型可以(例如，通过调整连接等)被修改成生成隐私筛查推理模型5523，该隐私筛查推理模型5523输出枪、刀或炸弹的分类。

[0338] 电路5510-3可以执行指令5528-3以例如从输入设备5552接收输入数据5524。附加地，电路5510-3可以执行指令5528-3，以从隐私筛查推理模型5523和输入数据5524生成经隐私筛查的输出数据5527。例如，电路5510-3可以在输入数据5524上执行隐私筛查推理模型5523以生成经隐私筛查的输出数据5527。在一些示例中，输入设备5352可以是被布置成捕获图像帧(例如，静止帧、视频帧等)的图像传感器。推理模型5322可以被布置成接收输入数据5324，并且从输入数据5324中检测对象(例如，在图像帧中表示的对象等)并且将对象进行分类。

[0339] *隐私感知多位置模型执行*

本公开提供了采用分布式方式的推理模型执行的示例。例如，推理模型的一部分可以在耦合到服务器的边缘设备上执行。作为另一个示例，可以使用多个服务器以分布式方式来执行推理模型，也就是说，每个服务器执行该模型的一部分。另外，本公开提供了查问(interrogate)分布式计算平台(例如，服务器、边缘设备等)以确定该平台的安全性能力，并且然后基于所确定的安全性能力、以及推理模型的不同部分的安全性要求来拆分该推理模型并且将各部分(或切片)指派给平台。

[0340] 因此，可以缓解与推理模型的分布式执行相关联的风险。例如，推理模型的不同部分可能具有不同的值或隐私问题。作为特定示例，神经网络中的初始层(例如，靠近输入的层)通常会花费更多的训练迭代来恢复(例如，重新创建)；而稍后的层(例如，靠近输出的层)通常更容易暴露关于在其上训练了推理模型的数据的信息。因此，推理模型的稍后的层可以被拆分并且被指派给比分布式执行环境中的其他平台具有更高安全性能力的平台。

[0341] 图56图示了包括客户端设备5601以及服务器5603和5605的示例推理环境5600。通常，服务器5603和5605被布置成以分布式方式来执行推理模型的一部分。服务器5603和5605可以是基于云的(或可访问的)计算资源、边缘设备、专门布置的推理模型执行加速器等。此外，尽管在该图中描绘了两个服务器5603和5605，但是应当领会的是，可以在推理环境5600中提供任何数量(2个、3个、4个或更多)的服务器。如上所指出，推理环境5600的服务器5603和5605可以是边缘设备。在一些示例的情况下，推理环境5600可以是多边缘环境，其中客户端设备5601利用“边缘”(例如，边缘设备5603、5605等)中的空闲计算功率。

[0342] 客户端设备5601以及服务器5603和5605可以经由网络(例如，其可以包括互联网等)而耦合。更具体地，接口5630-1、5630-3和5630-5可以耦合到网络以便于设备之间的数据通信或交换。

[0343] 客户端设备5601包括电路5610-1、存储器5620-1和接口5630-1。存储器5620-1存储推理模型5622、输入数据5624、输出数据5626、模型安全性要求5627、指令5628和服务器能力5629。通常，客户端设备5601进行操作以基于模型安全性要求5627和服务器能力5629将推理模型5622拆分成模型切片5623，并且然后基于模型安全性要求5627和服务器能力5629将这些模型切片5623指派给服务器5603或5605之一。因此，推理模型5622可以由服务器5603和5605以分布式方式来执行，同时基于服务器的服务器能力5629和推理模型5622的安全性要求5627来提供对该模型的部分的安全性保护。

[0344] 服务器5603包括电路5610-3、存储器5620-3和接口5630-3。存储器5620-3存储模型切片5622-1和指令5628-3。同样地，服务器5605包括电路5610-5、存储器5620-5和接口5630-5。存储器5620-5存储模型切片5622-2和指令5628-5。通常，服务器5603和5605提供暴露相应服务器的安全性能力(例如，篡改保护、安全执行规定等)的应用编程接口(API)。此外，服务器5603和105被布置成以分布式方式来执行推理模型5622的部分(例如，模型切片5623)。

[0345] 电路5610-1可以执行指令5628-1以查询服务器5603和5605从而确定相应服务器中的每一个的安全性能力。响应于该查询，电路5610-3和5610-5可以分别执行指令5628-3和5628-5，以将相应服务器5603或5605的安全性能力发送到客户端设备5601。电路5610-1可以执行指令5628-1，以将从相应服务器接收到的安全性能力的指示存储为服务器能力5629。

[0346] 电路5610-1可以执行指令5628-1,以基于服务器能力和模型安全性要求5627将推理模型5622拆分成模型切片5623。通常,模型安全性要求5627可以包括针对推理模型5622的相应层、部分或部件的安全性要求的指示。电路5610-1可以执行指令5628-1,以基于模型安全性要求5627和服务器能力5629来确定哪个服务器5603或5605执行哪个模型切片5623。此外,电路5610-1可以执行指令5628-1,以(例如,经由接口5630-1等)向服务器5603和5605中的每一个发送如下信息元素:该信息元素包括所指派的模型切片5623的指示、以及以分布式方式来执行模型切片5623的指令。

[0347] 电路5610-3和5610-5可以分别执行指令5628-3和5628-5,从而以分布式方式来执行接收到的模型切片5623。

[0348] 图57图示了示例技术5700,该技术可以在推理环境(诸如,环境5600)中实现,以基于分布式计算平台的安全性能力以分布式方式来执行推理模型。要注意的是,为了方便和清楚起见,参考图56的推理环境5600描述了技术5700。然而,这并不意图是限制性的。

[0349] 技术5700可以在圆圈57.1处开始。在圆圈57.1处,客户端设备5601可以向服务器5603和5605查询服务器5603和5605的安全性能力。例如,客户端设备5601可以经由接口5630-1向边缘设备5303发送接收服务器5603和5605的服务器能力5629的请求。

[0350] 在圆圈57.2和57.3处,服务器5603和5605可以分别将其相应安全性能力的指示发送到客户端设备5601。例如,响应于在圆圈57.1处接收到请求,服务器5603可以经由接口5630-3将与服务器5603相对应的服务器能力5629发送到客户端设备5601。同样地,响应于在圆圈57.1处接收到请求,服务器5605可以经由接口5630-5将与服务器5605相对应的服务器能力5629发送到客户端设备5601。

[0351] 在圆圈57.4处,客户端设备5601可以部分地基于模型安全性要求5627和接收到的服务器能力5629将推理模型5622拆分成模型切片5623。而且,在圆圈57.4处,客户端设备5601可以基于服务器能力5629和模型安全性要求5627将模型切片5623指派给服务器5603或5605之一以用于以分布式方式来执行。

[0352] 在圆圈57.5和57.6处,客户端设备5601可以分别向服务器5603和5605发送被指派给相应服务器的模型切片5623的指示、以及以分布式方式来执行接收到的模型切片5623的指示。例如,在圆圈57.5处,客户端设备5601可以经由接口5630-1将模型切片5623-1发送到服务器5603。同样地,在圆圈57.6处,客户端设备5601可以经由接口5630-1将模型切片5623-2发送到服务器5605。在圆圈57.7和57.8处,服务器5603和5605可以以分布式方式来执行接收到的模型切片5623。

[0353] *小节VI-一般的*

*用以补偿输入质量的迭代推理*

当前推理系统的一个问题是输入数据的质量和/或数量可能是不合期望的。例如,数据集中的样本的过多数量可能被评估为不可接受的。在这种实例中,推理模型可能无法在没有置信度的阈值水平的情况下对不可接受的样本进行分类。此外,可以过滤掉这些不可接受的样本,从而减少输入数据的量(例如,出于训练或测试目的)。因此,本公开提供了一种推理系统,该推理系统能够增强输入数据中不可接受的样本以努力将它们转换成可接受的样本。

[0354] 图58图示了可以根据本公开来实现的示例强化的系统5801。强化的系统5801包括

电路5810、存储器5820、接口5830和输入设备5852。存储器5820存储一组一个或多个推理模型5822-1、5822-2、5822-N(或推理模型5822)、输入数据5824、增强的输入数据5825、输出数据5826、分类数据5828、增强的分类数据5829和指令5828。在许多示例中,输入数据5824可以包括一个或多个样本的数据集,以用于由一个或多个推理模型5822进行分类。在操作期间,电路5810可以执行指令5828,以响应于由一个或多个推理模型5822基于输入数据5824生成的分类数据5828被评估为不可接受而生成一组一个或多个增强的输入数据5825-1、5825-2、5825-N(或一组增强的输入数据5825)。

[0355] 此外,电路5810可以执行指令5828,以利用一个或多个推理模型5822对增强的输入数据5825进行分类,以产生增强的分类数据5829。在若干个示例中,该循环可以重复直到增强的分类数据5829被评估为可接受为止。有时,电路5810可以执行指令5828以执行一个或多个迭代推理,诸如以补偿输入数据5824的质量和/或数量。在许多示例中,将分类数据评估为可接受或不可接受可以基于分类数据5828中包括的置信度水平。例如,输入数据中的伪像(例如,视频/图像伪像)可能导致低置信度水平。一旦分类数据被评估为可接受的,则可以提供输入数据5824、分类数据5828、增强的输入数据5825和增强的分类数据5829或其指示中的一个或多个作为输出数据5826。

[0356] 在一些示例中,投票5832可以在该组推理模型5822当中发生以确定分类或推理。在各种示例中,流动编排5830可以映射出数据如何流动经过该组推理模型5822。在一个或多个示例中,可以诸如基于一个或多个推理模型5822进行的分类来生成警报5834。在一些示例中,例如,电路5810可以执行指令5828,以在由推理模型522对输入数据5824进行的分类中的置信度水平超过阈值(例如75%)时将分类数据(例如,分类数据5828和/或增强的分类数据5829)评估为可接受的。在各种示例中,可以通过增强不可接受的样本的一个或多个方面以使得它们在增强之后被评估为可接受的,从而增加数据集中的可用样本(例如,被评估为可接受的样本)的数量。在若干个示例中,对不可接受的样本的增强可以包括伽马的改变、像素移位、去噪、错误校正等中的一个或多个。有时,可以并行地对一个或多个增强进行执行和/或分类。此外,可以对该组一个或多个增强的版本进行比较和评估,诸如以标识操作异常。通常,增加可用样本的数量可以改进强化的系统的准确性。例如,可以在较大的样本集上重新训练推理模型5822,这通常导致改进的推理模型。

[0357] 有时,电路5810可以执行指令5828,以基于存储器5820利用其耦合到输入设备5852的接口5830、或基于接口5830通过其与输入设备5852进行通信的通信方案来标识和/或接收输入数据5824。在许多示例中,输入设备5852可以包括对环境和/或情形的高级和/或挑战性访问。例如,输入设备5852可以包括卫星、间谍相机、物联网(IoT)设备、或捕获罕见或有限的事件(例如,海啸、地震、冲突、市场异常、有限的历史记录等)的其他换能器。要注意的是,描绘了单个接口5830。然而,在一些示例中,可以为不同的输入设备152或具有不同特性(例如,不同的安全性要求等)的输入设备152提供不同的接口5830。

[0358] 电路5810可以执行指令5828以利用推理模型5822来基于输入数据5824生成分类数据5828。然后,可以评估分类数据5828以(例如,通过超过置信度阈值)确定输入数据5824是否是可接受的。当诸如基于分类数据5828而认为输入数据5824是不可接受的时候,电路5810可以执行指令5828从而以一种或多种方式来增强输入数据5824以生成增强的输入数据5825。在一些示例中,该增强可以包括保留局部属性的变换。可以由推理模型5822对增强

的输入数据5825进行分类以生成增强的分类数据5829。可以评估增强的分类5829，以确定增强的分类数据5829是否是可接受的以作为输出数据来提供。

[0359] 在各种示例中，可以比较分类数据5828以及增强的分类数据5829的一个或多个版本，诸如以检测操作异常。例如，推理模型5822针对其生成了不同分类和/或置信度水平的分类数据5828以及增强的分类数据5829的一个或多个版本可能指示对于欺骗强化的系统5801的尝试。例如，如果增强之前和之后的相同输入的两个输出导致了相互矛盾的预测(例如，<<Stop>>与<<Limit 90>>，则这可能指示对于欺骗该系统的尝试。有时，可以基于对输入数据5824的评估来生成通知(诸如，警报或日志)。附加地或可替代地，可以基于对输入数据5824的评估来实现一个或多个校正动作。

[0360] 图59图示了逻辑流程5900。逻辑流程5900可以表示由本文中描述的一个或多个设备执行的一些或全部操作。例如，强化的系统5801(或其组件)可以执行逻辑流程5900中的操作来增强输入数据以改进输入数据质量(例如，推理模型对输入数据进行分类的准确性或置信度)。

[0361] 逻辑流程5900可以在框5910“接收输入数据”处开始，其中强化的系统5801可以接收输入数据5824。例如，可以经由接口5830从输入设备5852接收输入数据5824。在一些示例中，可以在框5912之前对输入数据5824进行预处理。进行到框5912“基于推理模型对输入数据进行分类以产生分类数据”，可以基于推理模型对输入数据进行分类以产生分类数据。例如，推理模型5822可以基于输入数据5824来生成分类数据5828。在判定框5914“输入数据质量是否可接受”处，可以基于分类数据将输入数据评估为可接受的或不可接受的。例如，可以基于分类数据5828将输入数据5824评估为可接受的或不可接受的。如果输入数据被评估为不可接受的，则逻辑流程5900可以进行到框5916，并且如果输入数据被评估为可接受的，则逻辑流程5900可以进行到框5918。

[0362] 在框5916“增强输入数据”处，可以增强输入数据。例如，可以修改输入数据5824以产生增强的输入数据5825。在许多示例中，每当逻辑流程5900通过框5916时，可以将不同的增强应用于输入数据。逻辑流程5900然后可以返回到框5912并且如上所描述的那样重复。然而，如果输入数据被评估为可接受的并且逻辑流程5900进行到框5918，则可以将分类数据提供为输出数据。例如，可以将增强的分类数据5829提供为输出数据5826。在一些示例中，除了增强的分类数据5829之外或者替代于增强的分类数据5829，可以提供输入数据5824、分类数据5828和增强的输入数据5825中的一个或多个作为输出数据5826。从框5918，逻辑流程5900可以结束，或者可以基于接收到新的输入而重复。

[0363] 图60图示了逻辑流程6000。逻辑流程6000可以表示由本文中描述的一个或多个设备执行的一些或全部操作。例如，强化的系统5801(或其组件)可以执行逻辑流程6000中的操作，以拆分和/或增强输入数据，并且使用投票来改进输入数据质量(例如，推理模型对输入数据进行分类的准确性或置信度)。

[0364] 逻辑流程6000可以在框6010“接收输入数据”处开始，其中强化的系统5801可以接收输入数据5824。例如，可以经由接口5830从输入设备5852接收输入数据5824。在一些示例中，可以在框6012之前对输入数据5824进行预处理。进行到框6012“基于输入数据来生成一组增强的输入数据，其中该组中的每个增强的输入数据被不同地增强”，生成一组增强的输入数据。在一些示例中，该组增强的输入数据可以包括增强的数据5825-1、5825-2、5825-N。

例如，该组增强的输入数据可以包括各种增强的结果(例如，伽马的改变、像素移位、去噪、错误校正等)。

[0365] 进行到框6014“利用一个或多个推理模型对该组中的每个增强的输入数据进行分类”，可以利用一个或多个推理模型对各种增强中的每一个进行分类。例如，可以由推理模型5822-1、5822-2、5822-n中的一个或多个对输入数据的每个增强的版本进行分类。继续到判定框6016“是否经由基于不同增强输入数据的分类的投票而标识出可接受的输入数据”，可以经由基于不同增强输入数据的分类的投票来确定输入数据(或其一个或多个增强)是否是可接受的。如果判定框6016评估为“是”，则逻辑流程6000可以进行到框6020，并且如果判定框6016评估为“否”，则逻辑流程6000可以进行到框6018。

[0366] 在框6018“标识输入数据的一组新的增强”，可以标识输入数据的一组新的增强。例如，可以利用输入数据5824的一组新的增强来更新增强的输入数据5825。在框6018之后，逻辑流程6000可以返回到框6012，并且如上所描述的那样以该组新的增强而重复。参考回到当框6016评估为“是”时，逻辑流程6000可以继续到框6020。在框6020“提供具有最高置信度得分的分类数据作为输出数据”，可以提供具有最高置信度得分的分类作为输出数据。在一些示例中，附加地或替代地，这可以基于投票来确定。从框6020，逻辑流程6000可以结束，或者可以基于接收到新的输入而重复。

[0367] 图61图示了逻辑流程6100。逻辑流程6100可以表示由本文中描述的一个或多个设备执行的一些或全部操作。例如，强化的系统5801(或其组件)可以执行逻辑流程6100中的操作以拆分和/或增强输入数据，并且使用投票来检测异常并且生成警报和/或过滤掉对抗方。

[0368] 逻辑流程6100可以在框6110“接收输入数据”处开始，其中强化的系统5801可以接收输入数据5824。例如，可以经由接口5830从输入设备5852接收输入数据5824。在一些示例中，可以在框512之前对输入数据5824进行预处理。进行到框6112“基于输入数据来生成一组增强的输入数据，其中该组中的每个增强的输入数据被不同地增强”，生成一组增强的输入数据。在一些示例中，该组增强的输入数据可以包括增强的数据5825-1、5825-2、5825-N。例如，该组增强的输入数据可以包括各种增强的结果(例如，伽马的改变、像素移位、去噪、错误校正等)。

[0369] 进行到框6114“利用一个或多个推理模型对该组中的每个增强的输入数据进行分类”，可以利用一个或多个推理模型对各种增强中的每一个进行分类。例如，可以由推理模型5822-1、5822-2、5822-n中的一个或多个对输入数据的每个增强的版本进行分类。继续到判定框6116“是否经由基于输入数据的不同增强的分类的投票而检测到异常”，基于不同增强的输入数据的分类的投票可以用于检测异常。如果判定框6116评估为“是”，则逻辑流程6100可以进行到框6120，并且如果判定框6116评估为“否”，则逻辑流程6100可以进行到框6118。

[0370] 在框6118“提供具有最高置信度得分的分类数据作为输出数据”，可以提供具有最高置信度得分的分类作为输出数据。在一些示例中，附加地或替代地，这可以基于投票来确定。参考回到当框6116评估为“是”时，逻辑流程6100可以继续到框6120。在框6120“过滤异常和/或生成警报”处，可以响应于对异常的检测来实现警报和/或过滤。例如，警报可以包括例外状况的生成。从框6120，逻辑流程6100可以结束，或者可以基于接收到新的输入而重

复。

[0371] *用于加密模型的激活函数*

当前推理系统的一个问题是AI网络中的节点的激活可能是低效的。因此，本公开提供了一种推理系统，该推理系统使得能够实现高效的激活而不需要将在推理时间期间在输入数据上完成的计算，从而产生数据不可知系统。数据不可知系统可以与同态加密结合地使用。在各种示例中，同态加密可以包括允许在密文上进行计算的加密形式，从而生成加密结果，该加密结果在被解密时与操作的结果相匹配，就好像它们已经在明文上被执行一样。通常，同态加密支持多项式计算。因此，不可能按原样地使用常见的激活函数(例如ReLU、TanH、S形、Softmax)。因此，通常在同态加密网络中使用近似(例如，泰勒级数、切比雪夫)，但是它们不太准确并且运行时昂贵。

[0372] 在许多示例中，本文中所描述的推理系统(例如，强化的系统6201)针对模型存储可能需要忽略不计的开销，同时如利用一般的激活函数那样保留了准确性(例如，在+/-1％的得分改变之内)。在给定输入或一组输入的情况下，激活函数可以定义节点或神经元的输出。通常，该输出然后用作下一个节点的输入，并且依此类推，直到找到解决方案为止。在若干个示例中，强化的系统6201可以比基于近似的激活快几个数量级地运行。在各种示例中，强化的系统6201可以将该方法事后应用于已经训练的模型。此外，该方法可能不需要精细调整。有时，可能包括专用硬件以改进效率。

[0373] 图62图示了可以根据本公开来实现的示例强化的系统6201。强化的系统6201包括电路6210、存储器6220、接口6230和输入设备6252。存储器6220存储输入数据6224、推理模型6222、输出特征图6225、参数6227、激活统计信息6223、输出数据6226、激活矩阵6229和指令6228。在许多示例中，输入数据6224可以包括训练数据集。在操作期间，电路6210可以执行指令6228，以依赖于新颖的数据不可知激活来实现用于加密模型的专用激活函数。

[0374] 如将在下面更详细地描述的那样，电路6210可以执行指令6228以执行以下各项中的一个或多个：收集激活统计信息6228、分析激活统计信息6228、以及在推理时进行处理。例如，首先，电路6210可以执行指令6228以采用推理模型6222(例如，经训练的深度神经网络(DNN)模型)和输入数据6224(例如，具有样本的训练数据集)，并且初始化空的激活矩阵6229，该激活矩阵6229对应于所有层的每个输出特征图6225。通常，输入数据6224中的每个训练样本可以正向传播，同时所有层的所有激活被累积在对应的激活矩阵6229中以产生激活图6231。在各种示例中，激活统计信息6223可以包括完成的激活矩阵6229和/或激活图。

[0375] 接下来，电路6210可以执行指令6228以分析激活统计信息6223。在一些示例中，可以基于一个或多个参数6227来分析该统计信息。有时，一个或多个参数6227可以由用户提供。通常，一个或多个参数6227可以包括第一、第二和第三参数。第一参数可以包括所激发的激活中将被固定的那些激活的百分比。第二参数可以包括应当被激发的神经元的总百分比。第三参数可以包括要从其中取得该固定的第一参数索引的总顶部(total top)激活的百分位数。例如，用户可能选择第一参数为50、第二参数为80、以及第三参数为70。在这种示例中，这意味着总体上将激发80％的神经元；它们中的50％每次运行被随机化；并且剩余的50％被预先固定，并且是从顶部70％的激活中随机选择的。电路6210可以执行指令6228以扫描每个激活图6231，并且从顶部激活神经元(例如，最激活的特征图的值)中选择的随机第一参数百分比的索引被保存。

[0376] 为了在运行时进行推理，电路6210可以执行指令6228以分析激活统计信息6223。推理时的实际激活操作可以包括传递处于经处理的层的对应保存的位置图(例如，激活图6231)中的所有神经元。可以通过随机地激发所需数量的神经元来实现剩余(第二参数——第一参数)激活百分比。该方法可以实现高效的激活，但是在推理时间期间对数据本身不进行任何计算(因此数据不可知)。此外，针对模型存储可能需要忽略不计的开销，同时保留了准确性，并且其可以比基于近似的激活快几个数量级地运行。而且，该方法可以被事后应用于已经训练的模型，并且不需要精细调整。

[0377] 图63图示了逻辑流程6300。逻辑流程6300可以表示由本文中描述的一个或多个设备执行的一些或全部操作。例如，强化的系统6201(或其组件)可以执行逻辑流程6300中的操作以用于高效的推理系统。

[0378] 逻辑流程6300可以在框6302处开始。在框6302“初始化与推理模型中的每一层的输出特征图相对应的一个或多个矩阵”，可以初始化与推理模型中的每一层的输出特征图相对应的一个或多个矩阵。例如，可以初始化与一个或多个输出特征图6225相对应的一个或多个激活矩阵6229。

[0379] 继续到框6304“将数据集中的样本的激活累积到一个或多个矩阵中以产生激活统计信息”，可以将数据集中的样本的激活累积到对应的矩阵中以产生激活统计信息。例如，每个样本的每次激活可以被累积到激活矩阵6229中。在一些示例中，每个样本可以正向传播，同时所有层的所有激活图被保存。

[0380] 在框6306“确定第一、第二和第三参数，其中第一参数包括被固定的激活的百分比，第二参数包括推理模型中应当被激发的节点的总百分比，并且第三参数包括要从其中取得固定激活的总顶部激活的百分位数”处，可以确定第一、第二和第三参数。例如，参数6227可以包括第一、第二和第三参数。在各种示例中，可以基于用户输入来确定参数6227。

[0381] 继续到框6308“基于激活统计信息、第一参数、第二参数和第三参数来实现专用激活函数”，可以基于激活统计信息、第一参数、第二参数和第三参数来实现专用激活函数。例如，可以基于参数6227和激活统计信息6223来实现专用激活函数。在这种实例中，激活统计信息6223可以包括激活矩阵6229和/或激活图6231。

[0382] 图64图示了逻辑流程6400。逻辑流程6400可以表示由本文中描述的一个或多个设备执行的一些或全部操作。例如，强化的系统401(或其组件)可以执行逻辑流程6400中的操作以用于高效的推理系统，。

[0383] 逻辑流程6400可以在框6402处开始。在框6402“加载保存的固定激活”处，可以加载被保存的固定激活。例如，可以加载激活矩阵6229和激活图6231中的一个或多个中的固定激活。进行到框6404“配置:要激发的输出神经元的固定位置和百分比”，可以确定、实现和/或配置一个或多个配置参数，该配置参数包括要激发的输出神经元的固定位置和百分比。例如，指令6228可以基于输入数据6224、推理模型6222、激活统计信息6223、输出特征图6225、参数6227、输出数据6226和激活图6231中的一个或多个来确定/配置输出神经元的固定位置和百分比。

[0384] 参考回到框6402，一旦加载了保存的固定激活，逻辑流程6400就可以进行到框6406。在框6406“激发固定位置处的神经元加上随机百分比的神经元”处，固定位置处的神经元和随机百分比的神经元、或随机神经元的百分比。例如，可以基于参数6227中的一个或

多个来确定该随机百分比的神经元。在一些示例中，本文中描述的一个或多个激活函数可以用于利用多方计算(例如，MPC方案和/或同态加密等)加密的神经网络。

[0385] *用于模型保护的多方执行*

如本文所提供的，推理模型可以在分布式环境中执行，诸如例如通过云计算资源、通过边缘设备等来执行。在这种环境中执行推理系统的部分可能会引入对数据泄漏的风险。作为简单的示例，推理系统可以被布置成利用私有组份(例如，健康记录、财务信息等)来推理信息。给定来自在分布式环境中的平台上执行的推理模型的一个或多个“切片”的输入和/或输出，对抗方可能能够推理出该私有信息。此外，当在云和/或边缘设备上执行时，该模型的部分经常处于被提取的风险之下。

[0386] 本公开提供了在训练期间使用附加的推理模型(被称为“对抗性模型”或“对抗方”)，其中主推理模型是基于来自对抗性模型的反馈来训练的。例如，可以生成对抗性模型来表示每个分布式计算平台(例如，每个模型切片等)。对抗方有权访问来自其他对抗方的输入和输出。对抗性模型与推理模型的切片耦合以用于训练。在训练期间，通过应用联合损失(例如，一般任务特定的损失，诸如分类损失等)和对抗性损失来训练推理模型，以保证该模型的每个组份都是关键的。也就是说，训练基于所有模型的联合损失将整个推理模型(或所有切片的组合)引导到局部最小值。然而，每个切片被训练成使对抗性损失最小化。由此，由于对抗性损失训练，这些切片在被单独地采用时被引导到不太准确的局部最小值。

[0387] 因此，一旦被训练，模型切片就可以由不同的计算平台以分布式方式来执行，同时降低了数据泄漏或模型提取的风险。也就是说，作为多方对抗性训练的结果，模型切片中的每一个可以展现出隐私性和/或安全性特征，同时维持推理准确性。即使执行模型切片的计算平台具有对彼此的I/O访问权也是如此。换句话说，对抗性训练可以用于引导模型切片作为秘密共享者进行协作，其中该秘密是模型函数(例如，推理模型输出到输入的映射)。也就是说，该秘密是$f(x)\rightarrow y$ ，其中$x$是输入样本，并且$y$是结果(例如，分类、预测、推理、所生成的输出等)。因此，每个模型切片对于确定该秘密(例如$f(x)\rightarrow y$)是必需的。然而，每个模型切片在被单独使用时不足以确定该秘密($f(x)\rightarrow y$)。

[0388] 图65图示了包括客户端设备6501以及服务器6503、6505和6507的示例推理环境6500。通常，服务器6503、6505和6507被布置成以分布式方式来执行推理模型的一部分。服务器6503、6505和6507可以是基于云的(或可访问的)计算资源、边缘设备、专门布置的推理模型执行加速器等。此外，尽管在该图中描绘了三个服务器6503、6505和6507，但是应当领会的是，可以在推理环境6500中提供任何数量(1个、2个、3个、4个或更多)的服务器。如上所指出，推理环境6500的服务器6503、6505和6507可以是边缘设备。在一些示例的情况下，推理环境6500可以是多边缘环境，其中客户端设备6501利用“边缘”(例如，边缘设备6503、6505等)中的空闲计算功率来执行模型切片。

[0389] 客户端设备6501以及服务器6503、6505和6507可以经由网络(例如，其可以包括互联网等)而耦合。更具体地，接口6530-1、6530-3、6530-5和6530-7可以耦合到网络，以便于设备之间的数据通信或交换。

[0390] 客户端设备6501包括电路6510-1、存储器6520-1和接口6530-1。存储器6520-1存储推理模型6522、输入数据6524、输出数据6526。推理模型6522包括模型切片6523。通常，客户端设备6501进行操作以协调并且提供推理模型6522的分布式执行。

[0391] 服务器6503包括电路6510-3、存储器6520-3和接口6530-3。存储器6520-3存储模型切片6522-1和指令6528-3。同样地，服务器6505包括电路6510-5、存储器6520-5和接口6530-5。存储器6520-5存储模型切片6522-2和指令6528-5。附加地，服务器6507包括电路6510-7、存储器6520-7和接口6530-7。存储器6520-7存储模型切片6522-3和指令6528-7。通常，服务器6503、6505和6507针对推理模型6522的部分(例如，模型切片6523-1、6523-2和6523-3等)提供执行环境。

[0392] 通常，服务器6503、6505和6507均可以以分布式方式来执行模型切片6523之一。例如，电路6510-3可以执行指令6528-3以执行模型切片6523-1，电路6510-5可以执行指令6528-5以执行模型切片6523-2，并且电路6510-7可以执行指令6528-7以执行模型切片6523-3。

[0393] 如上所指出，本公开提供了：可以利用对抗性模型来训练模型切片，该对抗性模型代替了分布式系统中的每一方。

[0394] 图66图示了系统6601，该系统可以被实现为训练分区推理模型以用于在如上所描述的分布式环境中的安全执行。系统6601包括电路6610、存储器6620和接口6630。存储器6620存储被拆分成多个部分或模型切片6623的推理模型6622。例如，描绘了包括模型切片6623-1、6623-2和6623-3的推理模型6622。通常，推理模型可以被拆分成与分布式计算平台一样多的部分(或切片)。例如，图65的推理环境6500描绘了可以执行模型部分的三(3)个计算平台(或参与方)，具体地是服务器6503、6505和6507。存储器6620还包括用于每个模型切片6623的对抗性模型6625。例如，该图描绘了对抗性模型6625-1、6625-2和6625-3。

[0395] 电路6610可以执行指令6628以使用对抗性训练技术、使用对抗性模型和替代性子模型(见图67和68)来训练模型切片6623。

[0396] 图67图示了示例推理模型6722，其被拆分成模型切片6723-1、6723-2和6723-3，这些模型切片耦合到为了对抗性训练而布置的对抗性模型6725和替代性子模型6726。一般而言，该图描绘了其中对抗性模型与模型切片之间存在1:1对应关系的实现方式。如前所陈述，训练进行操作以引导模型切片6723在利用对应的替代性子模型6726执行时单独地达到不准确或不太准确的局部最小值。在训练期间，对抗性模型6725具有对不与对抗性模型6725相对应的其他模型切片6723的读取访问权、以及对与对抗性模型6725相对应的替代性子模型6726的读取和写入访问权。例如，对抗性模型6725-1具有对模型切片6723-2和6723-3的读取访问权、以及对替代性子模型6725-1的读取和写入访问权。

[0397] 图68A、68B和68C图示了示例推理模型6822，其被拆分成模型切片6823-1、6823-2和6823-3，这些模型切片耦合到为了对抗性训练而布置的对抗性模型6825和替代性子模型6826。一般而言，该图描绘了其中对抗性模型与模型切片之间存在N:1对应关系的实现方式。换句话说，该图描绘了多个对抗性模型，这些对抗性模型耦合到单个模型切片以用于对抗性训练。

[0398] 更特别地转向图68A，描绘了对抗性模型6825-11、6825-12和6825-13、以及对应的替代性子模型6826-11、6826-12和6826-13。对抗性模型6825-11、6825-12和6825-13、以及替代性子模型6826-11、6826-12和6826-13与对模型切片6823-2和6823-3的读取访问权连接，以提供对模型切片6823-1的对抗性训练。同样地，图68B描绘了对抗性模型6825-21、6825-22和6825-23、以及对应的替代性子模型6826-21、6826-22和6826-23。对抗性模型

6825-21、6825-22和6825-23、以及替代性子模型6826-21、6826-22和6826-23与对模型切片6823-1和6823-3的读取访问权连接，以提供对模型切片6823-2的对抗性训练。最后，图68C描绘了对抗性模型6825-31、6825-32和6825-33、以及对应的替代性子模型6836-31、6836-32和6836-33。对抗性模型6825-31、6825-32和6825-33、以及替代性子模型6826-31、6826-32和6826-33与对模型切片6823-1和6823-2的读取访问权连接，以提供对模型切片6823-3的对抗性训练。

[0399] *示例物理实施例*

一些示例可以包括制品或至少一个计算机可读介质。计算机可读介质可以包括存储逻辑的非暂时性存储介质。在一些示例中，非暂时性存储介质可以包括能够存储电子数据的一个或多个类型的计算机可读存储介质，包括易失性存储器或非易失性存储器、可移除或不可移除存储器、可擦除或不可擦除存储器、可写入或可重新写入的存储器等。在一些示例中，逻辑可以包括各种软件元件，诸如软件组件、程序、应用、计算机程序、应用程序、系统程序、机器程序、操作系统软件、中间件、固件、软件模块、例程、子例程、函数、方法、进程、软件接口、API、指令集、计算代码、计算机代码、代码段、计算机代码段、字、值、符号或其任何组合。

[0400] 根据一些示例，计算机可读介质可以包括存储或维护指令的非暂时性存储介质，所述指令在由机器、计算设备或系统执行时使得机器、计算设备或系统执行根据所描述的示例的方法和/或操作。指令可以包括任何合适类型的代码，诸如源代码、经编译的代码、经解释的代码、可执行代码、静态代码、动态代码等。可以根据预定义的计算机语言、方式或语法来实现指令，以用于指令机器、计算设备或系统执行某个功能。可以使用任何合适的高级、低级、面向对象的、视觉、经编译的和/或经解释的编程语言来实现指令。

[0401] 图69图示了存储介质6900。存储介质6900可以包括任何非暂时性计算机可读存储介质或机器可读存储介质，诸如光学、磁性或半导体存储介质。在各种实施例中，存储介质6900可以包括制品。在一些实施例中，存储介质6900可以存储计算机可执行指令，诸如用以实现本文中所描述的技术、逻辑流程或操作中的一个或多个的计算机可执行指令，诸如关于200、300、500、700、1200、1300、1600、1700、2100、2500、2800、3000、3200、3300、3500、3700、3900、4100、4300、4600、4900、5100、5400、5700、5900、6000、6100、6300或6400的计算机可执行指令。附加地，存储介质6900可以存储表示输入数据、输出数据、推理模型、测试/训练数据集等的数据结构。计算机可读存储介质或机器可读存储介质的示例可以包括能够存储电子数据的任何有形介质，包括易失性存储器或非易失性存储器、可移除或不可移除存储器、可擦除或不可擦除存储器、可写入或可重新写入的存储器等等。计算机可执行指令的示例可以包括任何合适类型的代码，诸如源代码、经编译的代码、经解释的代码、可执行代码、静态代码、动态代码、面向对象的代码、视觉代码等。实施例在该情境中不受限制。

[0402] 适合于存储和/或执行程序代码的强化的系统将包括通过系统总线直接或间接耦合到存储器元件的至少一个处理器。存储器元件可以包括在程序代码的实际执行期间采用的本地存储器、大容量存储装置、以及提供对至少一些程序代码的临时存储以减少在执行期间必须从大容量存储装置中检索代码的次数的高速缓冲存储器。术语“代码”涵盖了宽范围的软件组件和构造，包括应用、驱动器、过程、例程、方法、模块、固件、微代码和子程序。因此，术语“代码”可以用于指代指令的任何集合，这些指令在由处理系统执行时执行期望的

一个或多个操作。

[0403] 本文中描述的逻辑电路、设备和接口可以执行以硬件实现的功能、以及利用在一个或多个处理器上执行的代码实现的功能。逻辑电路指代实现一个或多个逻辑功能的硬件或硬件和代码。电路是硬件，并且可以指代一个或多个电路。每个电路可以执行特定功能。电路中的电路可以包括与一个或多个导体互连的分立电子组件、集成电路、芯片封装、芯片组、存储器等。集成电路包括在衬底(诸如，硅晶圆)上创建的电路，并且可以包括组件。并且集成电路、处理器封装、芯片封装和芯片组可以包括一个或多个处理器。

[0404] 处理器可以在(一个或多个)输入处接收信号(诸如，指令和/或数据)，并且处理该信号以生成至少一个输出。在执行代码时，代码会改变构成处理器流水线的晶体管的物理状态和特性。晶体管的物理状态转换成存储在处理器内的寄存器中的1和0的逻辑位。处理器可以将晶体管的物理状态传递到寄存器中，并且将晶体管的物理状态传递到另一个存储介质。

[0405] 处理器可以包括用以执行一个或多个子功能的电路，该子功能被实现为执行处理器的总体功能。处理器的一个示例是包括至少一个输入和至少一个输出的状态机或专用集成电路(ASIC)。状态机可以通过在至少一个输入上执行预定系列的串行和/或并行操纵或变换，来操纵至少一个输入以生成至少一个输出。

[0406] 如上所描述的逻辑可以是用于集成电路芯片的设计的一部分。芯片设计以图形计算机编程语言而创建，并且存储在计算机存储介质或数据存储介质(诸如，磁盘、磁带、物理硬盘驱动器、或诸如存储访问网络中的虚拟硬盘驱动器)中。如果设计者不制造芯片或用于制造芯片的光刻掩模，则设计者将通过物理手段(例如，通过提供存储该设计的存储介质的副本)或以电子方式(例如，通过互联网)将所得的设计直接或间接地传输到这种实体。然后，将所存储的设计转换成适当的格式(例如GDSII)来进行制造。

[0407] 制造商可以以原始晶圆形式(也就是说，具有多个未封装芯片的单个晶圆)作为裸管芯、或以封装的形式来分布所得的集成电路芯片。在后一种情况下，芯片被安装在单芯片封装(诸如，塑料载体，其中引线被固定到母板或其他更高级别的载体)中，或者被安装在多芯片封装(诸如，具有表面互连或埋入互连中的任一个或两者的陶瓷载体)中。在任何情况下，然后将该芯片与其他芯片、分立电路元件和/或其他信号处理设备集成在一起，作为(a)中间产品(诸如，处理器板、服务器平台或母板)、或(b)最终产品的一部分。

[0408] 图70图示了系统7000。系统7000是具有多个处理器核的计算机系统，诸如分布式计算系统、超级计算机、高性能计算系统、计算集群、大型计算机、迷你计算机、客户端-服务器系统、个人计算机(PC)、工作站、服务器、便携式计算机、膝上型计算机、平板计算机、手持设备(诸如，个人数字助理(PDA))、或用于处理、显示或传输信息的其他设备。类似的实施例可以包括例如娱乐设备，诸如便携式音乐播放器或便携式视频播放器、智能电话或其他蜂窝电话、电话机、数字视频相机、数字静物相机、外部存储设备等。进一步的实施例实现了更大规模的服务器配置。在其他实施例中，系统7000可以具有带有一个核的单个处理器或多于一个处理器。要注意的是，术语“处理器”指代具有单个核的处理器、或具有多个处理器核的处理器封装。在至少一个实施例中，计算系统7000表示本文中所描述的强化的系统、客户端设备、服务器、边缘设备或其他计算设备。更一般地，计算系统7000被配置成实现本文中参考图1至图69而描述的所有逻辑、系统、逻辑流程、方法、装置和功能。

[0409] 如本申请中所使用的,术语“系统”和“组件”以及“模块”意图指代与计算机有关的实体,该实体是硬件、硬件和软件的组合、软件、或执行中的软件,这些的示例由示例性系统7000来提供。例如,组件可以是但不限于在处理器上运行的过程、处理器、硬盘驱动器、(光学和/或磁性存储介质的)多个存储驱动器、对象、可执行文件、执行的线程、程序和/或计算机。通过说明,在服务器上运行的应用和服务器两者都可以是组件。一个或多个组件可以驻留在执行的线程和/或过程内,并且组件可以位于一个计算机上和/或可以被分布在两个或更多个计算机之间。此外,组件可以通过各种类型的通信介质彼此可通信地耦合以协调操作。协调可以涉及信息的单向或双向交换。例如,组件可以以通过通信介质传送的信号的形式来传送信息。该信息可以实现为被分配给各种信号线路的信号。在这种分配中,每个消息是一信号。然而,进一步的实施例可以替代地采用数据消息。这种数据消息可以跨各种连接而被发送。示例性连接包括并行接口、串行接口和总线接口。

[0410] 如该图中所示,系统7000包括用于安装平台组件的母板7005。母板7005是点对点互连平台,该平台包括经由点对点互连7056(诸如,超路径互连(UPI)、无穷结构(Infinity Fabric)(IF)等)耦合的第一处理器7010和第二处理器7030。在其他实施例中,系统7000可以属于另一个总线架构,诸如多分支总线。此外,处理器7010和7030中的每一个可以是具有多个处理器核的处理器封装,分别包括(一个或多个)处理器核7020和7040。尽管系统7000是两插槽(2S)平台的示例,但是其他实施例可以包括多于两个的插槽或一个插槽。例如,一些实施例可以包括四插槽(4S)平台或八插槽(8S)平台。每个插槽是用于处理器的安装座,并且可以具有插槽标识符。要注意的是,术语“平台”指代安装了某些组件(诸如,处理器7010和芯片组7060)的母板。一些平台可以包括附加组件,并且一些平台可以包括用于安装处理器和/或芯片组的插槽。

[0411] 处理器7010、7020可以是各种商业可得的处理器中的任一种,包括但不限于Intel®Celeron®、Core®、Core(2)Duo®、Itanium®,Pentium®、Xeon®和XScale®处理器;AMD® Athlon®、Duron®和Opteron®处理器;ARM®应用、嵌入式和安全处理器;IBM®和Motorola®DragonBall®和PowerPC®处理器;IBM和Sony®Cell处理器;和类似的处理器。双微处理器、多核处理器和其他多处理器架构也可以被采用作为处理器7010、7020。通常,处理器7010和/或7020可以对应于本文中讨论的任何电路,诸如例如关于图1至图68而描绘和描述的电路。

[0412] 第一处理器7010包括集成存储器控制器(IMC)7014以及点对点(P-P)接口7018和7052。类似地,第二处理器7030包括IMC 7034和P-P接口7038和7054。IMC 7014和7034分别将处理器7010和7030耦合到相应的存储器——存储器7012和存储器7032。通常,处理器7010和/或7020可以对应于本文中讨论的任何存储器,诸如例如关于图1至图68而描绘和描述的存储器。存储器7012和7032可以是用于平台的主存储器(例如,动态随机存取存储器(DRAM))的部分,诸如双倍数据速率类型3(DDR3)或类型4(DDR4)同步DRAM(SDRAM)。在本实施例中,存储器7012和7032本地附接到相应的处理器7010和7030。在其他实施例中,主存储器可以经由总线和共享存储器集线器与处理器耦合。

[0413] 处理器7010和7030分别包括与(一个或多个)处理器核7020和7040中的每一个耦合的高速缓存。在本实施例中,处理器7010的(一个或多个)处理器核7020和处理器7030的(一个或多个)处理器核7040包括上面描述的电路(例如,参考图1至图69而描述的电路

N10)。也就是说,功能也可以全部或部分地作为电路而驻留在处理器7010内,并且可以例如在处理器或7010、7030内的寄存器或缓冲器(诸如,寄存器7016)内执行操作,或者在处理器7010、7030的指令流水线内执行操作。

[0414] (一个或多个)处理器核7020和7040可以表示处理器内的电路与介质的组合,该介质用以将上面描述的电路、逻辑或技术的全部或部分功能存储在存储器(诸如,高速缓存、存储器7012、缓冲器、寄存器等)中。在若干个实施例中,功能全部或部分地作为代码而驻留在存储器(诸如,经由芯片组7060而附接到处理器7010、7030的存储介质700)中。附加地,功能还可以全部或部分地驻留在诸如存储器7012和/或7032之类的存储器(例如,参考图1至图69而描述的存储器N20)中。

[0415] 如所陈述的,处理器7010和7030中的多于一个可以包括经由接口(I/F)7066与芯片组7060耦合的硬件(例如,加速器等)中的所描述的功能。I/F 7066可以是例如增强的外围组件互连(PCI-e)。

[0416] 第一处理器7010经由P-P互连7052和7062耦合到芯片组7060,并且第二处理器7030经由P-P互连7054和7064耦合到芯片组7060。直接媒体接口(DMI)7057和7058可以分别耦合P-P互连7052和7062以及P-P互连7054和7064。DMI可以是便于例如每秒八次千兆传输(GT/s)的高速互连,诸如DMI 3.0。在其他实施例中,处理器7010和7030可以经由总线而互连。

[0417] 芯片组7060可以包括控制器集线器,诸如平台控制器集线器(PCH)。芯片组7060可以包括执行时钟功能的系统时钟,并且包括用于I/O总线的接口(诸如,通用串行总线(USB)、外围组件互连(PCI)、串行外围互连(SPI)、集成互连(I2C)等)以便于连接该平台上的外围设备。在其他实施例中,芯片组8060可以包括多于一个的控制器集线器,诸如具有存储器控制器集线器、图形控制器集线器和输入/输出(I/O)控制器集线器的芯片组。

[0418] 在本实施例中,芯片组7060经由接口(I/F)7070与可信平台模块(TPM)7072和UEFI、BIOS、闪存组件7074耦合。TPM 7072是专用微控制器,其被设计成通过将密码密钥集成到设备中来保护硬件。UEFI、BIOS、闪存组件7074可以提供预启动代码。

[0419] 此外,芯片组7060包括I/F 7066以将芯片组7060与高性能图形引擎——图形卡7065耦合。在其他实施例中,系统7000可以在处理器7010和7030与芯片组7060之间包括柔性显示接口(FDI)。FDI将处理器中的图形处理器核与芯片组7060互连。

[0420] 各种I/O设备7092耦合到总线7081、以及将总线7081耦合到第二总线7091的总线桥7080和将总线7081与芯片组7060连接的I/F 7068。在一个实施例中,第二总线7091可以是低引脚数(LPC)总线。各种设备可以耦合到第二总线7091,包括例如键盘7082、鼠标7084、通信设备7086、以及可以存储如本文中先前描述的计算机可执行代码的存储介质6900。此外,音频I/O 7090可以耦合到第二总线7091。I/O设备7092、通信设备7086和存储介质6900中的许多可以驻留在母板7005上,而键盘7082和鼠标7084可以是附加外围设备。在其他实施例中,I/O设备7092、通信设备7086和存储介质6900中的一些或全部是附加外围设备,并且不驻留在母板7005上。通常,音频设备7090和/或I/O设备7092可以对应于本文中讨论的任何输入和/或输出设备,诸如例如关于图1至68而描绘和描述的输入设备、输出设备(或类似物)。

[0421] 图71图示了系统7100。系统7100是计算机系统,其可以被实现为处理本文中所描

述的各种工作负载，诸如例如训练推理模型、从推理模型生成输出等。该图的处理资源可以由上面描述的强化的系统、推理环境等中的任一个来使用。系统7100包括边缘设备7110、数据中心或云服务7120、节点7130和网络7140。

[0422] 节点7130可以包括具有处理资源和/或输入资源的系统7100的任何组件或设备。这些组件可以包括类似于上面描述的那些特征的特征(例如，系统100、系统7000等)。例如，节点7130可以包括利用处理资源实现的数据中心和/或云服务7120，该处理资源被布置成处理推理模型工作负载(例如，数据中心中的机器学习加速器、利用处理资源实现的边缘设备7110(例如，最终用户设备7112、物联网(IoT)设备7114、网关7116)等等)。作为另一个示例，节点7130可以包括利用被布置成生成用于推理模型的输入的组件(例如，相机、传感器等)来实现的边缘设备7110。下面将进一步讨论系统7100的这些各种组件。

[0423] 边缘设备7110可以包括靠近通信系统(例如，系统7100、网络7140等)的“边缘”而部署或连接的任何装备和/或设备。边缘设备7110可以通过一个或多个网络和/或通信协议(诸如，网络7140)彼此通信和/或与其他远程网络和服务(例如，数据中心和/或云服务7120)进行通信。在一些实施例中，某些边缘设备7110可以包括遍及本公开而描述的处理资源，并且因此可以用作节点7130。在所图示的实施例中，边缘设备7110包括最终用户设备7112(例如，台式计算机、膝上型计算机、移动设备)、IoT设备7114、以及网关、路由器、交换机和/或集线器7116以及其他设备。

[0424] 最终用户设备7112可以包括实现或便于与计算系统7100进行用户交互的任何设备，包括例如台式计算机、膝上型计算机、平板设备、移动电话和其他移动设备、以及可穿戴设备(例如，智能手表、智能眼镜、头戴式耳机)以及其他示例。

[0425] IoT设备7114可以包括能够在IoT系统或网络中进行通信和/或参与在其中的任何设备。IoT系统可以指代由对于特定应用或用例相互操作的多个不同设备(例如，IoT设备7114)组成的新的或改进的自组织系统和网络。这种自组织系统随着越来越多的产品和装备演进到变得“智能”而出现，这意味着它们由计算机处理器控制或监测并且能够与其他设备进行通信。例如，IoT设备7114可以包括用以允许与系统7100的其他组件相互操作(诸如，与数据中心、和/或云服务7120、和/或其他边缘设备7110相互操作)的计算机处理器和/或通信接口。IoT设备7114可以是全新开发的具有IoT能力的“绿地”设备，或者可以是通过将IoT能力集成到最初开发的不具有IoT能力的现有传统设备中而产生的“棕地”设备。例如，在某些情况下，可以从集成在“物件(things)”中或者附着到“物件”的传感器和通信模块来建立IoT设备7114，该“物件”诸如是装备、玩具、工具、车辆、活物(例如，植物、动物、人类)等等。可替代地或附加地，某些IoT设备7114可以依赖于中间组件(诸如，边缘网关或路由器7116)来与系统7100的各种组件进行通信。

[0426] IoT设备7114可以包括用于监测、检测、测量和生成与其环境的特性相关联的传感器数据和信号的各种类型的传感器。例如，给定的传感器可以被配置成检测一个或多个相应的特性，诸如移动、重量、物理接触、温度、风、噪声、光、位置、湿度、辐射、液体、特定化学成分、电池寿命、无线信号、计算机通信、以及带宽和其他示例。传感器可以包括物理传感器(例如，物理监测组件)和虚拟传感器(例如，基于软件的监测组件)。IoT设备7114还可以包括用以在其相应的环境中执行各种动作的致动器。例如，致动器可以被用来选择性地激活某个功能，诸如切换安全性系统(例如，警报、相机、锁)或家用电器(例如，音频系统、照明、

HVAC电器、车库门)的功率或操作以及其他示例。

[0427] 实际上,本公开设想到使用IoT设备7114和相关联的传感器/致动器的潜在地无限的领域。IoT设备7114例如可以包括与任何类型的系统7100和/或行业相关联的任何类型的装备和/或设备,这些行业包括运输(例如汽车、航空公司)、工业制造、能源(例如,发电厂)、电信(例如,互联网、蜂窝和电视服务提供商)、医疗(例如,健康护理、制药)、食品加工和/或零售行业等等。在运输行业中,例如,IoT设备7114可以包括与飞机、汽车或船舶相关联的装备和设备,诸如导航系统、自主飞行或驾驶系统、交通传感器和控制器、和/或由传感器监测的任何内部机械或电气组件(例如,引擎)。IoT设备7114还可以包括与工业制造和生产、装运(例如,货物跟踪)、通信网络(例如,网关、路由器、服务器、蜂窝塔)、服务器机群、发电厂、风力发电场、油气管道、水处理和配送、废水收集和处理、以及天气监测(例如,温度、风和湿度传感器)以及其他示例相关联的装备、设备和/或基础设施。IoT设备7114例如还可以包括任何类型的“智能”设备或系统,诸如智能娱乐系统(例如,电视、音频系统、视频游戏系统)、智能家用或办公室电器(例如,供热通风和空调(HVAC)电器、冰箱、洗衣机和烘干机、咖啡机)、功率控制系统(例如,自动电力、光和HVAC控制)、安全性系统(例如,警报、锁、相机、运动检测器、指纹扫描器、面部识别系统)、以及其他家庭自动化系统和其他示例。IoT设备7114可以被静态地定位,诸如安装在建筑物、墙壁、地板、地面、路灯柱、指示牌、水塔、或任何其他固定或静态结构上。IoT设备7114也可以是移动的,诸如车辆或飞机、无人机、包裹(例如,用于跟踪货物)、移动设备、以及可穿戴设备和其他示例中的设备。此外,IoT设备7114还可以是任何类型的边缘设备7110,其包括最终用户设备7112以及边缘网关和路由器7116。

[0428] 边缘网关和/或路由器7116可以被提供为节点7130。附加地,边缘网关和/或路由器7116可以被用来便于去往和来自边缘设备7110的通信。例如,网关7116可以给最初开发的不具有任何通信能力的现有传统设备(例如,“棕地”IoT设备)提供这种通信能力。网关7116还可以被利用来扩展具有短距离、专有或其他方面受限的通信能力的边缘设备7110的地理范围,该边缘设备110诸如是具有蓝牙或ZigBee通信能力的IoT设备7114。例如,网关7116可以充当IoT设备7114与远程网络或服务之间的中介(通过使用IoT设备7114固有的通信能力(蓝牙、ZigBee)为IoT设备7114提供前传(front-haul),并且使用另一有线或无线通信介质(例如,以太网、Wi-Fi、蜂窝)为其他网络7140和/或数据中心和/或云服务7120提供回传)。在一些实施例中,网关7116可以通过专用的网关设备或者通过通用设备(诸如,另一个IoT设备7114、最终用户设备7112或者其他类型的边缘设备7110)来实现。

[0429] 在某些实例中,网关7116还可以单独地或者与其他组件(诸如,数据中心和/或云服务7120和/或其他边缘设备7110)结合地实现某个网络管理和/或应用功能(例如,用于IoT设备7114的IoT管理和/或IoT应用功能)。例如,在一些实施例中,可以向或者从网关设备7116推送或拉取配置参数和/或应用逻辑,从而允许针对特定的IoT应用或用例来配置处于网关7116的范围内或其附近的IoT设备7114(或其他边缘设备7110)。

[0430] 数据中心和/或云服务120可以包括通过网络7140或在“云”中远程地(例如,在数据中心、多个数据中心处等)托管的服务。在一些实施例中,例如,数据中心和/或云服务7120可以被远程地托管在数据中心中的服务器(例如,应用服务器、数据库服务器、具有加速器的服务器等)上。数据中心和/或云服务7120可以包括能够由边缘设备7110利用或者为

边缘设备7110利用的任何服务，包括但不限于数据存储、计算服务(例如，数据分析、搜索、诊断和故障管理)、安全性服务(例如，监控、警报、用户认证)、地图绘制和导航、地理位置服务、网络或基础设施管理、IoT应用和管理服务、支付处理、音频和视频流送、消息传送、社交网络、新闻和天气以及其他示例。在一些实施例中，某些数据中心和/或云服务7120可以包括遍及本公开所描述的矩阵处理功能，并且因此可以被用作节点7130。

[0431] 网络7140可以被用来便于计算系统7100的组件之间的通信。例如，边缘设备7110(诸如，最终用户设备7112和IoT设备7114)可以使用网络7140彼此通信和/或访问一个或多个远程服务(例如，数据中心和/或云服务7120)。网络7140可以包括任何数量或类型的通信网络，包括例如局域网、广域网、公共网络、互联网、蜂窝网络、Wi-Fi网络、短距离网络(例如，蓝牙或ZigBee)和/或任何其他有线或无线网络或通信介质。

[0432] 系统7100的任何、全部或一些计算设备可以被适配成执行任何操作系统，包括Linux或其他基于UNIX的操作系统、Microsoft Windows、Windows Server、MacOS、Apple iOS、Google Android或者任何定制的和/或专有的操作系统、以及适配于特定操作系统的虚拟化执行的虚拟机。

[0433] 虽然图71被描述成包含多个单元或者与多个单元相关联，但是在本公开的每一种替代实现方式中可能并不会利用在图71的系统7100内图示的所有单元。附加地，结合图71的实例描述的一个或多个单元可以位于系统7100的外部，而在其他实例中，某些单元可以被包括在所描述的一个或多个其他单元以及未在所图示的实现方式中描述的其他单元之内或者作为其一部分。此外，在图71中图示的某些单元可以与其他组件相组合，以及被用于除了本文中所描述的那些目的之外的替代或附加目的。

[0434] 图72图示了包括经由数据中心网络7210互连的节点的系统7200。例如，如所示出的，多个节点7202、7204和7206可以经由数据中心网络7210而耦合。这种节点可以被实例化并且用于将推理模型(例如，神经网络等)的多个层并行化。在一个示例中，每个节点可以被实现为服务器，并且可以进一步包括至少一个硬件节点或电路(例如，FPGA、ASIC、处理器等)。因此，节点7202可以包括电路7222，节点7204可以包括电路7224，并且节点7206可以包括电路7226。

[0435] 该电路可以经由基于光传输层协议的系统而互连。在一个示例中，电路7222的第一实例可以经由传输链路132与电路7224的第一实例耦合，并且电路7222的第一实例可以进一步经由传输链路7234与电路7224的第二实例耦合。电路7224的第一实例可以经由传输链路7236与电路7226的第一实例耦合，并且电路7224的第一实例可以进一步经由传输链路7240与电路7226的第一实例耦合。类似地，电路7224的第二实例可以经由传输链路7242与电路7226的第一实例耦合，并且电路7224的第二实例可以进一步经由传输链路7238与电路7226的第二实例耦合。光传输层协议可以向该电路提供经由数据中心网络7210彼此传输或接收分组或其他这种数据的能力。该电路也可以以其他配置而互连。例如，电路7222的若干个实例可以经由多个传输链路7252耦合到电路7224的若干个实例。类似地，电路7224的若干个实例可以经由传输链路7254耦合到电路7226的若干个实例。尽管图72示出了包括电路的某个数量和布置的节点，但是可以存在更多或更少数量的被不同布置的节点。

[0436] 图73图示了包括分布式节点的系统7300，该系统可以被实现为数据中心(例如，数据中心7120)中的服务器机架(例如，系统7000、系统100等)。尽管该图示出了以某种方式布

置的系统的某个数量的组件，但是可以存在更多或更少数量的被不同布置的组件。

[0437] 每个服务器可以耦合到架顶式(TOR)交换机。尽管未示出，但是其他机架可以具有类似的配置。每个服务器可以包括至少一个节点或多个节点。每个节点可以包括服务器(例如，服务器7304、服务器7306或服务器7308)，并且每个服务器可以耦合到TOR交换机(例如，TOR交换机7310)。服务器7304可以包括主机组件，该主机组件包括处理器，诸如处理器7314和处理器7316，这些处理器可以经由本地链路7320耦合到硬件节点、例如电路7318。每个硬件节点也可以通过网络接口控制器7322而耦合(例如，用于跨数据中心的网络基础设施进行通信)。系统7300可以允许节点对从TOR交换机或其他交换机接收到(和/或发送到TOR交换机或其他交换机)的推理工作负载执行处理。在使用该示例系统的情况下，个体节点可以用于训练推理模型和/或从推理模型生成推理，如上所描述的那样。例如，可以跨系统7300内的各种处理资源(例如，处理器、电路等)来对推理模型工作负载的处理进行分区。

[0438] 通过拆分跨多个节点的层或通过拆分跨多个节点的层内的神经权重来对推理模型(例如，神经网络等)进行并行化。作为示例，单个推理模型(例如，RNN、LSTM权重矩阵等)可以被分区并且跨多个节点而固定。在该示例的实现方式中，RNN模型可以跨多个电路中的每个电路的存储器(例如，片上存储器)而分布。

[0439] 可以使用表述“在一个示例中”或“示例”以及其派生物来描述一些示例。这些术语意味着结合该示例描述的特定特征、结构或特性被包括在至少一个示例中。短语“在一个示例中”在说明书中的各种地方的出现不一定全部指代相同的示例。

[0440] 可以使用表述“耦合”和“连接”以及其派生物来描述一些示例。这些术语不一定意图作为彼此的同义词。例如，使用术语“连接”和/或“耦合”的描述可以指示两个或更多个元件与彼此直接物理或电气接触。然而，术语“耦合”还可以意味着两个或更多个元件不与彼此直接接触，但是仍然与彼此协作或交互。

[0441] 此外，在前述具体实施方式中，在单个示例中将各种特征分组在一起以简化本公开。本公开的该方法不应当被解释为反映以下意图：所要求保护的示例需要比明确记载在每个权利要求中的特征更多的特征。而是，如以下权利要求所反映的，发明主题在于少于单个所公开的示例的全部特征。因此，以下权利要求在此并入到具体实施方式中，其中每个权利要求独立地作为单独的实施例。在所附权利要求中，术语“包含”和“在其中”分别用作相应的术语“包括”和“其中”的简明英语等同物。而且，术语“第一”、“第二”、“第三”等仅仅用作标记，并且不意图对其对象强加数值要求。

[0442] 尽管已经以特定于结构特征和/或方法动作的语言描述了该主题，但是要理解的是，在所附权利要求中限定的主题不一定限于上面所描述的具体特征或动作。而是，上面描述的具体特征和动作作为实现权利要求的示例形式而公开。

[0443] 可以使用表述“一个实施例”或“实施例”以及其派生物来描述一些实施例。这些术语意味着结合该实施例描述的特定特征、结构或特性被包括在至少一个实施例中。短语“在一个实施例中”在说明书中的各种地方的出现不一定全都指代相同的实施例。此外，可以使用表述“耦合”和“连接”以及其派生物来描述一些实施例。这些术语不一定意图作为彼此的同义词。例如，可以使用术语“连接”和/或“耦合”来描述一些实施例，以指示两个或更多个元件与彼此直接物理接触或电气接触。然而，术语“耦合”还可以意味着两个或更多个元件不与彼此直接接触，但是仍然与彼此协作或交互。此外，来自不同实施例的方面或元件可以

被组合。

[0444] 要强调的是，提供了本公开的摘要以允许读者快速确定技术公开的性质。所提出的理解是，它将不被用来解释或限制权利要求的范围或含义。另外，在前述具体实施方式中，可以看出的是，出于简化本公开的目的，在单个实施例中将各种特征分组在一起。本公开的该方法不应被解释为反映以下意图：所要求保护的示例需要比明确记载在每个权利要求中的特征更多的特征。而是，如以下权利要求所反映的，发明主题在于少于单个所公开的实施例的所有特征。因此，以下权利要求在此被并入到具体实施方式中，其中每个权利要求独立地作为单独的实施例。在所附权利要求中，术语“包含”和“在其中”分别用作相应的术语“包括”和“其中”的简明英语等同物。而且，术语“第一”、“第二”、“第三”等仅仅用作标记，并且不意图对其对象强加数值要求。

[0445] 上面已经描述的内容包括所公开的架构的示例。当然，不可能描述组件和/或方法的每个可想到的组合，但是本领域的普通技术人员可以认识到，许多进一步的组合和排列是可能的。因此，新颖的架构意图涵盖落入所附权利要求的精神和范围内的所有这种变更、修改和变化。详细的公开现在转向提供关于进一步实施例的示例。下面提供的示例不意图是限制性的。

[0446] 示例1.一种装置，包括：电路；以及耦合到所述电路的存储器，存储器存储指令，所述指令在由所述电路执行时使所述电路：部分地基于输入数据和推理模型来生成输出数据；确定所生成的输出的推理概率；确定所述推理概率是否大于阈值；以及基于所述推理概率不大于阈值的确定而提供空的输出。

[0447] 示例2.一种装置，包括：电路；耦合到所述电路的存储器，存储器存储指令，所述指令在由所述电路执行时使所述电路：接收输入数据；经由推理模型、基于接收到的输入数据来生成输出；以及部分地基于输入数据和输出数据来提供对所述推理模型的使用未经授权的指示。

[0448] 示例3.一种装置，包括：电路；耦合到所述电路的接口，所述接口耦合到网络；以及耦合到所述电路的存储器，存储器存储指令，所述指令在由所述电路执行时使所述电路：从输入设备接收输入数据；经由所述接口从耦合到网络的服务器接收净化有效负载，所述净化有效负载包括净化协议的指示；部分地基于通过将所述净化协议应用于输入数据来生成净化后的输入数据；以及经由所述接口和网络将净化后的输入数据发送到服务器。

[0449] 示例4.一种装置，包括：电路；以及耦合到所述电路的存储器，存储器存储指令，所述指令在由所述电路执行时使所述电路：从输入设备接收输入数据；部分地基于利用输入数据执行推理模型来生成输出数据，输出数据包括可见类或隐藏类之一的指示；确定输出数据是否包括隐藏类的指示；以及基于输出数据不包括隐藏类的指示的确定而将所生成的输出提供给输出消费者。

[0450] 示例5.一种装置，包括：电路；以及存储指令的存储器，所述指令在由所述电路执行时使所述电路：部分地基于密钥来将输入数据混淆；以及执行推理模型以使用混淆的输入数据来生成输出数据。

[0451] 示例6.一种装置，包括：电路；以及存储指令的存储器，所述指令在由所述电路执行时使所述电路：使用输入数据来执行推理模型以生成(1)包括推理的指示的输出数据和(2)推理增强图；基于输入数据和推理增强图来生成输入区域影响图；以及至少基于输入区

域影响图来生成可靠性得分，所述可靠性得分包括所述推理的可靠性的指示。

[0452] 示例7.一种装置，包括：电路；以及存储指令的存储器，所述指令在由所述电路执行时使所述电路：确定输入设备的输入设备特性；基于输入设备特性来选择经训练的推理模型；以及利用基于输入设备特性而选择的推理模型对来自输入设备的输入数据进行分类。

[0453] 示例8.一种装置，包括：电路；耦合到所述电路的存储器，存储器存储指令，所述指令在由所述电路执行时使所述电路：从输入设备接收输入数据；部分地基于利用输入数据执行推理模型来生成输出数据；以及对输出数据进行过滤以生成经筛查的输出数据。

[0454] 示例9.一种装置，包括：电路；以及存储指令的存储器，所述指令在由所述电路执行时使所述电路：查询第一服务器以确定第一服务器的能力；部分地基于第一服务器的能力将推理模型拆分成多个模型切片；以及向第一服务器发送信息元素，所述信息元素包括所述多个模型切片之一的指示、以及在第一服务器处以分布式方式执行所述多个模型切片之一的指令。

[0455] 示例10.一种装置，包括：电路；以及存储指令的存储器，所述指令在由所述电路执行时使所述电路：在第一服务器处接收用以确定第一服务器的能力的查询；以及在第一服务器处接收信息元素，所述信息元素包括所述多个模型切片之一的指示、以及在第一服务器处以分布式方式执行所述多个模型切片之一的指令，所述推理模型部分地基于第一服务器的能力而被拆分成多个模型切片。

[0456] 示例11.一种装置，包括：电路；以及存储指令的存储器，所述指令在由所述电路执行时使所述电路：接收输入数据；基于推理模型对输入数据进行分类以产生分类数据；基于分类数据将输入数据评估为不可接受的；当分类数据被评估为不可接受时，修改输入数据以生成增强的输入数据；基于所述推理模型对增强的输入数据进行分类以产生增强的分类数据；基于增强的分类数据将增强的输入数据评估为可接受的；以及提供增强的分类数据作为输出数据。

[0457] 示例12.一种装置，包括：电路；以及存储指令的存储器，所述指令在由所述电路执行时使所述电路：初始化与推理模型中的每一层的每个输出特征图相对应的一个或多个矩阵；将数据集中的样本的激活累积到所述一个或多个矩阵中以产生激活统计信息；以及确定第一、第二和第三参数，其中第一参数包括被固定的激活的百分比，第二参数包括推理模型中应当被激发的节点的总百分比，并且第三参数包括要从其中取得固定激活的总顶部激活的百分位数；以及基于激活统计信息、第一参数、第二参数和第三参数来实现专用激活函数。

[0458] 示例13.一种装置，包括：电路；以及耦合到所述电路的存储器，存储器存储指令，所述指令在由所述电路执行时使所述电路：将推理模型拆分成多个模型切片；以及部分地基于多个对抗性模型来训练所述多个模型切片。

[0459] 示例14. 示例1至13中任一项所述的装置，包括耦合到所述电路的可信执行环境(TEE)，所述TEE包括TEE电路和可信存储器。

[0460] 示例15.一种装置，包括：电路；耦合到所述电路的存储器，存储器存储指令，所述指令在由所述电路执行时使所述电路部分地基于输入数据和推理模型来生成输出数据；以及可信执行环境(TEE)，包括：电路；以及耦合到所述电路的可信存储器，所述可信存储器存

储水印和TEE指令，所述TEE指令在由所述电路执行时使所述电路：确定所述输入是否与所述水印匹配；以及基于所述输入与所述水印匹配的确定来向所述电路发送不生成输出数据的第一指示。

[0461] 示例16.一种装置，包括：电路；存储指令的存储器，所述指令在由所述电路执行时使所述电路：确定与所述装置相关联的唯一标识(UUID)；部分地基于UUID对输入数据进行变换；以及可信执行环境(TEE)，其包括电路，用以：部分地基于预加载的UUID和经变换的输入数据来生成经恢复的输入数据，经恢复的输入数据用于经由推理模型来生成输出。

[0462] 示例17．示例2至16中任一项所述的装置，存储器进一步存储指令，所述指令在由所述电路执行时使所述电路：部分地基于输入数据和推理模型来生成输出数据；确定所生成的输出的推理概率；确定所述推理概率是否大于阈值；以及基于所述推理概率不大于阈值的确定而提供空的输出。

[0463] 示例18．示例1或3至16中任一项所述的装置，存储器进一步存储指令，所述指令在由所述电路执行时使所述电路：接收输入数据；经由推理模型、基于接收到的输入数据来生成输出；以及部分地基于输入数据和输出数据来提供对所述推理模型的使用未经授权的指示。

[0464] 示例19．示例1至2中任一项或示例4至16中任一项所述的装置，包括：耦合到所述电路的接口，所述接口耦合到网络；存储器进一步存储指令，所述指令在由所述电路执行时使所述电路：从输入设备接收输入数据；经由所述接口从耦合到网络的服务器接收净化有效负载，所述净化有效负载包括净化协议的指示；部分地基于通过将所述净化协议应用于输入数据来生成净化后的输入数据；以及经由所述接口和网络将净化后的输入数据发送到服务器。

[0465] 示例20．示例1至3中任一项或示例5至16中任一项所述的装置，存储器进一步存储指令，所述指令在由所述电路执行时使所述电路：从输入设备接收输入数据；部分地基于利用输入数据执行推理模型来生成输出数据，输出数据包括可见类或隐藏类之一的指示；确定输出数据是否包括隐藏类的指示；以及基于输出数据不包括隐藏类的指示的确定而将所生成的输出提供给输出消费者。

[0466] 示例21．示例1至4中任一项或示例6至16中任一项所述的装置，存储器进一步存储指令，所述指令在由所述电路执行时使所述电路：部分地基于密钥来将输入数据混淆；以及执行推理模型以使用混淆的输入数据来生成输出数据。

[0467] 示例22．示例1至5中任一项或示例7至16中任一项所述的装置，存储器进一步存储指令，所述指令在由所述电路执行时使所述电路：使用输入数据来执行推理模型以生成(1)包括推理的指示的输出数据和(2)推理增强图；基于输入数据和推理增强图来生成输入区域影响图；以及至少基于输入区域影响图来生成可靠性得分，所述可靠性得分包括所述推理的可靠性的指示。

[0468] 示例23．示例1至6中任一项或示例8至16中任一项所述的装置，存储器进一步存储指令，所述指令在由所述电路执行时使所述电路：确定输入设备的输入设备特性；基于输入设备特性来选择经训练的推理模型；以及利用基于输入设备特性而选择的推理模型对来自输入设备的输入数据进行分类。

[0469] 示例24．示例1至7中任一项或示例9至16中任一项所述的装置，存储器进一步存

储指令，所述指令在由所述电路执行时使所述电路：从输入设备接收输入数据；部分地基于利用输入数据执行推理模型来生成输出数据；以及对输出数据进行过滤以生成经筛查的输出数据。

[0470] 示例25．示例1至8中任一项或示例10至16中任一项所述的装置，存储器进一步存储指令，所述指令在由所述电路执行时使所述电路：查询第一服务器以确定第一服务器的能力；部分地基于第一服务器的能力将推理模型拆分成多个模型切片；以及向第一服务器发送信息元素，所述信息元素包括所述多个模型切片之一的指示、以及在第一服务器处以分布式方式执行所述多个模型切片之一的指令。

[0471] 示例26．示例1至9中任一项或示例11至16中任一项所述的装置，存储器进一步存储指令，所述指令在由所述电路执行时使所述电路：在第一服务器处接收用以确定第一服务器的能力的查询；以及在第一服务器处接收信息元素，所述信息元素包括所述多个模型切片之一的指示、以及在第一服务器处以分布式方式执行所述多个模型切片之一的指令，所述推理模型部分地基于第一服务器的能力而被拆分成多个模型切片。

[0472] 示例27．示例1至10中任一项或示例12至16中任一项所述的装置，存储器进一步存储指令，所述指令在由所述电路执行时使所述电路：接收输入数据；基于推理模型对输入数据进行分类以产生分类数据；基于分类数据将输入数据评估为不可接受的；当分类数据被评估为不可接受时，修改输入数据以生成增强的输入数据；基于所述推理模型对增强的输入数据进行分类以产生增强的分类数据；基于增强的分类数据将增强的输入数据评估为可接受的；以及提供增强的分类数据作为输出数据。

[0473] 示例28．示例1至11中任一项或示例13至16中任一项所述的装置，存储器进一步存储指令，所述指令在由所述电路执行时使所述电路：初始化与所述推理模型中的每一层的每个输出特征图相对应的一个或多个矩阵；将数据集中的样本的激活累积到所述一个或多个矩阵中以产生激活统计信息；以及确定第一、第二和第三参数，其中第一参数包括被固定的激活的百分比，第二参数包括所述推理模型中应当被激发的节点的总百分比，并且第三参数包括要从其中取得固定激活的总顶部激活的百分位数；以及基于激活统计信息、第一参数、第二参数和第三参数来实现专用激活函数。

[0474] 示例29．示例1至12中任一项或示例14至16中任一项所述的装置，存储器进一步存储指令，所述指令在由所述电路执行时使所述电路：将推理模型拆分成多个模型切片；以及部分地基于多个对抗性模型来训练所述多个模型切片。

[0475] 示例30．示例14或16中任一项所述的装置，所述TEE指令在由所述TEE电路执行时使所述TEE电路：确定所述输入是否与所述水印匹配；以及基于所述输入与所述水印匹配的确定来向所述电路发送不生成输出数据的第一指示。

[0476] 示例31．示例14至15中任一项所述的装置，存储器进一步存储指令，所述指令在由所述电路执行时使所述电路：确定与所述装置相关联的唯一标识(UUID)，以及部分地基于UUID对输入数据进行变换；所述TEE指令在由所述TEE电路执行时使所述TEE电路：部分地基于预加载的UUID和经变换的输入数据来生成经恢复的输入数据，经恢复的输入数据用于经由推理模型来生成输出。

[0477] 示例32．示例1至31中任一项所述的装置，存储器进一步存储指令，所述指令在由所述电路执行时使所述电路：接收输入数据；以及响应于接收到输入数据而生成输出。

[0478] 示例33．示例1至32中任一项所述的装置，存储器进一步存储指令，所述指令在由所述电路执行时使所述电路：基于所述推理概率大于阈值的确定而向输出设备提供所生成的输出。

[0479] 示例34．示例1至33中任一项所述的装置，存储器存储所述推理模型。

[0480] 示例35．示例1至34中任一项所述的装置，其中所述推理模型是神经网络或模式分类器网络。

[0481] 示例36．示例14至35中任一项所述的装置，所述可信存储器进一步存储TEE指令，所述TEE指令在由所述电路执行时使所述电路：基于所述输入与所述水印匹配的确定来向所述电路发送第二指示以提供对所述推理模型的数字权利侵犯的指示。

[0482] 示例37．示例36所述的装置，存储器进一步存储指令，所述指令在由所述电路执行时使所述电路：响应于接收到第二指示来向输出设备提供对所述推理模型的数字权利侵犯的指示；以及响应于接收到第一指示而不生成输出数据。

[0483] 示例38．示例14至37中任一项所述的装置，所述可信存储器进一步存储TEE指令，所述TEE指令在由所述电路执行时使所述电路：将输入数据与所述水印进行逐位比较；以及部分地基于所述逐位比较来确定输入数据与所述水印匹配。

[0484] 示例39．示例14至38中任一项所述的装置，所述可信存储器进一步存储TEE指令，所述TEE指令在由所述电路执行时使所述电路：向所述电路发送第三指示以响应于接收到第三指示而生成输出数据。

[0485] 示例40．示例14至39中任一项所述的装置，存储器进一步存储指令，所述指令在由所述电路执行时使所述电路：响应于接收到第一指示而生成输出数据。

[0486] 示例41．示例14至40中任一项所述的装置，存储器存储所述推理模型。

[0487] 示例42．示例14至41中任一项所述的装置，其中所述推理模型是神经网络。

[0488] 示例43．示例14至42中任一项所述的装置，所述输出包括多个分类之一的指示，所述指令在由所述电路执行时使所述电路：基于输入数据来确定所述多个分类之一是否正确；以及部分地基于根据输入数据确定了所述多个分类之一是正确的，从而确定对所述推理模型的使用未经授权。

[0489] 示例44．示例14至43中任一项所述的装置，所述指令在由所述电路执行时使所述电路：基于对所述推理模型的使用并不是未经授权的确定而提供所述输出。

[0490] 示例45．示例14至44中任一项所述的装置，所述指令在由所述电路执行时使所述电路：接收多个输入数据；经由所述推理模型、基于接收到的多个输入数据中的每一个来生成输出，所述多个输出中的每一个包括多个分类之一的指示；针对所述多个输出中的每一个、基于与所述多个输出之一相对应的所述多个输入数据之一来确定所述多个分类之一是否正确；基于根据与所述多个输出之一相对应的所述多个输入数据之一确定了所述多个分类之一是正确的，从而使计数器递增；以及部分地基于所述计数器超过阈值数量来确定对所述推理模型的使用未经授权。

[0491] 示例46．示例14至45中任一项所述的装置，包括用以捕获输入数据的输入设备。

[0492] 示例47．示例14至46中任一项所述的装置，存储器存储所述推理模型，所述指令在由所述电路执行时使所述电路：执行所述推理模型以部分地基于经恢复的输入数据来生成输出数据。

[0493] 示例48．示例14至47所述的装置，所述TEE存储所述推理模型，所述TEE电路被布置成执行所述推理模型以部分地基于经恢复的输入数据来生成输出数据。

[0494] 示例49．示例1至48中任一项所述的装置，存储器进一步存储指令，所述指令在由所述电路执行时使所述电路：从服务器接收查询，所述查询包括对报告所述电路的计算能力的指示；以及将报告发送给服务器，所述报告包括所述电路的计算能力的指示。

[0495] 示例50．示例1至49中任一项所述的装置，存储器进一步存储指令，所述指令在由所述电路执行时使所述电路：通过所述网络与服务器建立可信连接。

[0496] 示例51．示例1至50中任一项所述的装置，存储器进一步存储指令，所述指令在由所述电路执行时使所述电路：部分地基于将输入数据从第一数据格式转换到第二数据格式来生成经解码的输入数据；部分地基于将过滤器应用于经解码的输入数据来生成过滤后的输入数据；以及部分地基于将过滤后的输入数据从第二数据格式转换到第一数据格式来生成净化后的输入数据。

[0497] 示例52．示例1至51中任一项所述的装置，所述净化有效负载包括第二数据格式和过滤器的指示。

[0498] 示例53．示例1至52中任一项所述的装置，存储器进一步存储指令，所述指令在由所述电路执行时使所述电路：基于输出数据包括隐藏类的指示的确定而向监测系统发送警报。

[0499] 示例54．示例1至53中任一项所述的装置，存储器进一步存储指令，所述指令在由所述电路执行时使所述电路：响应于输出数据包括隐藏类的指示的确定而生成混淆的输出，混淆的输出包括可见类的指示；以及响应于输出数据包括隐藏类的指示的确定而将混淆的输出提供给输出消费者。

[0500] 示例55．示例1至54中任一项所述的装置，存储器进一步存储指令，所述指令在由所述电路执行时使所述电路：响应于输出数据包括隐藏类的指示的确定而收集取证数据，所述取证数据包括输入数据和输入设备中的至少一个的指示；以及响应于输出数据包括隐藏类的指示的确定而向监测系统提供所述取证数据。

[0501] 示例56．示例1至55中任一项所述的装置，可见类包括多个可见类，输出数据包括隐藏类或者所述多个可见类之一的指示。

[0502] 示例57．示例1至56中任一项所述的装置，隐藏类包括多个隐藏类，输出数据包括所述多个隐藏类之一或所述多个可见类之一的指示。

[0503] 示例58．示例1至57中任一项所述的装置，其中所述多个隐藏类中的至少一个对应于列入黑名单的输入。

[0504] 示例59．示例1至58中任一项所述的装置，存储器进一步存储指令，所述指令在由所述电路执行时使所述电路：响应于输出数据包括隐藏类的指示的确定而标识与输出数据所指示的所述多个隐藏类之一相关联的所述多个可见类之一；以及生成混淆的输出以包括所述多个可见类中的所标识的可见类的指示。

[0505] 示例60．示例1至59中任一项所述的装置，包括被布置成捕获输入数据的输入设备，所述指令在由所述电路执行时使所述电路：从输入设备接收输入数据。

[0506] 示例61．示例1至60中任一项所述的装置，包括输入设备，所述指令在由所述电路执行时使所述电路：从输入设备接收输入数据。

[0507] 示例62．示例1至61中任一项所述的装置，其中所述推理增强图是模型激活图。

[0508] 示例63．示例1至62中任一项所述的装置，存储器存储指令，所述指令在由所述电路执行时使所述电路：从输入设备接收输入数据和输入设备特性。

[0509] 示例64．示例1至63中任一项所述的装置，存储器存储指令，所述指令在由所述电路执行时使所述电路：从输入设备请求输入设备特性。

[0510] 示例65．示例1至64中任一项所述的装置，其中所述未经训练的推理模型包括前馈神经网络(FFNN)、递归神经网络(RNN)、卷积神经网络(CNN)、长/短期记忆(LSTM)网络、深度神经网络(DNN)、贝叶斯信念网络和玻尔兹曼机器中的一个或多个。

[0511] 示例66．示例1至65中任一项所述的装置，输出数据包括多个类中的至少一个的指示，存储器进一步存储指令，所述指令在由所述电路执行时使所述电路：确定所述多个类中的至少一个是否被包括在所述多个类的子集中；以及移除对被包括在所述多个类的子集中的所述多个类中的至少一个的指示以对输出数据进行过滤。

[0512] 示例67．示例1至66中任一项所述的装置，包括：耦合到所述电路的接口，所述接口耦合到网络，并且存储器进一步存储指令，所述指令在由所述电路执行时使所述电路：经由所述接口从服务器接收隐私级别规范，所述隐私级别规范包括所述多个类的子集的指示。

[0513] 示例68．示例67所述的装置，存储器进一步存储指令，所述指令在由所述电路执行时使所述电路：经由所述接口向服务器提供经筛查的输出。

[0514] 示例69．示例67至68中任一项所述的装置，存储器进一步存储指令，所述指令在由所述电路执行时使所述电路：从服务器接收查询，所述查询包括对报告所述电路的筛查能力的指示；以及将报告发送给服务器，所述报告包括所述电路的筛查能力的指示。

[0515] 示例70．示例67至69中任一项所述的装置，存储器进一步存储指令，所述指令在由所述电路执行时使所述电路：查询第二服务器以确定第二服务器的能力；部分地基于第一服务器和第二服务器的能力将所述推理模型拆分成多个模型切片；以及向第二服务器发送信息元素，所述信息元素包括所述多个模型切片中的另一个的指示、以及在第二服务器处以分布式方式执行所述多个模型切片中的另一个的指令。

[0516] 示例71．示例67至70中任一项所述的装置，其中第一服务器和第二服务器的能力包括相应的第一服务器和第二服务器的安全性特征集合的指示。

[0517] 示例72．示例67至71中任一项所述的装置，存储器进一步存储指令，所述指令在由所述电路执行时使所述电路：基于所述多个模型切片之一的漏洞级别、以及第一和第二服务器的安全性特征集合，将所述多个模型切片之一指派给第一服务器。

[0518] 示例73．示例1至72中任一项所述的装置，所述推理模型部分地基于第一服务器的能力和第二服务器的能力而被拆分成所述多个模型切片。

[0519] 示例74．示例1至73中任一项所述的装置，其中第一服务器和第二服务器的能力包括相应的第一服务器和第二服务器的安全性特征集合的指示。

[0520] 示例75．示例1至74中任一项所述的装置，基于所述多个模型切片之一的漏洞级别、以及第一和第二服务器的安全性特征集合，所述多个模型切片之一被分配给第一服务器。

[0521] 示例76．示例1至75中任一项所述的装置，分类数据包括输入数据的类以及与所

述类的准确性相关联的置信度水平，并且存储器存储指令，所述指令在由所述电路执行时使所述电路：当置信度水平落在阈值以下时将分类数据评估为不可接受的；以及当置信度水平超过阈值时将分类数据评估为可接受的。

[0522] 示例77．示例1至76中任一项所述的装置，存储器存储指令，所述指令在由所述电路执行时使所述电路：利用伽马的改变、像素移位、去噪和错误校正中的一个或多个来修改输入数据。

[0523] 示例78．示例1至77中任一项所述的装置，存储器存储指令，所述指令在由所述电路执行时使所述电路：比较分类数据和增强的分类数据以检测操作异常。

[0524] 示例79．示例1至78中任一项所述的装置，输入数据包括一个或多个样本的数据集以用于由所述推理模型进行分类。

[0525] 示例80．示例1至79中任一项所述的装置，其中通过将每个样本进行正向传播来累积数据集中的样本的激活，同时存储所述推理模型的每一层的激活图。

[0526] 示例81．示例1至80中任一项所述的装置，其中所述专用激活函数的实现方式包括：传递处于经处理的层的对应矩阵中的所有节点。

[0527] 示例82．示例1至81中任一项所述的装置，其中通过随机地激发所述推理模型中的一个或多个节点来实现激活百分比。

[0528] 示例83．示例1至82中任一项所述的装置，其中所述激活百分比等于第二参数减去第一参数。

[0529] 示例84．示例1至83中任一项所述的装置，存储器存储指令，所述指令在由所述电路执行时使所述电路：基于从用户接收到的输入来确定第一、第二和第三参数。

[0530] 示例85．示例1至84中任一项所述的装置，存储器存储指令，所述指令在由所述电路执行时使所述电路：部分地基于多个替代性子模型和多个对抗性模型来训练所述多个模型切片中的每一个。

[0531] 示例86．示例1至85中任一项所述的装置，其中所述多个替代性子模型包括针对所述多个模型切片中的每一个的替代性子模型。

[0532] 示例87．示例1至86中任一项所述的装置，其中所述多个对抗性模型包括针对所述多个模型切片中的每一个的对抗性模型。

[0533] 示例88．示例1至87中任一项所述的装置，其中所述推理模型包括前馈神经网络(FFNN)、递归神经网络(RNN)、卷积神经网络(CNN)、长/短期记忆(LSTM)网络、深度神经网络(DNN)、贝叶斯信念网络和玻尔兹曼机器中的一个或多个。

[0534] 示例89.一种输入设备，包括：输入捕获电路，其被布置成输出指示所捕获的输入的信号；以及输入混淆电路，其被配置成修改所述信号并且生成表示所修改的信号的混淆的输入数据。

[0535] 示例90．示例80所述的输入设备，所述输入捕获电路是图像捕获传感器，所述输入设备包括：耦合到所述图像捕获传感器的镜头；以及耦合到所述镜头的过滤器，所述过滤器被布置成将传输通过所述镜头的光混淆。

[0536] 示例91.一种非暂时性计算机可读存储介质，包括指令，所述指令在由计算设备的电路执行时使所述计算设备：部分地基于输入数据和推理模型来生成输出数据；确定所生成的输出的推理概率；确定所述推理概率是否大于阈值；以及基于所述推理概率不大于阈

值的确定而提供空的输出。

[0537] 示例92.一种非暂时性计算机可读存储介质,包括指令,所述指令在由计算设备执行时使所述计算设备:从输入设备接收输入数据;经由所述接口从耦合到网络的服务器接收净化有效负载,所述净化有效负载包括净化协议的指示;部分地基于通过将所述净化协议应用于输入数据来生成净化后的输入数据;以及经由接口和网络将净化后的输入数据发送到服务器。

[0538] 示例93.一种非暂时性计算机可读存储介质,包括指令,所述指令在由电路执行时使所述电路:从输入设备接收输入数据;部分地基于利用输入数据执行推理模型来生成输出数据,输出数据包括可见类或隐藏类之一的指示;确定输出数据是否包括隐藏类的指示;以及基于输出数据不包括隐藏类的指示的确定而将所生成的输出提供给输出消费者。

[0539] 示例94.一种非暂时性计算机可读存储介质,包括指令,所述指令在由电路执行时使所述电路:部分地基于密钥来将输入数据混淆;以及执行推理模型以使用混淆的输入数据来生成输出数据。

[0540] 示例95. 一种非暂时性计算机可读存储介质,包括指令,所述指令在由电路执行时使所述电路:使用输入数据来执行推理模型以生成(1)包括推理的指示的输出数据和(2)推理增强图;基于输入数据和推理增强图来生成输入区域影响图;以及至少基于输入区域影响图来生成可靠性得分,所述可靠性得分包括所述推理的可靠性的指示。

[0541] 示例96.一种非暂时性计算机可读存储介质,包括指令,所述指令在由电路执行时使所述电路:确定输入设备的输入设备特性;基于输入设备特性来选择经训练的推理模型;以及利用基于输入设备特性而选择的推理模型对来自输入设备的输入数据进行分类。

[0542] 示例97.一种非暂时性计算机可读存储介质,包括指令,所述指令在由电路执行时使所述电路:从输入设备接收输入数据;部分地基于利用输入数据执行推理模型来生成输出数据;以及对输出数据进行过滤以生成经筛查的输出数据。

[0543] 示例98.一种非暂时性计算机可读存储介质,包括指令,所述指令在由电路执行时使所述电路:查询第一服务器以确定第一服务器的能力;部分地基于第一服务器的能力将推理模型拆分成多个模型切片;以及向第一服务器发送信息元素,所述信息元素包括所述多个模型切片之一的指示、以及在第一服务器处以分布式方式执行所述多个模型切片之一的指令。

[0544] 示例99.一种非暂时性计算机可读存储介质,包括指令,所述指令在由电路执行时使所述电路:在第一服务器处接收用以确定第一服务器的能力的查询;以及在第一服务器处接收信息元素,所述信息元素包括所述多个模型切片之一的指示、以及在第一服务器处以分布式方式执行所述多个模型切片之一的指令,所述推理模型部分地基于第一服务器的能力而被拆分成多个模型切片。

[0545] 示例100.一种非暂时性计算机可读存储介质,包括指令,所述指令在由电路执行时使所述电路:接收输入数据;基于推理模型对输入数据进行分类以产生分类数据;基于分类数据将输入数据评估为不可接受的;当分类数据被评估为不可接受时,修改输入数据以生成增强的输入数据;基于所述推理模型对增强的输入数据进行分类以产生增强的分类数据;基于增强的分类数据将增强的输入数据评估为可接受的;以及提供增强的分类数据作为输出数据。

[0546] 示例101.一种非暂时性计算机可读存储介质,包括指令,所述指令在由电路执行时使所述电路:初始化与所述推理模型中的每一层的每个输出特征图相对应的一个或多个矩阵;将数据集中的样本的激活累积到所述一个或多个矩阵中以产生激活统计信息;以及确定第一、第二和第三参数,其中第一参数包括被固定的激活的百分比,第二参数包括所述推理模型中应当被激发的节点的总百分比,并且第三参数包括要从其中取得固定激活的总顶部激活的百分位数;以及基于激活统计信息、第一参数、第二参数和第三参数来实现专用激活函数。

[0547] 示例102.一种非暂时性计算机可读存储介质,包括指令,所述指令在由电路执行时使所述电路:将推理模型拆分成多个模型切片;以及部分地基于多个对抗性模型来训练所述多个模型切片。

[0548] 示例103. 示例91至102中任一项所述的非暂时性计算机可读存储介质,其中所述电路是可信执行环境(TEE)的一部分。

[0549] 示例104.一种非暂时性计算机可读存储介质,包括指令,所述指令在由计算设备执行时使所述计算设备部分地基于输入数据和推理模型来生成输出数据;并且所述指令在由存储标记的可信执行环境(TEE)执行时使所述TEE确定所述输入是否与所述标记匹配;以及基于所述输入与所述标记匹配的确定来向所述电路发送不生成输出数据的第一指示。

[0550] 示例105.一种非暂时性计算机可读存储介质,包括指令,所述指令在由计算设备执行时使所述计算设备确定与所述装置相关联的唯一标识(UUID);部分地基于UUID对输入数据进行变换;以及使可信执行环境(TEE)部分地基于预加载的UUID和经变换的输入数据来生成经恢复的输入数据,经恢复的输入数据用于经由推理模型来生成输出。

[0551] 示例106. 示例92至105中任一项所述的非暂时性计算机可读存储介质,进一步包括指令,所述指令在被执行时使所述计算设备:部分地基于输入数据和推理模型来生成输出数据;确定所生成的输出的推理概率;确定所述推理概率是否大于阈值;以及基于所述推理概率不大于阈值的确定而提供空的输出。

[0552] 示例107. 示例91或示例93至105中任一项所述的非暂时性计算机可读存储介质,进一步包括指令,所述指令在被执行时使所述计算设备:从输入设备接收输入数据;经由所述接口从耦合到网络的服务器接收净化有效负载,所述净化有效负载包括净化协议的指示;部分地基于通过将所述净化协议应用于输入数据来生成净化后的输入数据;以及经由接口和网络将净化后的输入数据发送到服务器。

[0553] 示例108. 示例91至92中任一项或示例94至105中任一项所述的非暂时性计算机可读存储介质,进一步包括指令,所述指令在被执行时使所述电路:从输入设备接收输入数据;部分地基于利用输入数据执行推理模型来生成输出数据,输出数据包括可见类或隐藏类之一的指示;确定输出数据是否包括隐藏类的指示;以及基于输出数据不包括隐藏类的指示的确定而将所生成的输出提供给输出消费者。

[0554] 示例109. 示例91至93中任一项或示例95至105中任一项所述的非暂时性计算机可读存储介质,进一步包括指令,所述指令在被执行时使所述电路:部分地基于密钥来将输入数据混淆;以及执行推理模型以使用混淆的输入数据来生成输出数据。

[0555] 示例110. 示例91至94中任一项或示例96至105中任一项所述的非暂时性计算机可读存储介质,进一步包括指令,所述指令在被执行时使所述电路:使用输入数据来执行推

理模型以生成(1)包括推理的指示的输出数据和(2)推理增强图;基于输入数据和推理增强图来生成输入区域影响图;以及至少基于输入区域影响图来生成可靠性得分,所述可靠性得分包括所述推理的可靠性的指示。

[0556] 示例111.示例91至95中任一项或示例97至105中任一项所述的非暂时性计算机可读存储介质,进一步包括指令,所述指令在被执行时使所述电路:确定输入设备的输入设备特性;基于输入设备特性来选择经训练的推理模型;以及利用基于输入设备特性而选择的推理模型对来自输入设备的输入数据进行分类。

[0557] 示例112.示例91至96中任一项或示例98至105中任一项所述的非暂时性计算机可读存储介质,进一步包括指令,所述指令在被执行时使所述电路:从输入设备接收输入数据;部分地基于利用输入数据执行推理模型来生成输出数据;以及对输出数据进行过滤以生成经筛查的输出数据。

[0558] 示例113.示例91至97中任一项或示例99至105中任一项所述的非暂时性计算机可读存储介质,进一步包括指令,所述指令在被执行时使所述电路:查询第一服务器以确定第一服务器的能力;部分地基于第一服务器的能力将推理模型拆分成多个模型切片;以及向第一服务器发送信息元素,所述信息元素包括所述多个模型切片之一的指示、以及在第一服务器处以分布式方式执行所述多个模型切片之一的指令。

[0559] 示例114.示例91至98中任一项或示例100至105中任一项所述的非暂时性计算机可读存储介质,进一步包括指令,所述指令在被执行时使所述电路:在第一服务器处接收用以确定第一服务器的能力的查询;以及在第一服务器处接收信息元素,所述信息元素包括所述多个模型切片之一的指示、以及在第一服务器处以分布式方式执行所述多个模型切片之一的指令,所述推理模型部分地基于第一服务器的能力而被拆分成多个模型切片。

[0560] 示例115.示例91至99中任一项或示例101至105中任一项所述的非暂时性计算机可读存储介质,进一步包括指令,所述指令在被执行时使所述电路:接收输入数据;基于推理模型对输入数据进行分类以产生分类数据;基于分类数据将输入数据评估为不可接受的;当分类数据被评估为不可接受时,修改输入数据以生成增强的输入数据;基于所述推理模型对增强的输入数据进行分类以产生增强的分类数据;基于增强的分类数据将增强的输入数据评估为可接受的;以及提供增强的分类数据作为输出数据。

[0561] 示例116.示例91至100中任一项或示例102至105中任一项所述的非暂时性计算机可读存储介质,进一步包括指令,所述指令在被执行时使所述电路:初始化与推理模型中的每一层的每个输出特征图相对应的一个或多个矩阵;将数据集中的样本的激活累积到所述一个或多个矩阵中以产生激活统计信息;以及确定第一、第二和第三参数,其中第一参数包括被固定的激活的百分比,第二参数包括所述推理模型中应当被激发的节点的总百分比,并且第三参数包括要从其中取得固定激活的总顶部激活的百分位数;以及基于激活统计信息、第一参数、第二参数和第三参数来实现专用激活函数。

[0562] 示例117.示例91至101中任一项或示例103至105中任一项所述的非暂时性计算机可读存储介质,进一步包括指令,所述指令在被执行时使所述电路:将推理模型拆分成多个模型切片;以及部分地基于多个对抗性模型来训练所述多个模型切片。

[0563] 示例118.示例103或105中任一项所述的非暂时性计算机可读存储介质,进一步包括指令,所述指令在被执行时使所述电路部分地基于输入数据和推理模型来生成输出数

据;并且所述指令在由存储标记的可信执行环境(TEE)执行时使所述TEE确定所述输入是否与所述标记匹配;以及基于所述输入与所述标记匹配的确定来向所述电路发送不生成输出数据的第一指示。

[0564] 示例119.示例103至104中任一项所述的非暂时性计算机可读存储介质,进一步包括指令,所述指令在被执行时使所述计算设备:确定与所述装置相关联的唯一标识(UUID);部分地基于UUID对输入数据进行变换;以及使可信执行环境(TEE):部分地基于预加载的UUID和经变换的输入数据来生成经恢复的输入数据,经恢复的输入数据用于经由推理模型来生成输出。

[0565] 示例120.示例91至119中任一项所述的非暂时性计算机可读存储介质,进一步包括指令,所述指令在由计算设备执行时使所述计算设备:接收输入数据;以及响应于接收到输入数据而生成输出。

[0566] 示例121.示例91至120中任一项所述的非暂时性计算机可读存储介质,进一步包括指令,所述指令在由计算设备执行时使所述计算设备:基于所述推理概率大于阈值的确定而向输出设备提供所生成的输出。

[0567] 示例122.示例91至121中任一项所述的非暂时性计算机可读存储介质,进一步包括指令,所述指令在由计算设备执行时使所述计算设备:将所述推理模型存储在耦合到所述电路的存储器中。

[0568] 示例123.示例91至122中任一项所述的非暂时性计算机可读存储介质,其中所述推理模型是神经网络。

[0569] 示例124.示例91至123中任一项所述的非暂时性计算机可读存储介质,其中所述推理模型是模式分类器网络。

[0570] 示例125.示例91至124中任一项所述的非暂时性计算机可读存储介质,进一步包括指令,所述指令在由计算设备执行时使所述计算设备:从服务器接收查询,所述查询包括对报告所述电路的计算能力的指示;以及将报告发送给服务器,所述报告包括所述电路的计算能力的指示。

[0571] 示例126.示例91至125中任一项所述的非暂时性计算机可读存储介质,进一步包括指令,所述指令在由计算设备执行时使所述计算设备:通过所述网络与服务器建立可信连接。

[0572] 示例127.示例91至126中任一项所述的非暂时性计算机可读存储介质,进一步包括指令,所述指令在由计算设备执行时使所述计算设备:部分地基于将输入数据从第一数据格式转换到第二数据格式来生成经解码的输入数据;部分地基于将过滤器应用于经解码的输入数据来生成过滤后的输入数据;以及部分地基于将过滤后的输入数据从第二数据格式转换到第一数据格式来生成净化后的输入数据。

[0573] 示例128.示例91至127中任一项所述的非暂时性计算机可读存储介质,所述净化有效负载包括第二数据格式和过滤器的指示。

[0574] 示例129.示例91至128中任一项所述的非暂时性计算机可读存储介质,进一步存储指令,所述指令在由所述电路执行时使所述电路:基于输出数据包括隐藏类的指示的确定而向监测系统发送警报。

[0575] 示例130.示例91至129中任一项所述的非暂时性计算机可读存储介质,进一步存

储指令,所述指令在由所述电路执行时使所述电路:响应于输出数据包括隐藏类的指示的确定而生成混淆的输出,混淆的输出包括可见类的指示;以及响应于输出数据包括隐藏类的指示的确定而将混淆的输出提供给输出消费者。

[0576] 示例131.示例91至130中任一项所述的非暂时性计算机可读存储介质,进一步存储指令,所述指令在由所述电路执行时使所述电路:响应于输出数据包括隐藏类的指示的确定而收集取证数据,所述取证数据包括输入数据和输入设备中的至少一个的指示;以及响应于输出数据包括隐藏类的指示的确定而向监测系统提供所述取证数据。

[0577] 示例132.示例91至131中任一项所述的非暂时性计算机可读存储介质,可见类包括多个可见类,输出数据包括隐藏类或者所述多个可见类之一的指示。

[0578] 示例133.示例91至132中任一项所述的非暂时性计算机可读存储介质,隐藏类包括多个隐藏类,输出数据包括所述多个隐藏类之一或所述多个可见类之一的指示。

[0579] 示例134.示例91至133中任一项所述的非暂时性计算机可读存储介质,其中所述多个隐藏类中的至少一个对应于列入黑名单的输入。

[0580] 示例135.示例91至134中任一项所述的非暂时性计算机可读存储介质,进一步存储指令,所述指令在由所述电路执行时使所述电路:响应于输出数据包括隐藏类的指示的确定而标识与输出数据所指示的所述多个隐藏类之一相关联的所述多个可见类之一;以及生成混淆的输出以包括所述多个可见类中的所标识的可见类的指示。

[0581] 示例136.示例91至135中任一项所述的非暂时性计算机可读存储介质,包括指令,所述指令在由所述电路执行时使所述电路:从输入设备接收输入数据。

[0582] 示例137.示例91至136中任一项所述的非暂时性计算机可读存储介质,其中所述推理增强图是模型激活图。

[0583] 示例138.示例91至137中任一项所述的非暂时性计算机可读存储介质,存储指令,所述指令在由所述电路执行时使所述电路:从输入设备接收输入数据和输入设备特性。

[0584] 示例139.示例91至138中任一项所述的非暂时性计算机可读存储介质,存储指令,所述指令在由所述电路执行时使所述电路:从输入设备请求输入设备特性。

[0585] 示例140.示例91至139中任一项所述的非暂时性计算机可读存储介质,其中所述未经训练的推理模型包括前馈神经网络(FFNN)、递归神经网络(RNN)、卷积神经网络(CNN)、长/短期记忆(LSTM)网络、深度神经网络(DNN)、贝叶斯信念网络和玻尔兹曼机器中的一个或多个。

[0586] 示例141.示例91至140中任一项所述的非暂时性计算机可读存储介质,输出数据包括多个类中的至少一个的指示,所述指令在由所述电路执行时使所述电路:确定所述多个类中的至少一个是否被包括在所述多个类的子集中;以及移除对被包括在所述多个类的子集中的所述多个类中的至少一个的指示以对输出数据进行过滤。

[0587] 示例142.示例91至141中任一项所述的非暂时性计算机可读存储介质,包括耦合到所述电路的接口,所述接口耦合到网络,并且所述指令在由所述电路执行时使所述电路:经由所述接口从服务器接收隐私级别规范,所述隐私级别规范包括所述多个类的子集的指示。

[0588] 示例143.示例91至142中任一项所述的非暂时性计算机可读存储介质,所述指令在由所述电路执行时使所述电路:经由所述接口向服务器提供经筛查的输出。

[0589] 示例144．示例91至143中任一项所述的非暂时性计算机可读存储介质，所述指令在由所述电路执行时使所述电路：从服务器接收查询，所述查询包括对报告所述电路的筛查能力的指示；以及将报告发送给服务器，所述报告包括所述电路的筛查能力的指示。

[0590] 示例145．示例91至144中任一项所述的非暂时性计算机可读存储介质，所述指令在由所述电路执行时使所述电路：查询第二服务器以确定第二服务器的能力；部分地基于第一服务器和第二服务器的能力将所述推理模型拆分成多个模型切片；以及向第二服务器发送信息元素，所述信息元素包括所述多个模型切片中的另一个的指示、以及在第二服务器处以分布式方式执行所述多个模型切片中的另一个的指令。

[0591] 示例146．示例91至145中任一项所述的非暂时性计算机可读存储介质，其中第一服务器和第二服务器的能力包括相应的第一服务器和第二服务器的安全性特征集合的指示。

[0592] 示例147．示例91至146中任一项所述的非暂时性计算机可读存储介质，所述指令在由所述电路执行时使所述电路：基于所述多个模型切片之一的漏洞级别、以及第一和第二服务器的安全性特征集合，将所述多个模型切片之一指派给第一服务器。

[0593] 示例148．示例91至147中任一项所述的非暂时性计算机可读存储介质，所述推理模型部分地基于第一服务器的能力和第二服务器的能力而被拆分成多个模型切片。

[0594] 示例149．示例91至148中任一项所述的非暂时性计算机可读存储介质，其中第一服务器和第二服务器的能力包括相应的第一服务器和第二服务器的安全性特征集合的指示。

[0595] 示例150．示例91至149中任一项所述的非暂时性计算机可读存储介质，基于所述多个模型切片之一的漏洞级别、以及第一和第二服务器的安全性特征集合，所述多个模型切片之一被分配给第一服务器。

[0596] 示例151．示例91至150中任一项所述的非暂时性计算机可读存储介质，所述分类数据包括输入数据的类以及与所述类的准确性相关联的置信度水平，并且所述介质存储指令，所述指令在由所述电路执行时使所述电路：当置信度水平落在阈值以下时将分类数据评估为不可接受的；以及当置信度水平超过阈值时将分类数据评估为可接受的。

[0597] 示例152．示例91至151中任一项所述的非暂时性计算机可读存储介质，所述指令在由所述电路执行时使所述电路：利用伽马的改变、像素移位、去噪和错误校正中的一个或多个来修改输入数据。

[0598] 示例153．示例91至152中任一项所述的非暂时性计算机可读存储介质，所述指令在由所述电路执行时使所述电路：比较分类数据和增强的分类数据以检测操作异常。

[0599] 示例154．示例91至153中任一项所述的非暂时性计算机可读存储介质，其中所述推理模型包括前馈神经网络(FFNN)、递归神经网络(RNN)、卷积神经网络(CNN)、长/短期记忆(LSTM)网络、深度神经网络(DNN)、贝叶斯信念网络和玻尔兹曼机器中的一个或多个。

[0600] 示例155．示例91至154中任一项所述的非暂时性计算机可读存储介质，输入数据包括一个或多个样本的数据集以用于由所述推理模型进行分类。

[0601] 示例156．示例91至155中任一项所述的非暂时性计算机可读存储介质，其中通过将每个样本进行正向传播来累积数据集中的样本的激活，同时存储所述推理模型的每一层的激活图。

[0602] 示例157．示例91至156中任一项所述的非暂时性计算机可读存储介质，其中所述专用激活函数的实现方式包括：传递处于经处理的层的对应矩阵中的所有节点。

[0603] 示例158．示例91至157中任一项所述的非暂时性计算机可读存储介质，其中通过随机地激发所述推理模型中的一个或多个节点来实现激活百分比。

[0604] 示例159．示例91至158中任一项所述的非暂时性计算机可读存储介质，其中所述激活百分比等于第二参数减去第一参数。

[0605] 示例160．示例91至159中任一项所述的非暂时性计算机可读存储介质，所述指令在由所述电路执行时使所述电路：基于从用户接收到的输入来确定第一、第二和第三参数。

[0606] 示例161．示例91至160中任一项所述的非暂时性计算机可读存储介质，其中所述推理模型包括前馈神经网络(FFNN)、递归神经网络(RNN)、卷积神经网络(CNN)、长/短期记忆(LSTM)网络、深度神经网络(DNN)、贝叶斯信念网络和玻尔兹曼机器中的一个或多个。

[0607] 示例162．示例91至161中任一项所述的非暂时性计算机可读存储介质，所述指令在由所述电路执行时使所述电路：部分地基于多个替代性子模型和多个对抗性模型来训练所述多个模型切片中的每一个。

[0608] 示例163．示例91至162中任一项所述的非暂时性计算机可读存储介质，其中所述多个替代性子模型包括针对所述多个模型切片中的每一个的替代性子模型。

[0609] 示例164．示例91至163中任一项所述的非暂时性计算机可读存储介质，其中所述多个对抗性模型包括针对所述多个模型切片中的每一个的对抗性模型。

[0610] 示例165．示例103至164中任一项所述的非暂时性计算机可读存储介质，进一步包括指令，所述指令在由所述TEE执行时使所述TEE：基于所述输入与所述标记匹配的确定来向所述电路发送第二指示以提供对所述推理模型的数字权利侵犯的指示。

[0611] 示例166．示例103至165中任一项所述的非暂时性计算机可读存储介质，进一步包括指令，所述指令在由所述电路执行时使所述电路：响应于接收到第二指示来向输出设备提供对所述推理模型的数字权利侵犯的指示；以及响应于接收到第一指示而不生成输出数据。

[0612] 示例167．示例103至166中任一项所述的非暂时性计算机可读存储介质，进一步包括指令，所述指令在由所述TEE执行时使所述TEE：将输入数据与所述标记进行逐位比较；以及部分地基于所述逐位比较来确定输入数据与所述标记匹配。

[0613] 示例168．示例103至167中任一项所述的非暂时性计算机可读存储介质，进一步包括指令，所述指令在由所述TEE执行时使所述TEE：向所述电路发送第三指示以响应于接收到第三指示而生成输出数据。

[0614] 示例169．示例103至168中任一项所述的非暂时性计算机可读存储介质，进一步包括指令，所述指令在由所述电路执行时使所述电路：响应于接收到第一指示而生成输出数据。

[0615] 示例170．示例103至169中任一项所述的非暂时性计算机可读存储介质，进一步包括所述推理模型。

[0616] 示例171．示例103至170中任一项所述的非暂时性计算机可读存储介质，其中所述推理模型是神经网络。

[0617] 示例172．示例103至171中任一项所述的非暂时性计算机可读存储介质，进一步

包括指令，所述指令在由计算设备执行时使所述计算设备：从输入设备接收输入数据。

[0618] 示例173．示例103至172中任一项所述的非暂时性计算机可读存储介质，进一步包括所述推理模型，所述指令在由所述电路执行时使所述电路：执行所述推理模型以部分地基于经恢复的输入数据来生成输出数据。

[0619] 示例174．示例103至173中任一项所述的非暂时性计算机可读存储介质，所述TEE存储所述推理模型，进一步包括指令，所述指令在由所述TEE执行时使所述TEE：部分地基于经恢复的输入数据来生成输出数据。

[0620] 示例175.一种计算机实现方法，包括：部分地基于输入数据和推理模型来生成输出数据；确定所生成的输出的推理概率；确定所述推理概率是否大于阈值；以及基于所述推理概率不大于阈值的确定而提供空的输出。

[0621] 示例176.一种计算机实现的方法，包括：接收输入数据；经由推理模型、基于接收到的输入数据来生成输出；以及部分地基于输入数据和输出数据来提供对所述推理模型的使用未经授权的指示。

[0622] 示例177.一种计算机实现方法，包括：从输入设备接收输入数据；经由接口从耦合到网络的服务器接收净化有效负载，所述净化有效负载包括净化协议的指示；部分地基于通过将所述净化协议应用于输入数据来生成净化后的输入数据；以及经由接口和网络将净化后的输入数据发送到服务器。

[0623] 示例178.一种计算机实现方法，包括：从输入设备接收输入数据；部分地基于利用输入数据执行推理模型来生成输出数据，输出数据包括可见类或隐藏类之一的指示；确定输出数据是否包括隐藏类的指示；以及基于输出数据不包括隐藏类的指示的确定而将所生成的输出提供给输出消费者。

[0624] 示例179.一种计算机实现方法，包括：部分地基于密钥来将输入数据混淆；以及执行推理模型以使用混淆的输入数据来生成输出数据。

[0625] 示例180.一种计算机实现方法，包括：使用输入数据来执行推理模型以生成(1)包括推理的指示的输出数据和(2)推理增强图；基于输入数据和推理增强图来生成输入区域影响图；以及至少基于输入区域影响图来生成可靠性得分，所述可靠性得分包括所述推理的可靠性的指示。

[0626] 示例181.一种计算机实现方法，包括：确定输入设备的输入设备特性；基于输入设备特性来选择经训练的推理模型；以及利用基于输入设备特性而选择的推理模型对来自输入设备的输入数据进行分类。

[0627] 示例182.一种计算机实现方法，包括：从输入设备接收输入数据；部分地基于利用输入数据执行推理模型来生成输出数据；以及对输出数据进行过滤以生成经筛查的输出数据。

[0628] 示例183.一种计算机实现方法，包括：查询第一服务器以确定第一服务器的能力；部分地基于第一服务器的能力将推理模型拆分成多个模型切片；以及向第一服务器发送信息元素，所述信息元素包括所述多个模型切片之一的指示、以及在第一服务器处以分布式方式执行所述多个模型切片之一的指令。

[0629] 示例184.一种计算机实现方法，包括：在第一服务器处接收用以确定第一服务器的能力的查询；以及在第一服务器处接收信息元素，所述信息元素包括所述多个模型切片

之一的指示、以及在第一服务器处以分布式方式执行所述多个模型切片之一的指令，所述推理模型部分地基于第一服务器的能力而被拆分成多个模型切片。

[0630] 示例185.一种计算机实现方法，包括：接收输入数据；基于推理模型对输入数据进行分类以产生分类数据；基于分类数据将输入数据评估为不可接受的；当分类数据被评估为不可接受时，修改输入数据以生成增强的输入数据；基于所述推理模型对增强的输入数据进行分类以产生增强的分类数据；基于增强的分类数据将增强的输入数据评估为可接受的；以及提供增强的分类数据作为输出数据。

[0631] 示例186.一种计算机实现方法，包括：初始化与推理模型中的每一层的每个输出特征图相对应的一个或多个矩阵；将数据集中的样本的激活累积到所述一个或多个矩阵中以产生激活统计信息；以及确定第一、第二和第三参数，其中第一参数包括被固定的激活的百分比，第二参数包括所述推理模型中应当被激发的节点的总百分比，并且第三参数包括要从其中取得固定激活的总顶部激活的百分位数；以及基于激活统计信息、第一参数、第二参数和第三参数来实现专用激活函数。

[0632] 示例187.一种计算机实现方法，包括：将推理模型拆分成多个模型切片；以及部分地基于多个对抗性模型来训练所述多个模型切片。

[0633] 示例188.一种计算机实现方法，包括：部分地基于输入数据和推理模型来生成输出数据；经由存储标记的可信执行环境(TEE)来确定所述输入是否与所述标记匹配；以及基于所述输入与所述标记匹配的确定来从所述TEE向所述电路发送不生成输出数据的第一指示。

[0634] 示例189.一种计算机实现方法，包括：确定与所述装置相关联的唯一标识(UUID)；部分地基于UUID对输入数据进行变换；以及使得可信执行环境(TEE)：部分地基于预加载的UUID和经变换的输入数据来生成经恢复的输入数据，经恢复的输入数据用于经由推理模型来生成输出。

[0635] 示例190. 示例176至189中任一项所述的计算机实现方法，包括：部分地基于输入数据和推理模型来生成输出数据；确定所生成的输出的推理概率；确定所述推理概率是否大于阈值；以及基于所述推理概率不大于阈值的确定而提供空的输出。

[0636] 示例191. 示例175或示例177至189中任一项所述的计算机实现方法，包括：接收输入数据；经由推理模型、基于接收到的输入数据来生成输出；以及部分地基于输入数据和输出数据来提供对所述推理模型的使用未经授权的指示。

[0637] 示例192. 示例175至176中任一项或示例178至189中任一项所述的计算机实现方法，包括：从输入设备接收输入数据；经由所述接口从耦合到网络的服务器接收净化有效负载，所述净化有效负载包括净化协议的指示；部分地基于通过将所述净化协议应用于输入数据来生成净化后的输入数据；以及经由接口和网络将净化后的输入数据发送到服务器。

[0638] 示例192. 示例175至177中任一项或示例179至189中任一项所述的计算机实现方法，包括：从输入设备接收输入数据；部分地基于利用输入数据执行推理模型来生成输出数据，输出数据包括可见类或隐藏类之一的指示；确定输出数据是否包括隐藏类的指示；以及基于输出数据不包括隐藏类的指示的确定而将所生成的输出提供给输出消费者。

[0639] 示例193. 示例175至178中任一项或示例180至189中任一项所述的计算机实现方法，包括：部分地基于密钥来将输入数据混淆；以及执行推理模型以使用混淆的输入数据来

生成输出数据。

[0640] 示例194．示例175至179中任一项或示例181至189中任一项所述的计算机实现方法，包括：使用输入数据来执行推理模型以生成(1)包括推理的指示的输出数据和(2)推理增强图；基于输入数据和推理增强图来生成输入区域影响图；以及至少基于输入区域影响图来生成可靠性得分，所述可靠性得分包括所述推理的可靠性的指示。

[0641] 示例195．示例175至180中任一项或示例182至189中任一项所述的计算机实现方法，包括：确定输入设备的输入设备特性；基于输入设备特性来选择经训练的推理模型；以及利用基于输入设备特性而选择的推理模型对来自输入设备的输入数据进行分类。

[0642] 示例196．示例175至181中任一项或示例183至189中任一项所述的计算机实现方法，包括：从输入设备接收输入数据；部分地基于利用输入数据执行推理模型来生成输出数据；以及对输出数据进行过滤以生成经筛查的输出数据。

[0643] 示例197．示例175至182中任一项或示例184至189中任一项所述的计算机实现方法，包括：查询第一服务器以确定第一服务器的能力；部分地基于第一服务器的能力将推理模型拆分成多个模型切片；以及向第一服务器发送信息元素，所述信息元素包括所述多个模型切片之一的指示、以及在第一服务器处以分布式方式执行所述多个模型切片之一的指令。

[0644] 示例198．示例175至183中任一项或示例185至189中任一项所述的计算机实现方法，包括：在第一服务器处接收用以确定第一服务器的能力的查询；以及在第一服务器处接收信息元素，所述信息元素包括所述多个模型切片之一的指示、以及在第一服务器处以分布式方式执行所述多个模型切片之一的指令，所述推理模型部分地基于第一服务器的能力而被拆分成多个模型切片。

[0645] 示例199．示例175至184中任一项或示例186至189中任一项所述的计算机实现方法，包括：接收输入数据；基于推理模型对输入数据进行分类以产生分类数据；基于分类数据将输入数据评估为不可接受的；当分类数据被评估为不可接受时，修改输入数据以生成增强的输入数据；基于所述推理模型对增强的输入数据进行分类以产生增强的分类数据；基于增强的分类数据将增强的输入数据评估为可接受的；以及提供增强的分类数据作为输出数据。

[0646] 示例200．示例175至185中任一项或示例187至189中任一项所述的计算机实现方法，包括：初始化与推理模型中的每一层的每个输出特征图相对应的一个或多个矩阵；将数据集中的样本的激活累积到所述一个或多个矩阵中以产生激活统计信息；以及确定第一、第二和第三参数，其中第一参数包括被固定的激活的百分比，第二参数包括所述推理模型中应当被激发的节点的总百分比，并且第三参数包括要从其中取得固定激活的总顶部激活的百分位数；以及基于激活统计信息、第一参数、第二参数和第三参数来实现专用激活函数。

[0647] 示例201．示例175至186中任一项或示例188至189中任一项所述的计算机实现方法，包括：将推理模型拆分成多个模型切片；以及部分地基于多个对抗性模型来训练所述多个模型切片。

[0648] 示例202．示例175至187或示例189中任一项所述的计算机实现方法，包括：部分地基于输入数据和推理模型来生成输出数据；经由存储标记的可信执行环境(TEE)来确定

所述输入是否与所述标记匹配;以及基于所述输入与所述标记匹配的确定来从所述TEE向所述电路发送不生成输出数据的第一指示。

[0649] 示例203．示例175至188中任一项所述的计算机实现方法,包括:确定与所述装置相关联的唯一标识(UUID);部分地基于UUID对输入数据进行变换;以及使得可信执行环境(TEE):部分地基于预加载的UUID和经变换的输入数据来生成经恢复的输入数据,经恢复的输入数据用于经由推理模型来生成输出。

[0650] 示例204．示例175至203中任一项所述的计算机实现方法,包括:接收输入数据;以及响应于接收到输入数据而生成输出。

[0651] 示例205．示例175至204中任一项所述的计算机实现方法,包括:基于所述推理概率大于阈值的确定而向输出设备提供所生成的输出。

[0652] 示例206．示例175至205中任一项所述的计算机实现方法,包括将所述推理模型存储在存储器中。

[0653] 示例207．示例175至206中任一项所述的计算机实现方法,其中所述推理模型是神经网络。

[0654] 示例208．示例175至207中任一项所述的计算机实现方法,其中所述推理模型是模式分类器网络。

[0655] 示例209．示例175至208中任一项所述的计算机实现方法,输出包括多个分类之一的指示,所述方法包括:基于输入数据来确定所述多个分类之一是否正确;以及部分地基于根据输入数据确定了所述多个分类之一是正确的,从而确定对所述推理模型的使用未经授权。

[0656] 示例210．示例175至209中任一项所述的计算机实现方法,包括:基于对所述推理模型的使用并不是未经授权的确定而提供所述输出。

[0657] 示例211．示例175至210中任一项所述的计算机实现方法,包括:接收多个输入数据;经由所述推理模型、基于接收到的多个输入数据中的每一个来生成输出,所述多个输出中的每一个包括多个分类之一的指示;针对所述多个输出中的每一个、基于与所述多个输出之一相对应的所述多个输入数据之一来确定所述多个分类之一是否正确;基于根据与所述多个输出之一相对应的所述多个输入数据之一确定了所述多个分类之一是正确的,从而使计数器递增;以及部分地基于所述计数器超过阈值数量来确定对所述推理模型的使用未经授权。

[0658] 示例212．示例175至211中任一项所述的计算机实现方法,包括:从服务器接收查询,所述查询包括对报告所述电路的计算能力的指示;以及将报告发送给服务器,所述报告包括所述电路的计算能力的指示。

[0659] 示例213．示例175至212中任一项所述的计算机实现方法,包括:通过所述网络与服务器建立可信连接。

[0660] 示例214．示例175至213中任一项所述的计算机实现方法,包括:部分地基于将输入数据从第一数据格式转换到第二数据格式来生成经解码的输入数据;部分地基于将过滤器应用于经解码的输入数据来生成过滤后的输入数据;以及部分地基于将过滤后的输入数据从第二数据格式转换到第一数据格式来生成净化后的输入数据。

[0661] 示例215．示例175至214中任一项所述的计算机实现方法,所述净化有效负载包

括第二数据格式和过滤器的指示。

[0662] 示例216．示例175至215中任一项所述的计算机实现方法，包括：基于输出数据包括隐藏类的指示的确定而向监测系统发送警报。

[0663] 示例217．示例175至216中任一项所述的计算机实现方法，包括：响应于输出数据包括隐藏类的指示的确定而生成混淆的输出，混淆的输出包括可见类的指示；以及响应于输出数据包括隐藏类的指示的确定而将混淆的输出提供给输出消费者。

[0664] 示例218．示例175至217中任一项所述的计算机实现方法，包括：响应于输出数据包括隐藏类的指示的确定而收集取证数据，所述取证数据包括输入数据和输入设备中的至少一个的指示；以及响应于输出数据包括隐藏类的指示的确定而向监测系统提供所述取证数据。

[0665] 示例219．示例175至218中任一项所述的计算机实现方法，可见类包括多个可见类，输出数据包括隐藏类或者所述多个可见类之一的指示。

[0666] 示例220．示例175至219中任一项所述的计算机实现方法，隐藏类包括多个隐藏类，输出数据包括所述多个隐藏类之一或所述多个可见类之一的指示。

[0667] 示例221．示例175至220中任一项所述的计算机实现方法，包括：响应于输出数据包括隐藏类的指示的确定而标识与输出数据所指示的所述多个隐藏类之一相关联的所述多个可见类之一；以及生成混淆的输出以包括所述多个可见类中的所标识的可见类的指示。

[0668] 示例222．示例175至221中任一项所述的计算机实现方法，包括：从输入设备接收输入数据。

[0669] 示例223．示例175至222中任一项所述的计算机实现方法，其中所述推理增强图是模型激活图。

[0670] 示例224．示例175至223中任一项所述的计算机实现方法，包括：从输入设备接收输入数据和输入设备特性。

[0671] 示例225．示例175至224中任一项所述的计算机实现方法，包括：从输入设备请求输入设备特性。

[0672] 示例226．示例175至225中任一项所述的计算机实现方法，其中未经训练的推理模型包括前馈神经网络(FFNN)、递归神经网络(RNN)、卷积神经网络(CNN)、长/短期记忆(LSTM)网络、深度神经网络(DNN)、贝叶斯信念网络和玻尔兹曼机器中的一个或多个。

[0673] 示例227．示例175至226中任一项所述的计算机实现方法，输出数据包括多个类中的至少一个的指示，所述方法包括：确定所述多个类中的至少一个是否被包括在所述多个类的子集中；以及移除对被包括在所述多个类的子集中的所述多个类中的至少一个的指示以对输出数据进行过滤。

[0674] 示例228．示例175至227中任一项所述的计算机实现方法，包括：经由接口从服务器接收隐私级别规范，所述隐私级别规范包括所述多个类的子集的指示。

[0675] 示例229．示例175至228中任一项所述的计算机实现方法，包括：经由所述接口向服务器提供经筛查的输出。

[0676] 示例230．示例175至229中任一项所述的计算机实现方法，包括：从服务器接收查询，所述查询包括对报告所述电路的筛查能力的指示；以及将报告发送给服务器，所述报告

包括所述电路的筛查能力的指示。

[0677] 示例231．示例175至230中任一项所述的计算机实现方法，包括：查询第二服务器以确定第二服务器的能力；部分地基于第一服务器和第二服务器的能力将所述推理模型划分成多个模型切片；以及向第二服务器发送信息元素，所述信息元素包括所述多个模型切片中的另一个的指示、以及在第二服务器处以分布式方式执行所述多个模型切片中的另一个的指令。

[0678] 示例232．示例175至231中任一项所述的计算机实现方法，其中第一服务器和第二服务器的能力包括相应的第一服务器和第二服务器的安全性特征集合的指示。

[0679] 示例233．示例175至232中任一项所述的计算机实现方法，包括：基于所述多个模型切片之一的漏洞级别、以及第一和第二服务器的安全性特征集合，将所述多个模型切片之一指派给第一服务器。

[0680] 示例234．示例175至233中任一项所述的计算机实现方法，所述推理模型部分地基于第一服务器的能力和第二服务器的能力而被拆分成所述多个模型切片。

[0681] 示例235．示例175至234中任一项所述的计算机实现方法，其中第一服务器和第二服务器的能力包括相应的第一服务器和第二服务器的安全性特征集合的指示。

[0682] 示例236．示例175至235中任一项所述的计算机实现方法，其中基于所述多个模型切片之一的漏洞级别、以及第一和第二服务器的安全性特征集合，所述多个模型切片之一被分配给第一服务器。

[0683] 示例237．示例175至236中任一项所述的计算机实现方法，所述分类数据包括输入数据的类以及与所述类的准确性相关联的置信度水平，并且所述介质存储指令，所述指令在由所述电路执行时使所述电路：当置信度水平落在阈值以下时将分类数据评估为不可接受的；以及当置信度水平超过阈值时将分类数据评估为可接受的。

[0684] 示例238．示例175至237中任一项所述的计算机实现方法，包括：利用伽马的改变、像素移位、去噪和错误校正中的一个或多个来修改输入数据。

[0685] 示例239．示例175至238中任一项所述的计算机实现方法，包括：比较分类数据和增强的分类数据以检测操作异常。

[0686] 示例240．示例175至239中任一项所述的计算机实现方法，其中所述推理模型包括前馈神经网络(FFNN)、递归神经网络(RNN)、卷积神经网络(CNN)、长/短期记忆(LSTM)网络、深度神经网络(DNN)、贝叶斯信念网络和玻尔兹曼机器中的一个或多个。

[0687] 示例241．示例175至240中任一项所述的计算机实现方法，输入数据包括一个或多个样本的数据集以用于由所述推理模型进行分类。

[0688] 示例242．示例175至241中任一项所述的计算机实现方法，其中通过将每个样本进行正向传播来累积数据集中的样本的激活，同时存储所述推理模型的每一层的激活图。

[0689] 示例243．示例175至242中任一项所述的计算机实现方法，其中所述专用激活函数的实现方式包括：传递处于经处理的层的对应矩阵中的所有节点。

[0690] 示例244．示例175至243中任一项所述的计算机实现方法，其中通过随机地激发所述推理模型中的一个或多个节点来实现激活百分比。

[0691] 示例245．示例175至244中任一项所述的计算机实现方法，其中所述激活百分比等于第二参数减去第一参数。

[0692] 示例246．示例175至245中任一项所述的计算机实现方法，包括：基于从用户接收到的输入来确定第一、第二和第三参数。

[0693] 示例247．示例175至246中任一项所述的计算机实现方法，其中所述推理模型包括前馈神经网络(FFNN)、递归神经网络(RNN)、卷积神经网络(CNN)、长/短期记忆(LSTM)网络、深度神经网络(DNN)、贝叶斯信念网络和玻尔兹曼机器中的一个或多个。

[0694] 示例248．示例175至247中任一项所述的计算机实现方法，包括：部分地基于多个替代性子模型和多个对抗性模型来训练所述多个模型切片中的每一个。

[0695] 示例249．示例175至248中任一项所述的计算机实现方法，其中所述多个替代性子模型包括针对所述多个模型切片中的每一个的替代性子模型。

[0696] 示例250．示例175至249中任一项所述的计算机实现方法，其中所述多个对抗性模型包括针对所述多个模型切片中的每一个的对抗性模型。

[0697] 示例251．示例188至250中任一项所述的计算机实现方法，包括：基于所述输入与所述标记匹配的确定来向所述电路发送第二指示以提供对所述推理模型的数字权利侵犯的指示。

[0698] 示例252．示例188至251中任一项所述的计算机实现方法，进一步包括指令，所述指令在由所述电路执行时使所述电路：响应于接收到第二指示来向输出设备提供对所述推理模型的数字权利侵犯的指示；以及响应于接收到第一指示而不生成输出数据。

[0699] 示例253．示例188至252中任一项所述的计算机实现方法，进一步包括指令，所述指令在由所述TEE执行时使所述TEE：将输入数据与所述标记进行逐位比较；以及部分地基于所述逐位比较来确定输入数据与所述标记匹配。

[0700] 示例254．示例188至253中任一项所述的计算机实现方法，进一步包括指令，所述指令在由所述TEE执行时使所述TEE：向所述电路发送第三指示以响应于接收到第三指示而生成输出数据。

[0701] 示例255．示例188至254中任一项所述的计算机实现方法，进一步包括指令，所述指令在由所述电路执行时使所述电路：响应于接收到第一指示而生成输出数据。

[0702] 示例256．示例188至255中任一项所述的计算机实现方法，进一步包括存储器中的所述推理模型。

[0703] 示例257．示例188至256中任一项所述的计算机实现方法，其中所述推理模型是神经网络。

[0704] 示例258．示例188至257中任一项所述的计算机实现方法，包括：从输入设备接收输入数据。

[0705] 示例259．示例188至258中任一项所述的计算机实现方法，包括：执行所述推理模型以部分地基于经恢复的输入数据来生成输出数据。

[0706] 示例260．示例188至259中任一项所述的计算机实现方法，所述TEE存储所述推理模型，所述方法包括：使所述TEE部分地基于经恢复的输入数据来生成输出数据。

[0707] 示例261．一种装置，包括：电路；耦合到所述电路的存储器，存储器存储指令，所述指令在由所述电路执行时使所述电路：接收用于推理模型的训练数据，所述训练数据包括多个输入和多个对抗性输入的指示，所述训练数据进一步包括针对所述多个输入和所述多个对抗性输入中的每一个的预期输出的指示；以及部分地基于所述训练数据来训练所述推

理模型。

[0708] 示例262.一种装置,包括:电路;以及存储指令的存储器,所述指令在由所述电路执行时使所述电路:接收具有多个样本的训练数据集;基于训练数据集中的样本来针对两个或更多个推理模型生成分类数据;确定针对所述两个或更多个推理模型中的每一个而生成的分类数据是否匹配;当针对所述两个或更多个推理模型中的每一个而生成的分类数据未能匹配时,将训练数据集中的样本添加到不相关的输出数据;以及当针对所述两个或更多个推理模型中的每一个而生成的分类数据匹配时,将训练数据集中的样本添加到相关的输出数据。

[0709] 示例263.一种装置,包括:电路;以及存储指令的存储器,所述指令在由所述电路执行时使所述电路:确定输入设备的输入设备特性;基于输入设备特性来选择未经训练的推理模型;以及利用来自输入设备的输入数据来训练基于输入设备特性而选择的未经训练的推理模型。

[0710] 示例264.一种装置,包括:电路;以及存储指令的存储器,所述指令在由所述电路执行时使所述电路:接收输入数据;对输入数据进行转码;基于经转码的输入数据来生成训练数据集完整性状态(TSIS)值;以及提供TSIS值作为输出,以启用关于基于所述输入数据训练了推理模型的验证。

[0711] 示例265.一种装置,包括:电路;以及存储指令的存储器,所述指令在由所述电路执行时使所述电路:在用于训练推理模型的训练数据集中检测至少一个错误;接收恢复训练数据;以及利用所述恢复训练数据来训练所述推理模型。

[0712] 示例266.一种装置,包括:电路;以及存储指令的存储器,所述指令在由所述电路执行时使所述电路:将推理模型拆分成至少公共模型分区和私有模型分区;以及部分地基于对抗性模型来训练公共模型分区。

[0713] 示例267. 示例262至266中任一项所述的装置,存储器进一步存储指令,所述指令在由所述电路执行时使所述电路:接收用于推理模型的训练数据,所述训练数据包括多个输入和多个对抗性输入的指示,所述训练数据进一步包括针对所述多个输入和所述多个对抗性输入中的每一个的预期输出的指示;以及部分地基于所述训练数据来训练所述推理模型。

[0714] 示例268. 示例261中任一项或示例263到266中任一项所述的装置,存储器进一步存储指令,所述指令在由所述电路执行时使所述电路:接收具有多个样本的训练数据集;基于训练数据集中的样本来针对两个或更多个推理模型生成分类数据;确定针对所述两个或更多个推理模型中的每一个而生成的分类数据是否匹配;当针对所述两个或更多个推理模型中的每一个而生成的分类数据未能匹配时,将训练数据集中的样本添加到不相关的输出数据;以及当针对所述两个或更多个推理模型中的每一个而生成的分类数据匹配时,将训练数据集中的样本添加到相关的输出数据。

[0715] 示例269. 示例261至262中任一项或示例264至266中任一项所述的装置,存储器进一步存储指令,所述指令在由所述电路执行时使所述电路:确定输入设备的输入设备特性;基于输入设备特性来选择未经训练的推理模型;以及利用来自输入设备的输入数据来训练基于输入设备特性而选择的未经训练的推理模型。

[0716] 示例270. 示例261至263中任一项或示例265至266中任一项所述的装置,存储器

进一步存储指令，所述指令在由所述电路执行时使所述电路：接收输入数据；对输入数据进行转码；基于经转码的输入数据来生成训练数据集完整性状态(TSIS)值；以及提供TSIS值作为输出，以启用关于基于所述输入数据训练了推理模型的验证。

[0717] 示例271．示例261至264或示例266中任一项所述的装置，存储器进一步存储指令，所述指令在由所述电路执行时使所述电路：在用于训练推理模型的训练数据集中检测至少一个错误；接收恢复训练数据；以及利用所述恢复训练数据来训练所述推理模型。

[0718] 示例272．示例261至265中任一项所述的装置，存储器进一步存储指令，所述指令在由所述电路执行时使所述电路：将推理模型拆分成至少公共模型分区和私有模型分区；以及部分地基于对抗性模型来训练公共模型分区。

[0719] 示例273．示例261至272中任一项所述的装置，所述多个预期输出中的每一个对应于多个类之一。

[0720] 示例274．示例261至273中任一项所述的装置，存储器存储指令，所述指令在由所述电路执行时使所述电路：利用相关的输出数据集来训练可信推理模型。

[0721] 示例275．示例261至274中任一项所述的装置，存储器存储指令，所述指令在由所述电路执行时使所述电路：在一个或多个训练流程中利用不相关的输出数据。

[0722] 示例276．示例261至275中任一项所述的装置，所述一个或多个训练流程包括弹性训练流程或对抗性训练流程。

[0723] 示例277．示例261至276中任一项所述的装置，存储器存储指令，所述指令在由所述电路执行时使所述电路：确定训练数据集中尚未被添加到不相关的输出数据或相关的输出数据之一的剩余样本；基于训练数据集中的所述剩余样本来针对两个或更多个推理模型生成附加分类数据；确定针对所述两个或更多个推理模型中的每一个而生成的附加分类数据是否匹配；当针对所述两个或更多个推理模型中的每一个而生成的附加分类数据未能匹配时，将训练数据集中的所述剩余样本添加到不相关的输出数据；以及当针对所述两个或更多个推理模型中的每一个而生成的附加分类数据匹配时，将训练数据集中的所述剩余样本添加到相关的输出数据。

[0724] 示例278．示例261至277中任一项所述的装置，其中所述两个或更多个推理模型包括前馈神经网络(FFNN)、递归神经网络(RNN)、卷积神经网络(CNN)、长/短期记忆(LSTM)网络、深度神经网络(DNN)、贝叶斯信念网络和玻尔兹曼机器中的两个或更多个。

[0725] 示例279．示例261至278中任一项所述的装置，存储器存储指令，所述指令在由所述电路执行时使所述电路：从输入设备接收输入数据和输入设备特性。

[0726] 示例280．示例261至279中任一项所述的装置，存储器存储指令，所述指令在由所述电路执行时使所述电路：从输入设备请求输入设备特性。

[0727] 示例281．示例261至280中任一项所述的装置，其中未经训练的推理模型包括前馈神经网络(FFNN)、递归神经网络(RNN)、卷积神经网络(CNN)、长/短期记忆(LSTM)网络、贝叶斯信念网络和玻尔兹曼机器中的一个或多个。

[0728] 示例282．示例261至281中任一项所述的装置，存储器存储指令，所述指令在由所述电路执行时使所述电路：基于输入数据来训练所述推理模型；以及在输出中提供所述推理模型。

[0729] 示例283．示例261至282中任一项所述的装置，存储器存储指令，所述指令在由所

述电路执行时使所述电路:基于所述推理模型来生成推理模型完整性状态(IMIS)值;以及在输出中提供IMIS值。

[0730] 示例284.示例261至283中任一项所述的装置,其中在输出中提供TSIS值、IMIS值和所述推理模型作为元组。

[0731] 示例285.示例261至284中任一项所述的装置,存储器存储指令,所述指令在由所述电路执行时使所述电路:将输入数据变换成数学表示以对输入数据进行转码。

[0732] 示例286.示例261至285中任一项所述的装置,存储器存储指令,所述指令在由所述电路执行时使所述电路:使用密码函数来生成TSIS值。

[0733] 示例287.示例261至286中任一项所述的装置,存储器存储指令,所述指令在由所述电路执行时使所述电路:接收训练数据集;以及利用训练数据集来训练所述推理模型。

[0734] 示例288.示例261至287中任一项所述的装置,所述恢复训练数据包括多个输入的指示,并且针对所述多个输入中的每一个包括对应的输出。

[0735] 示例289.示例261至288中任一项所述的装置,其中所述多个输出对应于至少一个通配符类,所述至少一个通配符类被布置成将与至少一个错误相关联的输入分类成所述至少一个通配符类。

[0736] 示例290.示例261至289中任一项所述的装置,存储器存储指令,所述指令在由所述电路执行时使所述电路:将所述推理模型的层$L_1$和$L_2$添加到私有模型分区;以及从公共模型分区中移除所述推理模型的层$L_1$和$L_2$。

[0737] 示例291.示例261至290中任一项所述的装置,存储器存储指令,所述指令在由所述电路执行时使所述电路:评估私有模型分区的性能成本,所述性能成本部分地基于在可信执行环境中执行私有模型分区;确定所述性能成本是否超过性能预算;以及基于所述性能成本没有超过性能预算的确定而将所述推理模型的层$L_3$添加到私有模型分区、并且从公共模型分区中移除所述推理模型的层$L_3$。

[0738] 示例292.示例261至291中任一项所述的装置,存储器存储指令,所述指令在由所述电路执行时使所述电路:基于所述性能成本超过了性能预算的确定而从私有模型分区中移除所述推理模型的层$L_2$、并且将所述推理模型的层$L_2$添加到公共模型分区。

[0739] 示例293.一种非暂时性计算机可读存储介质,包括指令,所述指令在由计算设备执行时使所述计算设备:接收用于推理模型的训练数据,所述训练数据包括多个输入和多个对抗性输入的指示,所述训练数据进一步包括针对所述多个输入和所述多个对抗性输入中的每一个的预期输出的指示;以及部分地基于所述训练数据来训练所述推理模型。

[0740] 示例294.一种非暂时性计算机可读存储介质,包括指令,所述指令在由电路执行时使所述电路:接收具有多个样本的训练数据集;基于训练数据集中的样本来针对两个或更多个推理模型生成分类数据;确定针对所述两个或更多个推理模型中的每一个而生成的分类数据是否匹配;当针对所述两个或更多个推理模型中的每一个而生成的分类数据未能匹配时,将训练数据集中的样本添加到不相关的输出数据;以及当针对所述两个或更多个推理模型中的每一个而生成的分类数据匹配时,将训练数据集中的样本添加到相关的输出数据。

[0741] 示例295.一种非暂时性计算机可读存储介质,包括指令,所述指令在由电路执行时使所述电路:确定输入设备的输入设备特性;基于输入设备特性来选择未经训练的推理

模型;以及利用来自输入设备的输入数据来训练基于输入设备特性而选择的未经训练的推理模型。

[0742] 示例296.一种非暂时性计算机可读存储介质,包括指令,所述指令在由电路执行时使所述电路:接收输入数据;对输入数据进行转码;基于经转码的输入数据来生成训练数据集完整性状态(TSIS)值;以及提供TSIS值作为输出,以启用关于基于所述输入数据训练了推理模型的验证。

[0743] 示例297.一种非暂时性计算机可读存储介质,包括指令,所述指令在由电路执行时使所述电路:在用于训练推理模型的训练数据集中检测至少一个错误;接收恢复训练数据;以及利用所述恢复训练数据来训练所述推理模型。

[0744] 示例298.一种非暂时性计算机可读存储介质,包括指令,所述指令在由电路执行时使所述电路:将推理模型拆分成至少公共模型分区和私有模型分区;以及部分地基于对抗性模型来训练公共模型分区。

[0745] 示例299. 示例294至298中任一项所述的非暂时性计算机可读存储介质,进一步存储指令,所述指令在由所述电路执行时使所述电路:接收用于推理模型的训练数据,所述训练数据包括多个输入和多个对抗性输入的指示,所述训练数据进一步包括针对所述多个输入和所述多个对抗性输入中的每一个的预期输出的指示;以及部分地基于所述训练数据来训练所述推理模型。

[0746] 示例300. 示例293或示例295至298中任一项所述的非暂时性计算机可读存储介质,进一步存储指令,所述指令在由所述电路执行时使所述电路:接收具有多个样本的训练数据集;基于训练数据集中的样本来针对两个或更多个推理模型生成分类数据;确定针对所述两个或更多个推理模型中的每一个而生成的分类数据是否匹配;当针对所述两个或更多个推理模型中的每一个而生成的分类数据未能匹配时,将训练数据集中的样本添加到不相关的输出数据;以及当针对所述两个或更多个推理模型中的每一个而生成的分类数据匹配时,将训练数据集中的样本添加到相关的输出数据。

[0747] 示例301. 示例293至294中任一项或示例296至297中任一项所述的非暂时性计算机可读存储介质,进一步包括指令,所述指令在由所述电路执行时使所述电路:确定输入设备的输入设备特性;基于输入设备特性来选择未经训练的推理模型;以及利用来自输入设备的输入数据来训练基于输入设备特性而选择的未经训练的推理模型。

[0748] 示例302. 示例293至295中任一项或示例297至298中任一项所述的非暂时性计算机可读存储介质,进一步包括指令,所述指令在由所述电路执行时使所述电路:接收输入数据;对输入数据进行转码;基于经转码的输入数据来生成训练数据集完整性状态(TSIS)值;以及提供TSIS值作为输出,以启用关于基于所述输入数据训练了推理模型的验证。

[0749] 示例303. 示例293至295或示例298中任一项所述的非暂时性计算机可读存储介质,进一步存储指令,所述指令在由所述电路执行时使所述电路:在用于训练推理模型的训练数据集中检测至少一个错误;接收恢复训练数据;以及利用所述恢复训练数据来训练所述推理模型。

[0750] 示例304. 示例293至297中任一项所述的非暂时性计算机可读存储介质,进一步存储指令,所述指令在由所述电路执行时使所述电路:将推理模型拆分成至少公共模型分区和私有模型分区;以及部分地基于对抗性模型来训练公共模型分区。

[0751] 示例305．示例293至304中任一项所述的非暂时性计算机可读存储介质，所述多个预期输出中的每一个对应于多个类之一。

[0752] 示例306．示例293至305中任一项所述的非暂时性计算机可读存储介质，存储指令，所述指令在由所述电路执行时使所述电路：利用相关的输出数据集来训练可信推理模型。

[0753] 示例307．示例293至306中任一项所述的非暂时性计算机可读存储介质，存储指令，所述指令在由所述电路执行时使所述电路：在一个或多个训练流程中利用不相关的输出数据。

[0754] 示例308．示例293至307中任一项所述的非暂时性计算机可读存储介质，所述一个或多个训练流程包括弹性训练流程或对抗性训练流程。

[0755] 示例309．示例293至308中任一项所述的非暂时性计算机可读存储介质，存储指令，所述指令在由所述电路执行时使所述电路：确定训练数据集中尚未被添加到不相关的输出数据或相关的输出数据之一的剩余样本；基于训练数据集中的所述剩余样本来针对两个或更多个推理模型生成附加分类数据；确定针对所述两个或更多个推理模型中的每一个而生成的附加分类数据是否匹配；当针对所述两个或更多个推理模型中的每一个而生成的附加分类数据未能匹配时，将训练数据集中的所述剩余样本添加到不相关的输出数据；以及当针对所述两个或更多个推理模型中的每一个而生成的附加分类数据匹配时，将训练数据集中的所述剩余样本添加到相关的输出数据。

[0756] 示例310．示例293至309中任一项所述的非暂时性计算机可读存储介质，其中所述两个或更多个推理模型包括前馈神经网络(FFNN)、递归神经网络(RNN)、卷积神经网络(CNN)、长/短期记忆(LSTM)网络、深度神经网络(DNN)、贝叶斯信念网络和玻尔兹曼机器中的两个或更多个。

[0757] 示例311．示例293至310中任一项所述的非暂时性计算机可读存储介质，存储指令，所述指令在由所述电路执行时使所述电路：从输入设备接收输入数据和输入设备特性。

[0758] 示例312．示例293至311中任一项所述的非暂时性计算机可读存储介质，存储指令，所述指令在由所述电路执行时使所述电路：从输入设备请求输入设备特性。

[0759] 示例313．示例293至312中任一项所述的非暂时性计算机可读存储介质，其中未经训练的推理模型包括前馈神经网络(FFNN)、递归神经网络(RNN)、卷积神经网络(CNN)、长/短期记忆(LSTM)网络、贝叶斯信念网络和玻尔兹曼机器中的一个或多个。

[0760] 示例314．示例293至313中任一项所述的非暂时性计算机可读存储介质，所述指令在由所述电路执行时使所述电路：基于输入数据来训练所述推理模型；以及在输出中提供所述推理模型。

[0761] 示例315．示例293至314中任一项所述的非暂时性计算机可读存储介质，所述指令在由所述电路执行时使所述电路：基于所述推理模型来生成推理模型完整性状态(IMIS)值；以及在输出中提供IMIS值。

[0762] 示例316．示例293至315中任一项所述的非暂时性计算机可读存储介质，其中在输出中提供TSIS值、IMIS值和所述推理模型作为元组。

[0763] 示例317．示例293至316中任一项所述的非暂时性计算机可读存储介质，所述指令在由所述电路执行时使所述电路：将输入数据变换成数学表示以对输入数据进行转码。

[0764] 示例318．示例293至317中任一项所述的非暂时性计算机可读存储介质，所述指令在由所述电路执行时使所述电路：使用密码函数来生成TSIS值。

[0765] 示例319．示例293至318中任一项所述的非暂时性计算机可读存储介质，所述指令在由所述电路执行时使所述电路：接收训练数据集；以及利用训练数据集来训练所述推理模型。

[0766] 示例320．示例293至319中任一项所述的非暂时性计算机可读存储介质，所述恢复训练数据包括多个输入的指示，并且针对所述多个输入中的每一个包括对应的输出。

[0767] 示例321．示例293至320中任一项所述的非暂时性计算机可读存储介质，其中所述多个输出对应于至少一个通配符类，所述至少一个通配符类被布置成将与至少一个错误相关联的输入分类成所述至少一个通配符类。

[0768] 示例322．示例293至321中任一项所述的非暂时性计算机可读存储介质，所述指令在由所述电路执行时使所述电路：将所述推理模型的层$L_1$和$L_2$添加到私有模型分区；以及从公共模型分区中移除所述推理模型的层$L_1$和$L_2$。

[0769] 示例323．示例293至322中任一项所述的非暂时性计算机可读存储介质，所述指令在由所述电路执行时使所述电路：评估私有模型分区的性能成本，所述性能成本部分地基于在可信执行环境中执行私有模型分区；确定所述性能成本是否超过性能预算；以及基于所述性能成本没有超过性能预算的确定而将所述推理模型的层$L_3$添加到私有模型分区、并且从公共模型分区中移除所述推理模型的层$L_3$。

[0770] 示例324．示例293至323中任一项所述的非暂时性计算机可读存储介质，所述指令在由所述电路执行时使所述电路：基于所述性能成本超过了性能预算的确定而从私有模型分区中移除所述推理模型的层$L_2$、并且将所述推理模型的层$L_2$添加到公共模型分区。

[0771] 示例325.一种计算机实现方法，包括：接收用于推理模型的训练数据，所述训练数据包括多个输入和多个对抗性输入的指示，所述训练数据进一步包括针对所述多个输入和所述多个对抗性输入中的每一个的预期输出的指示；以及部分地基于所述训练数据来训练所述推理模型。

[0772] 示例326.一种计算机实现方法，包括：接收具有多个样本的训练数据集；基于训练数据集中的样本来针对两个或更多个推理模型生成分类数据；确定针对所述两个或更多个推理模型中的每一个而生成的分类数据是否匹配；当针对所述两个或更多个推理模型中的每一个而生成的分类数据未能匹配时，将训练数据集中的样本添加到不相关的输出数据；以及当针对所述两个或更多个推理模型中的每一个而生成的分类数据匹配时，将训练数据集中的样本添加到相关的输出数据。

[0773] 示例327.一种计算机实现方法，包括：确定输入设备的输入设备特性；基于输入设备特性来选择未经训练的推理模型；以及利用来自输入设备的输入数据来训练基于输入设备特性而选择的未经训练的推理模型。

[0774] 示例328.一种计算机实现方法，包括：接收输入数据；对输入数据进行转码；基于经转码的输入数据来生成训练数据集完整性状态(TSIS)值；以及提供TSIS值作为输出，以启用关于基于所述输入数据训练了推理模型的验证。

[0775] 示例329.一种计算机实现方法，包括：在用于训练推理模型的训练数据集中检测至少一个错误；接收恢复训练数据；以及利用所述恢复训练数据来训练所述推理模型。

[0776] 示例330.一种计算机实现方法,包括:将推理模型拆分成至少公共模型分区和私有模型分区;以及部分地基于对抗性模型来训练公共模型分区。

[0777] 示例331.示例326至330中任一项所述的计算机实现方法,包括:接收用于推理模型的训练数据,所述训练数据包括多个输入和多个对抗性输入的指示,所述训练数据进一步包括针对所述多个输入和所述多个对抗性输入中的每一个的预期输出的指示;以及部分地基于所述训练数据来训练所述推理模型。

[0778] 示例332.示例325或示例327至330中任一项所述的计算机实现方法,包括:接收具有多个样本的训练数据集;基于训练数据集中的样本来针对两个或更多个推理模型生成分类数据;确定针对所述两个或更多个推理模型中的每一个而生成的分类数据是否匹配;当针对所述两个或更多个推理模型中的每一个而生成的分类数据未能匹配时,将训练数据集中的样本添加到不相关的输出数据;以及当针对所述两个或更多个推理模型中的每一个而生成的分类数据匹配时,将训练数据集中的样本添加到相关的输出数据。

[0779] 示例333.示例325至326中任一项或示例328至330中任一项所述的计算机实现方法,包括:确定输入设备的输入设备特性;基于输入设备特性来选择未经训练的推理模型;以及利用来自输入设备的输入数据来训练基于输入设备特性而选择的未经训练的推理模型。

[0780] 示例334.示例325至327中任一项或示例329至330中任一项所述的计算机实现方法,包括:接收输入数据;对输入数据进行转码;基于经转码的输入数据来生成训练数据集完整性状态(TSIS)值;以及提供TSIS值作为输出,以启用关于基于所述输入数据训练了推理模型的验证。

[0781] 示例335.示例325至328或示例330中任一项所述的计算机实现方法,包括在用于训练推理模型的训练数据集中检测至少一个错误;接收恢复训练数据;以及利用所述恢复训练数据来训练所述推理模型。

[0782] 示例336.示例325至329中任一项所述的计算机实现方法,包括:将推理模型拆分成至少公共模型分区和私有模型分区;以及部分地基于对抗性模型来训练公共模型分区。

[0783] 示例337.示例325至336中任一项所述的计算机实现方法,所述多个预期输出中的每一个对应于多个类之一。

[0784] 示例338.示例325至337中任一项所述的计算机实现方法,包括:利用相关的输出数据集来训练可信推理模型。

[0785] 示例339.示例325至338中任一项所述的计算机实现方法,包括:在一个或多个训练流程中利用不相关的输出数据。

[0786] 示例340.示例325至339中任一项所述的计算机实现方法,所述一个或多个训练流程包括弹性训练流程或对抗性训练流程。

[0787] 示例341.示例325至340中任一项所述的计算机实现方法,包括:确定训练数据集中尚未被添加到不相关的输出数据或相关的输出数据之一的剩余样本;基于训练数据集中的所述剩余样本来针对两个或更多个推理模型生成附加分类数据;确定针对所述两个或更多个推理模型中的每一个而生成的附加分类数据是否匹配;当针对所述两个或更多个推理模型中的每一个而生成的附加分类数据未能匹配时,将训练数据集中的所述剩余样本添加到不相关的输出数据;以及当针对所述两个或更多个推理模型中的每一个而生成的附加分

类数据匹配时，将训练数据集中的所述剩余样本添加到相关的输出数据。

[0788] 示例342．示例325至341中任一项所述的计算机实现方法，其中所述两个或更多个推理模型包括前馈神经网络(FFNN)、递归神经网络(RNN)、卷积神经网络(CNN)、长/短期记忆(LSTM)网络、深度神经网络(DNN)、贝叶斯信念网络和玻尔兹曼机器中的两个或更多个。

[0789] 示例343．示例325至342中任一项所述的计算机实现方法，包括：从输入设备接收输入数据和输入设备特性。

[0790] 示例344．示例325至343中任一项所述的计算机实现方法，包括：从输入设备请求输入设备特性。

[0791] 示例345．示例325至344中任一项所述的计算机实现方法，其中未经训练的推理模型包括前馈神经网络(FFNN)、递归神经网络(RNN)、卷积神经网络(CNN)、长/短期记忆(LSTM)网络、贝叶斯信念网络和玻尔兹曼机器中的一个或多个。

[0792] 示例346．示例325至345中任一项所述的计算机实现方法，包括：基于输入数据来训练所述推理模型；以及在输出中提供所述推理模型。

[0793] 示例347．示例325至346中任一项所述的计算机实现方法，包括：基于所述推理模型来生成推理模型完整性状态(IMIS)值；以及在输出中提供IMIS值。

[0794] 示例348．示例325至347中任一项所述的计算机实现方法，其中在输出中提供TSIS值、IMIS值和所述推理模型作为元组。

[0795] 示例349．示例325至348中任一项所述的计算机实现方法，包括：将输入数据变换成数学表示以对输入数据进行转码。

[0796] 示例350．示例325至349中任一项所述的计算机实现方法，包括：使用密码函数来生成TSIS值。

[0797] 示例351．示例325至350中任一项所述的计算机实现方法，包括：接收训练数据集；以及利用训练数据集来训练所述推理模型。

[0798] 示例352．示例325至351中任一项所述的计算机实现方法，所述恢复训练数据包括多个输入的指示，并且针对所述多个输入中的每一个包括对应的输出。

[0799] 示例353．示例325至352中任一项所述的计算机实现方法，其中所述多个输出对应于至少一个通配符类，所述至少一个通配符类被布置成将与至少一个错误相关联的输入分类成所述至少一个通配符类。

[0800] 示例354．示例325至353中任一项所述的计算机实现方法，包括：将所述推理模型的层$L_1$和$L_2$添加到私有模型分区；以及从公共模型分区中移除所述推理模型的层$L_1$和$L_2$。

[0801] 示例355．示例325至354中任一项所述的计算机实现方法，包括：评估私有模型分区的性能成本，所述性能成本部分地基于在可信执行环境中执行私有模型分区；确定所述性能成本是否超过性能预算；以及基于所述性能成本没有超过性能预算的确定而将所述推理模型的层$L_3$添加到私有模型分区、并且从公共模型分区中移除所述推理模型的层$L_3$。

[0802] 示例356．示例325至355中任一项所述的计算机实现方法，包括：基于所述性能成本超过了性能预算的确定而从私有模型分区中移除所述推理模型的层$L_2$、并且将所述推理模型的层$L_2$添加到公共模型分区。

[0803] 示例357.一种装置，包括被布置成实现示例175至260中任一项或示例325至356中

任一项所述的功能的部件。

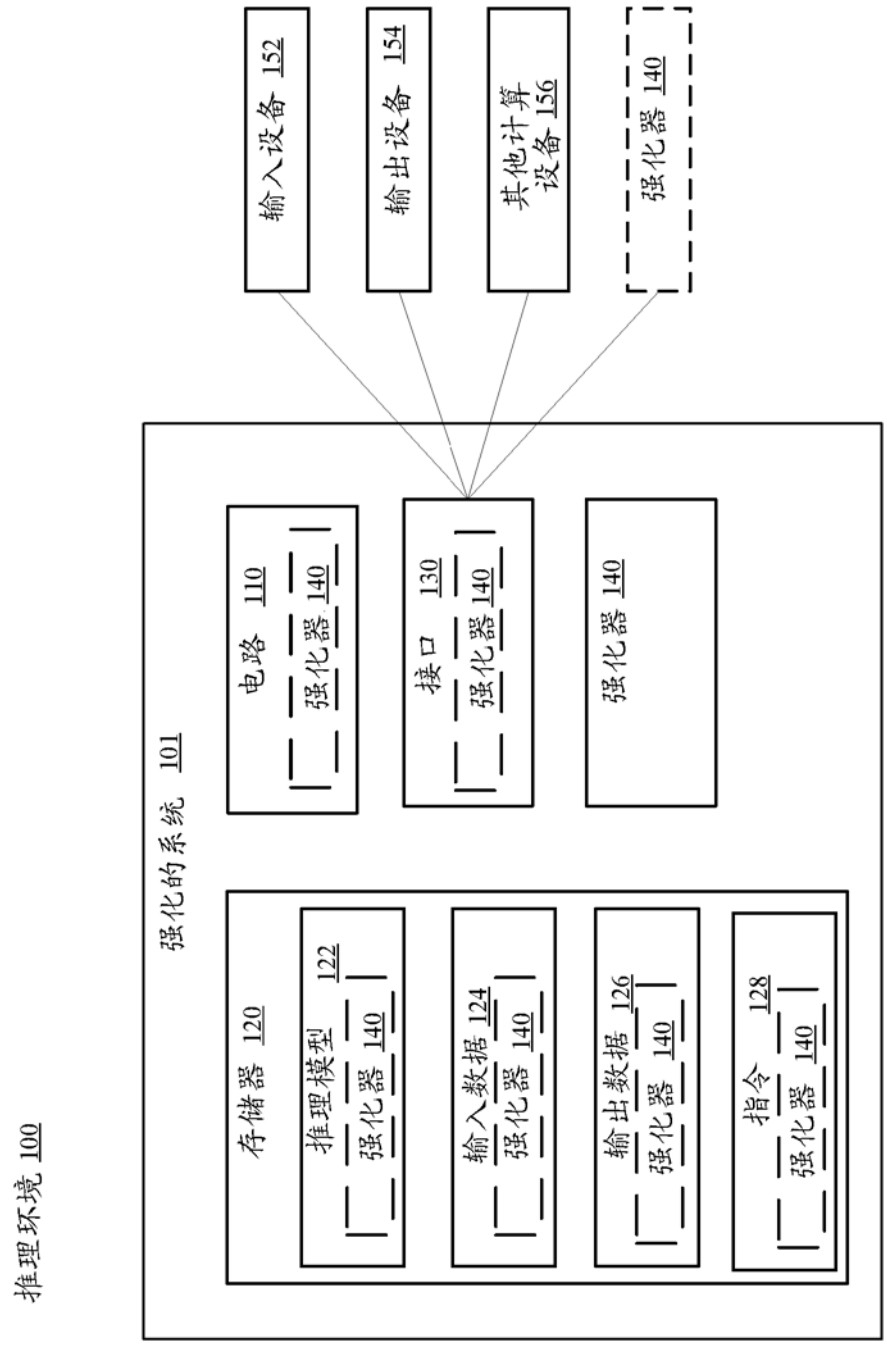

推理环境 100
强化的系统 101
存储器 120
推理模型 122
强化器 140
输入数据 124
强化器 140
输出数据 126
强化器 140
指令 128
强化器 140
电路 110
强化器 140
接口 130
强化器 140
强化器 140
输入设备 152
输出设备 154
其他计算设备 156
强化器 140

图 1

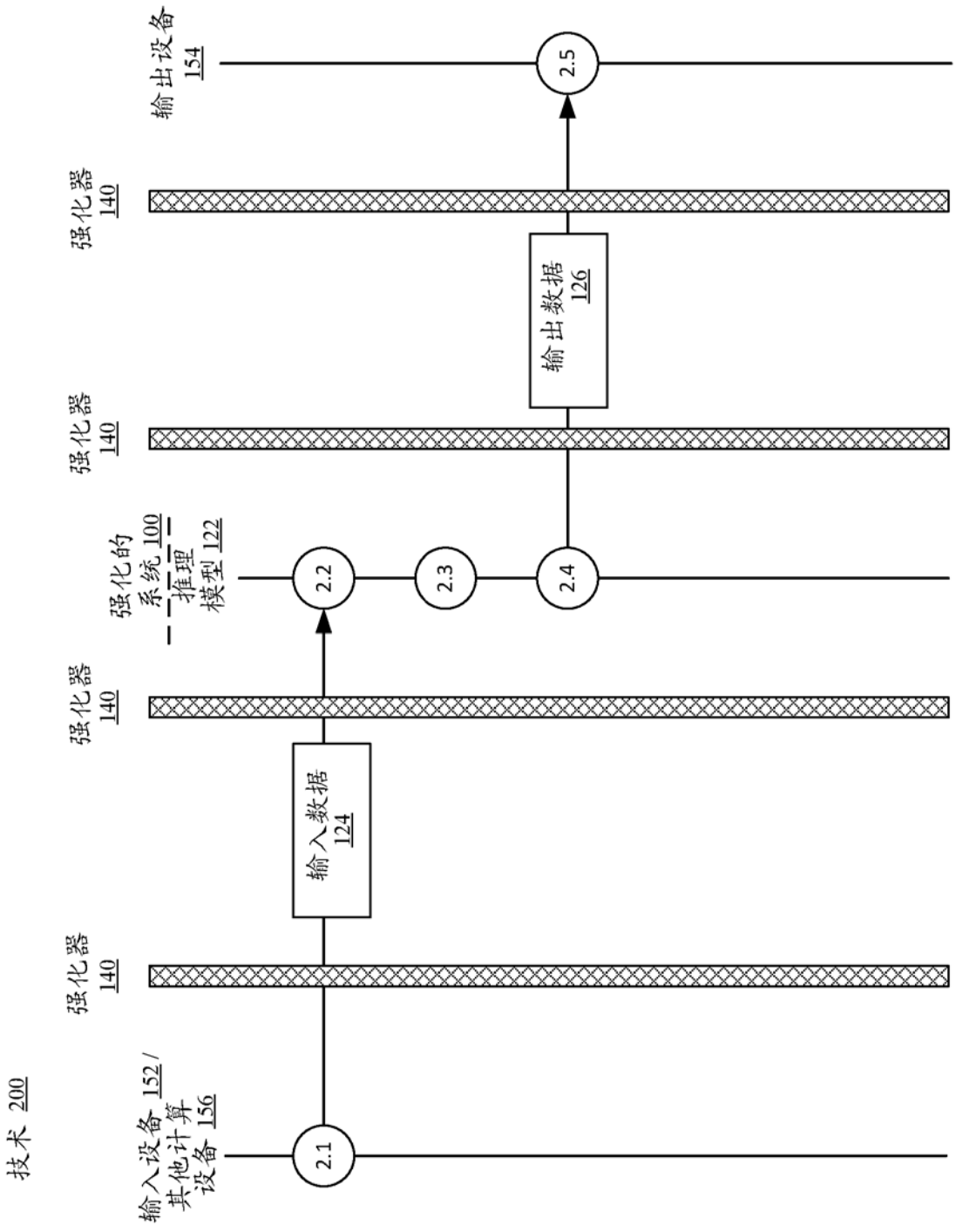

技术 200
输入设备 152 / 其他计算设备 156
强化器 140
输入数据 124
强化器 140
强化的系统 100 推理模型 122
强化器 140
输出数据 126
强化器 140
输出设备 154
2.1
2.2
2.3
2.4
2.5

图 2

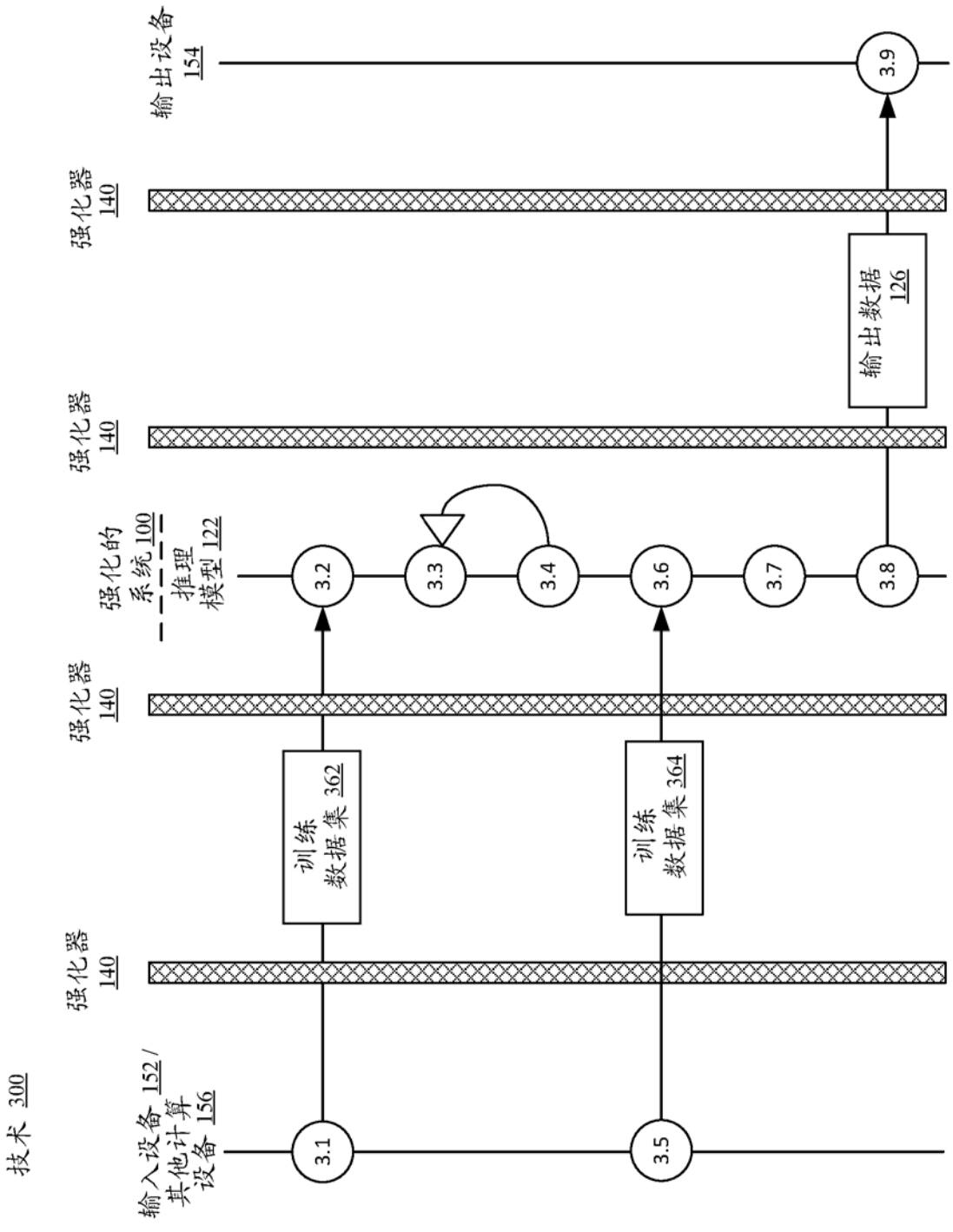

技术 300
输入设备 152 / 其他计算设备 156
强化器 140
训练数据集 362
训练数据集 364
强化的系统 100
推理模型 122
输出数据 126
输出设备 154
3.1
3.2
3.3
3.4
3.5
3.6
3.7
3.8
3.9

图 3

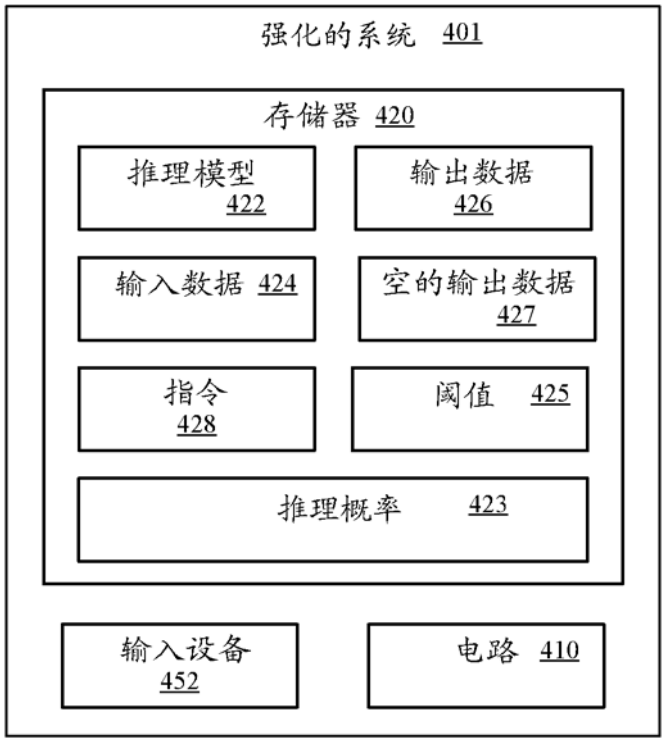

强化的系统 401
存储器 420
推理模型 422
输出数据 426
输入数据 424
空的输出数据 427
指令 428
阈值 425
推理概率 423
输入设备 452
电路 410

图 4

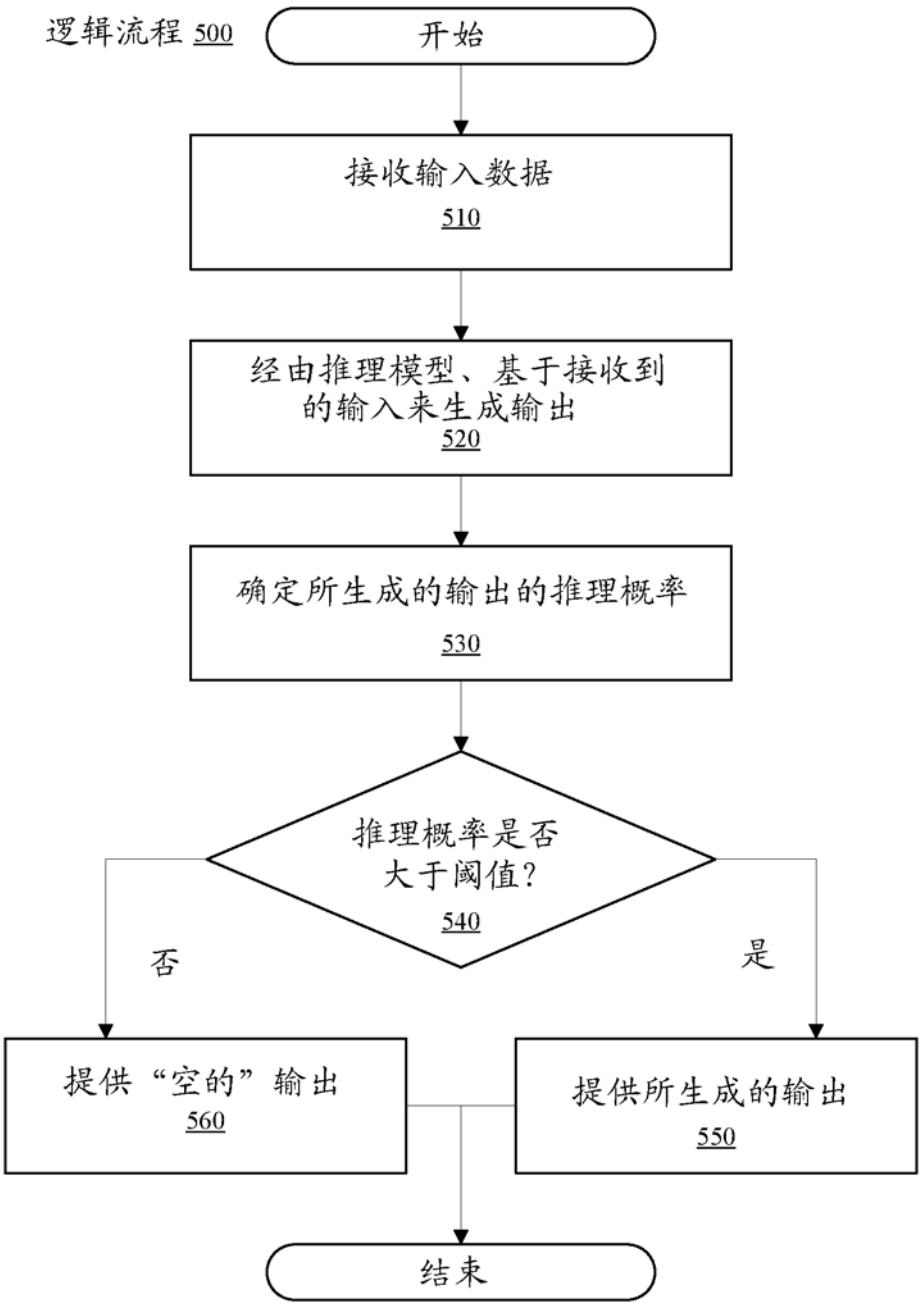

逻辑流程 500
开始
接收输入数据
510
经由推理模型、基于接收到的输入来生成输出
520
确定所生成的输出的推理概率
530
推理概率是否大于阈值?
540
否
是
提供“空的”输出
560
提供所生成的输出
550
结束

图 5

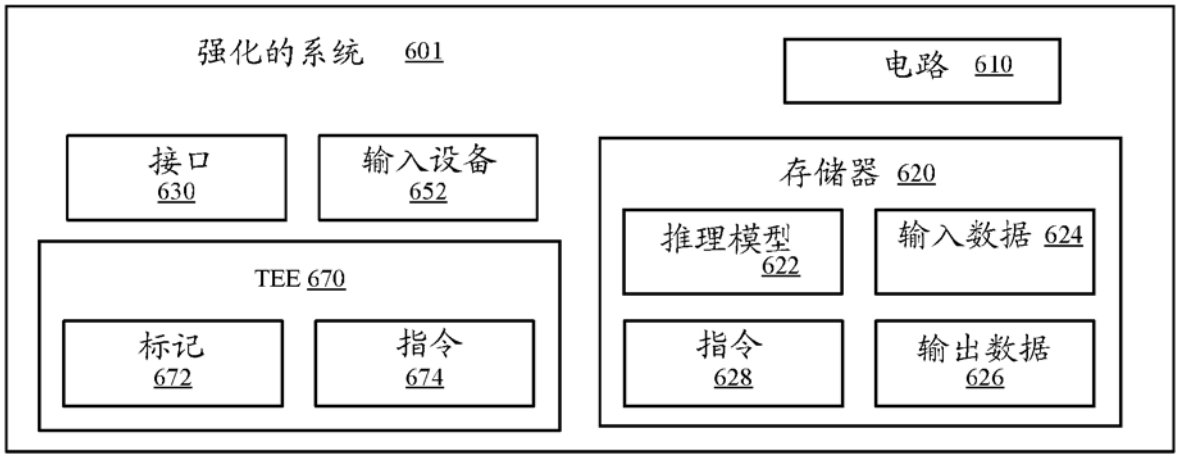

强化的系统 601
电路 610
接口 630
输入设备 652
存储器 620
推理模型 622
输入数据 624
TEE 670
标记 672
指令 674
指令 628
输出数据 626

图 6

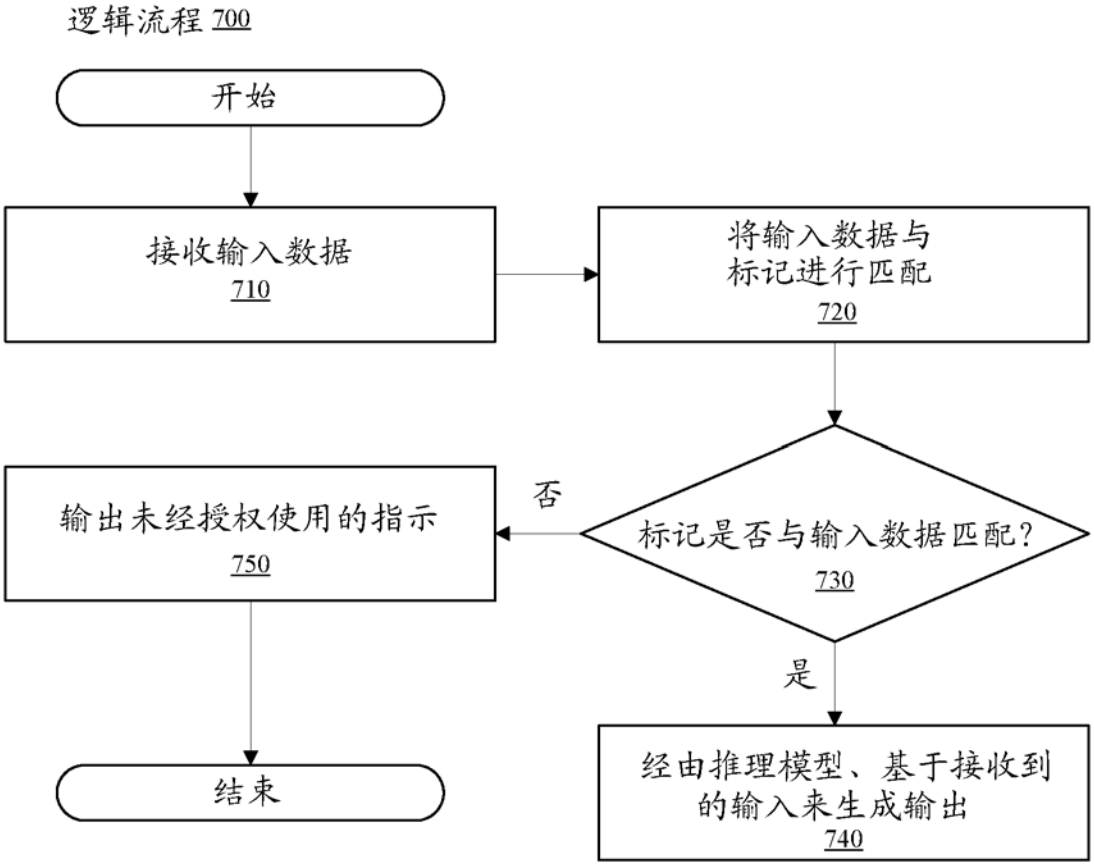

逻辑流程 700
开始
接收输入数据 710
将输入数据与标记进行匹配 720
标记是否与输入数据匹配? 730
否
是
输出未经授权使用的指示 750
经由推理模型、基于接收到的输入来生成输出 740
结束

图 7

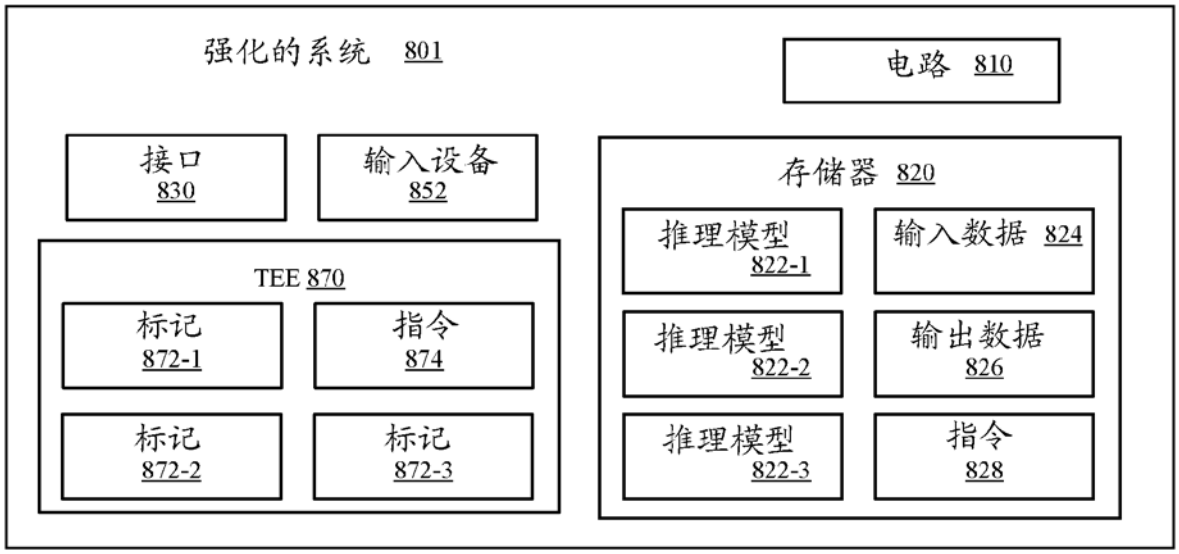

强化的系统 801
电路 810
接口 830
输入设备 852
存储器 820
TEE 870
标记 872-1
指令 874
标记 872-2
标记 872-3
推理模型 822-1
输入数据 824
推理模型 822-2
输出数据 826
推理模型 822-3
指令 828

图 8

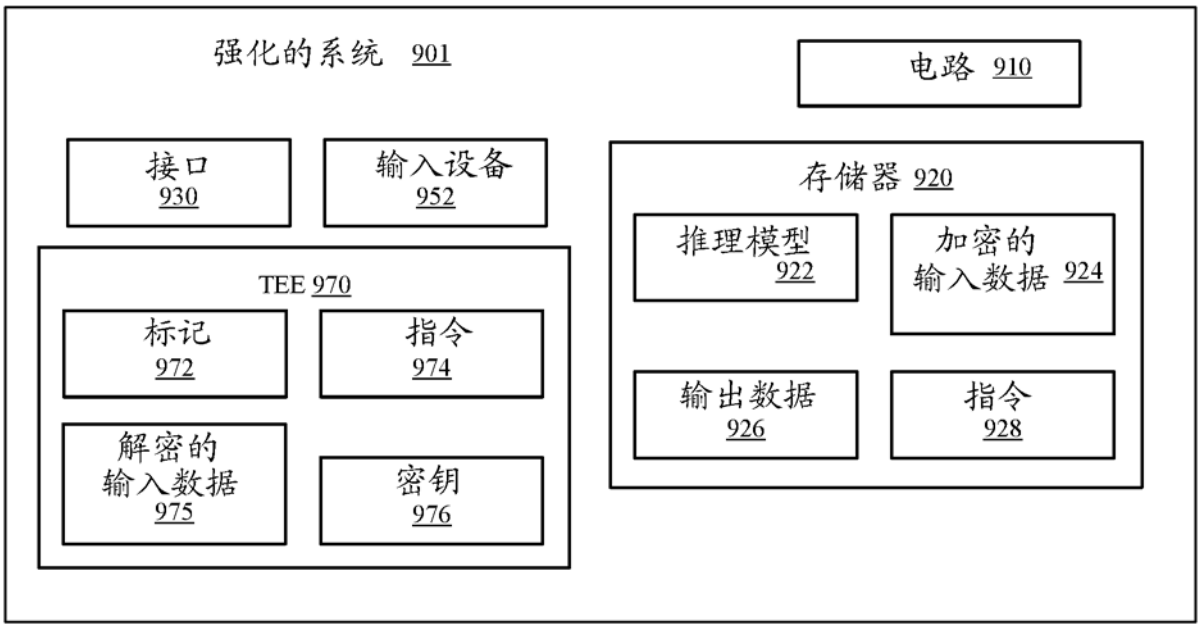

强化的系统 901
电路 910
接口 930
输入设备 952
存储器 920
TEE 970
标记 972
指令 974
解密的 输入数据 975
密钥 976
推理模型 922
加密的 输入数据 924
输出数据 926
指令 928

图 9

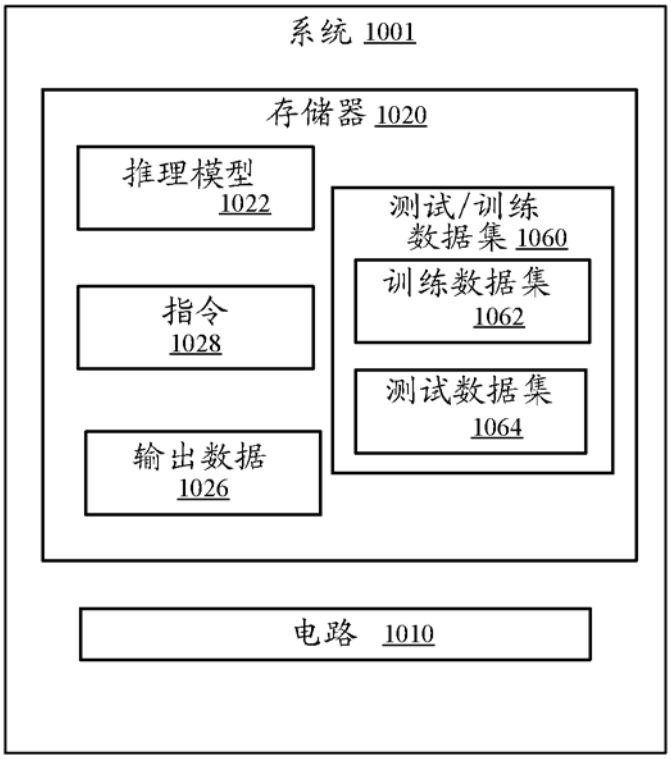

系统 1001
存储器 1020
推理模型 1022
测试/训练数据集 1060
训练数据集 1062
测试数据集 1064
指令 1028
输出数据 1026
电路 1010

图 10A

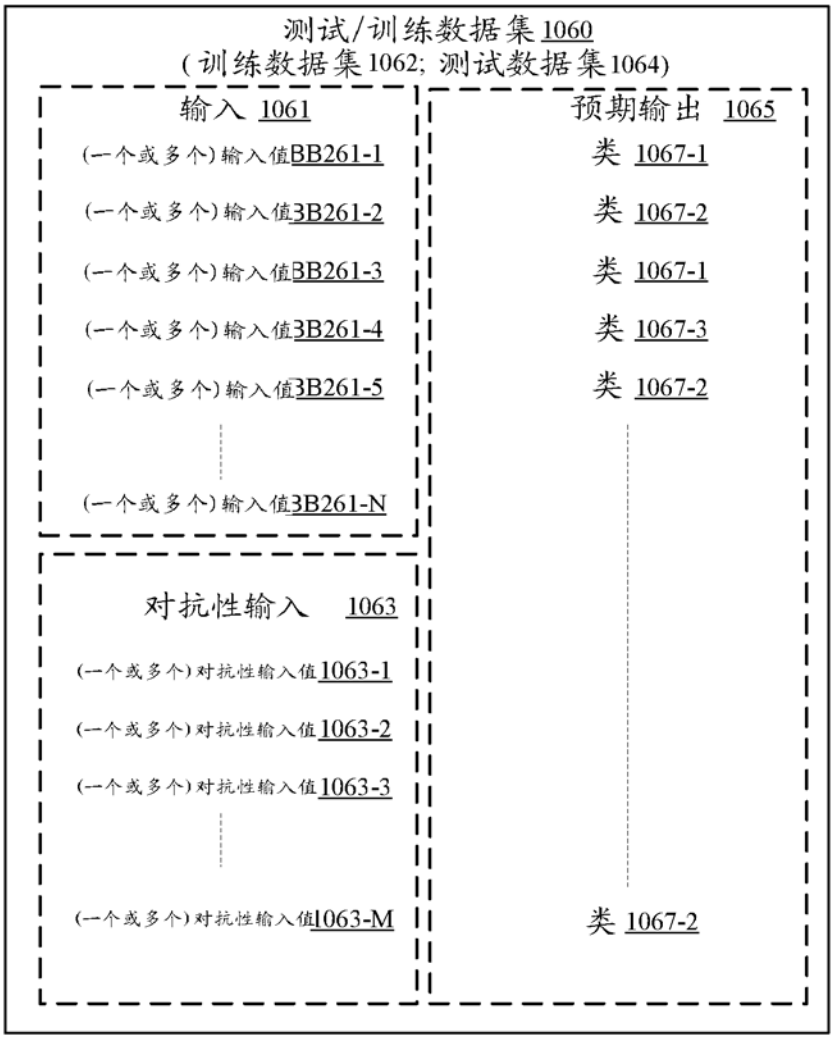

测试/训练数据集 1060
(训练数据集 1062; 测试数据集 1064)
输入 1061
(一个或多个)输入值BB261-1
(一个或多个)输入值3B261-2
(一个或多个)输入值3B261-3
(一个或多个)输入值3B261-4
(一个或多个)输入值3B261-5
(一个或多个)输入值3B261-N
预期输出 1065
类 1067-1
类 1067-2
类 1067-1
类 1067-3
类 1067-2
类 1067-2
对抗性输入 1063
(一个或多个)对抗性输入值1063-1
(一个或多个)对抗性输入值1063-2
(一个或多个)对抗性输入值1063-3
(一个或多个)对抗性输入值1063-M

图 10B

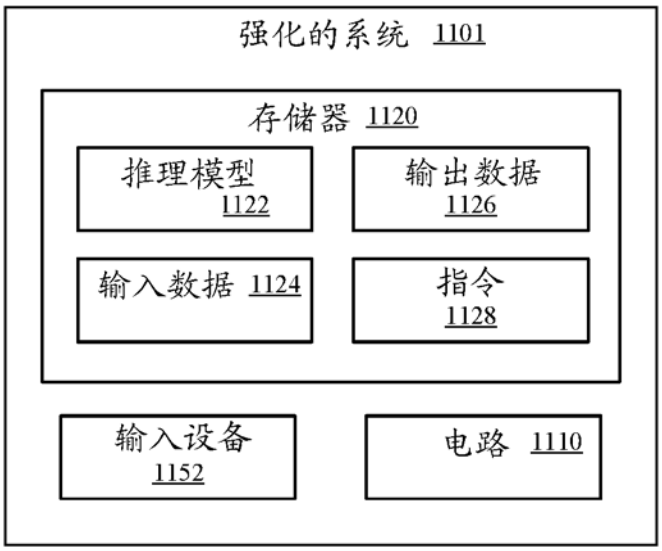

强化的系统 1101
存储器 1120
推理模型 1122
输出数据 1126
输入数据 1124
指令 1128
输入设备 1152
电路 1110

图 11

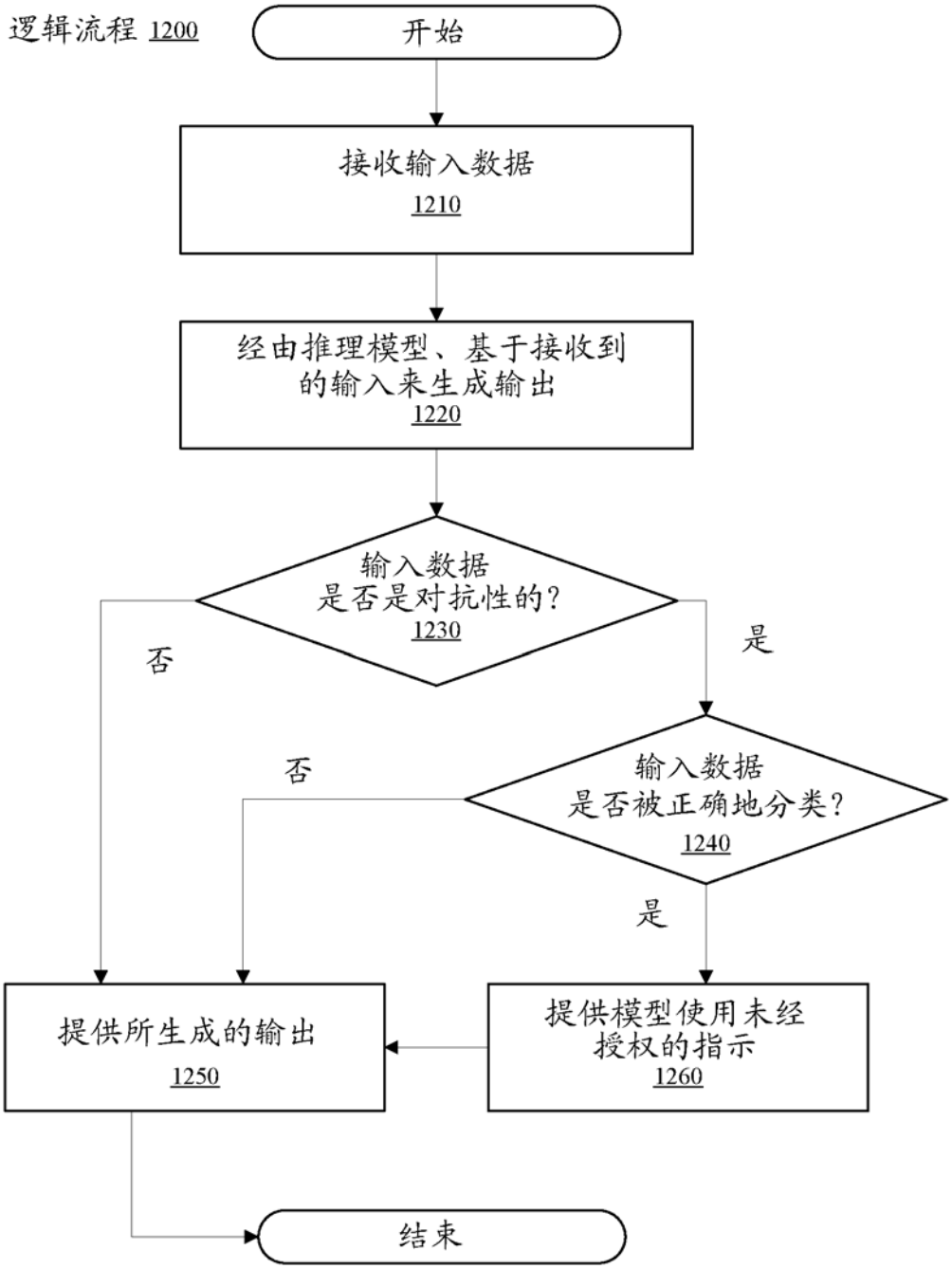

逻辑流程 1200
开始
接收输入数据
1210
经由推理模型、基于接收到的输入来生成输出
1220
输入数据是否是对抗性的?
1230
否
是
输入数据是否被正确地分类?
1240
否
是
提供所生成的输出
1250
提供模型使用未经授权的指示
1260
结束

图 12

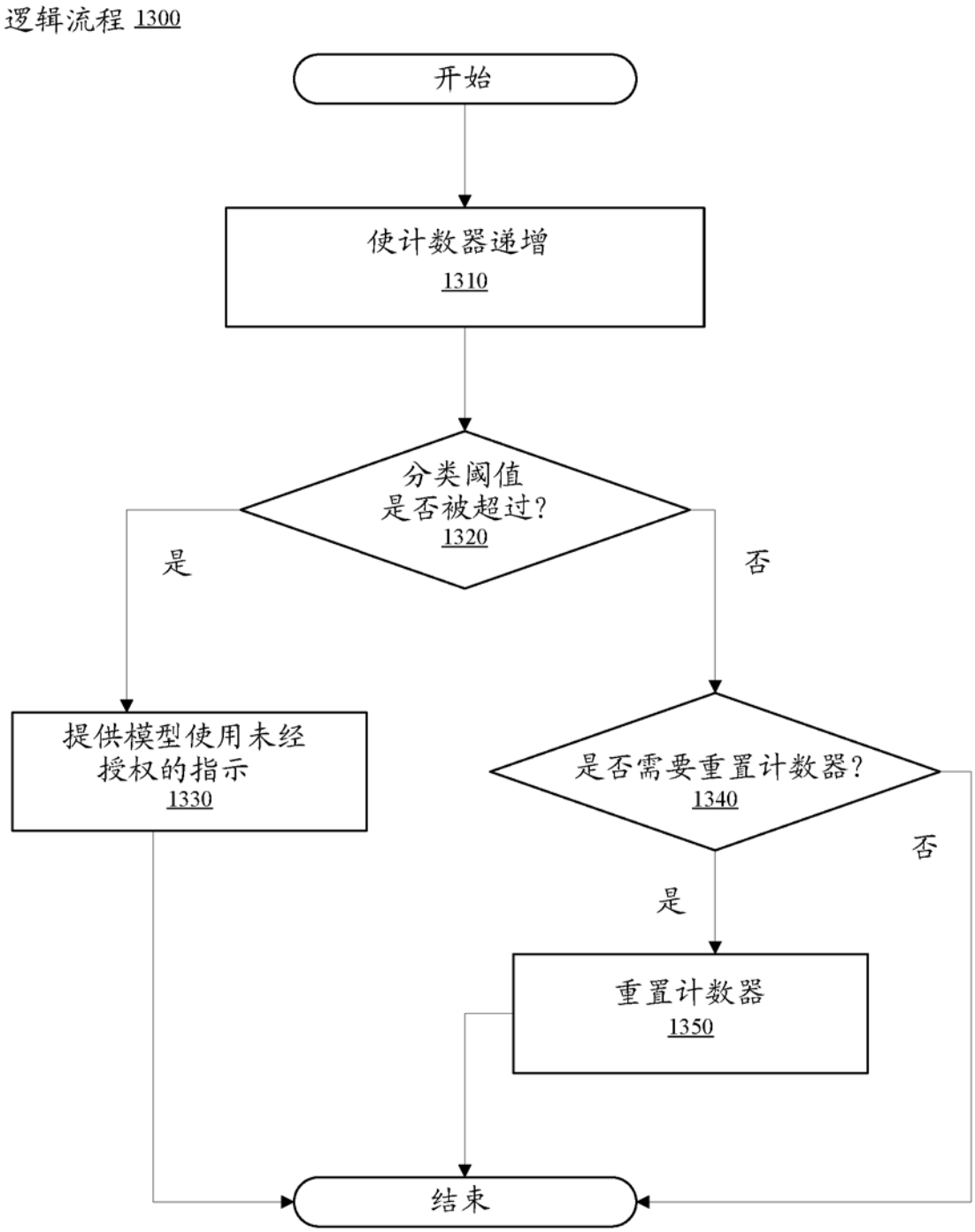

逻辑流程 1300
开始
使计数器递增
1310
分类阈值
是否被超过?
1320
是
否
提供模型使用未经
授权的指示
1330
是否需要重置计数器?
1340
否
是
重置计数器
1350
结束

图 13

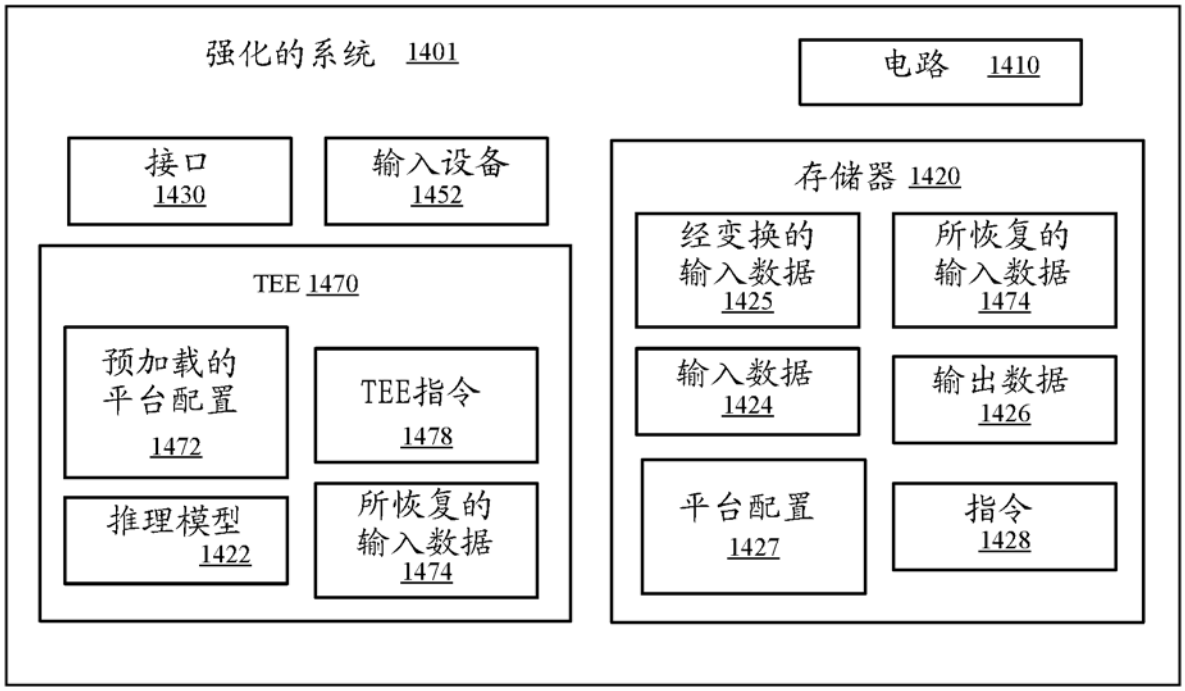

强化的系统 1401
电路 1410
接口 1430
输入设备 1452
存储器 1420
经变换的输入数据 1425
所恢复的输入数据 1474
TEE 1470
预加载的平台配置 1472
TEE指令 1478
输入数据 1424
输出数据 1426
推理模型 1422
所恢复的输入数据 1474
平台配置 1427
指令 1428

图 14

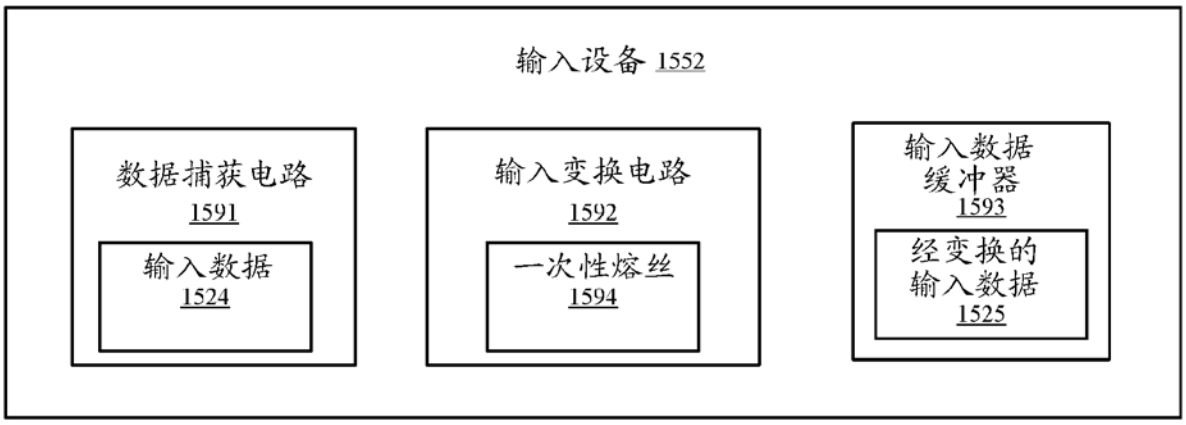

输入设备 1552
数据捕获电路 1591
输入数据 1524
输入变换电路 1592
一次性熔丝 1594
输入数据缓冲器 1593
经变换的输入数据 1525

图 15

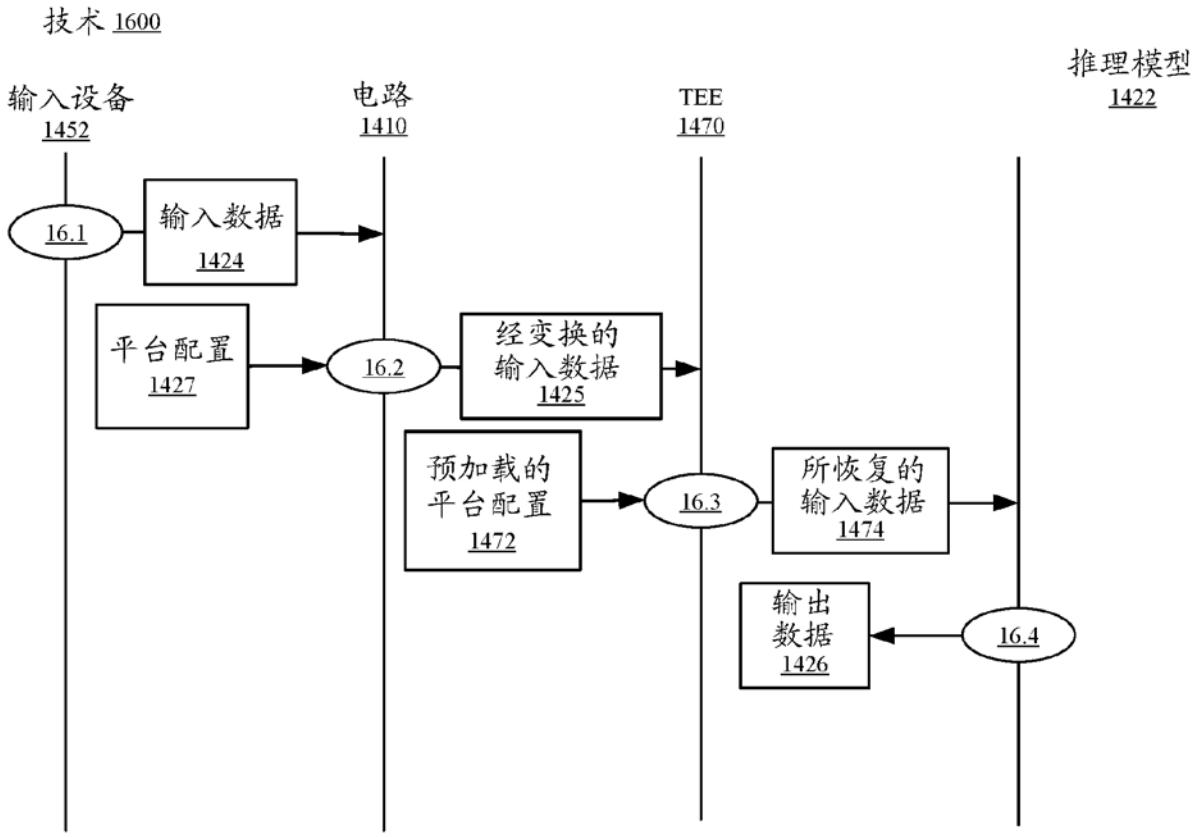

技术 1600
输入设备
1452
电路
1410
TEE
1470
推理模型
1422
16.1
输入数据
1424
平台配置
1427
16.2
经变换的
输入数据
1425
预加载的
平台配置
1472
16.3
所恢复的
输入数据
1474
输出
数据
1426
16.4

图 16

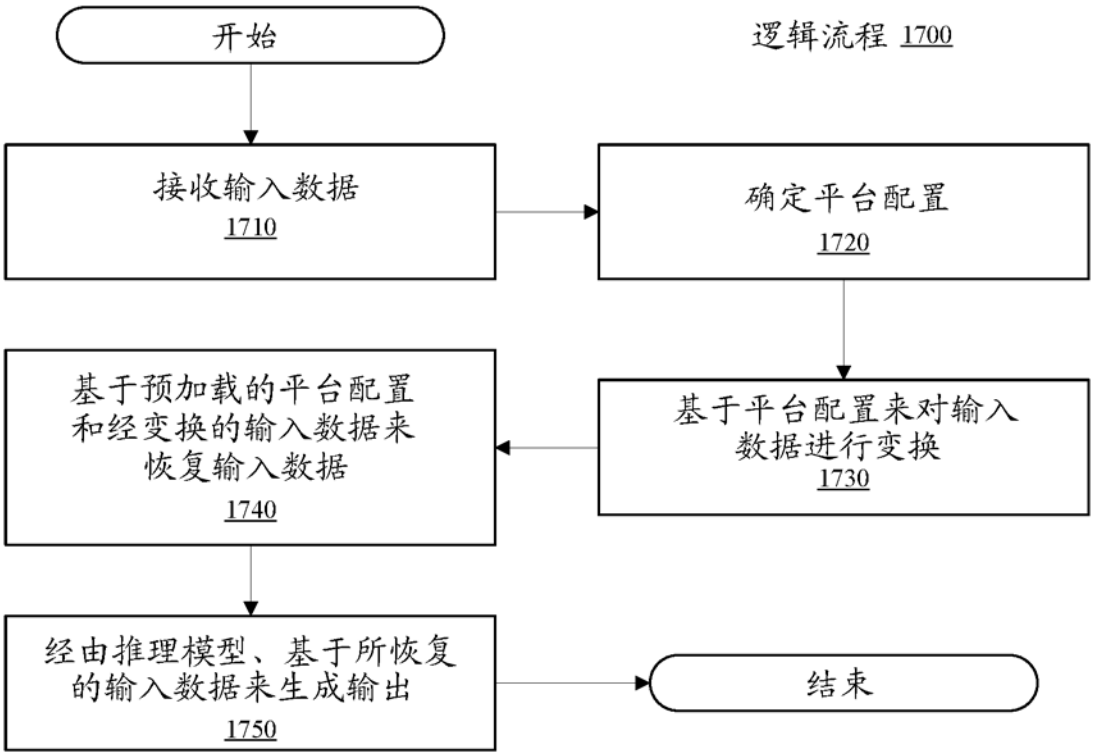

开始
逻辑流程 1700
接收输入数据
1710
确定平台配置
1720
基于预加载的平台配置
和经变换的输入数据来
恢复输入数据
1740
基于平台配置来对输入
数据进行变换
1730
经由推理模型、基于所恢复
的输入数据来生成输出
1750
结束

图 17

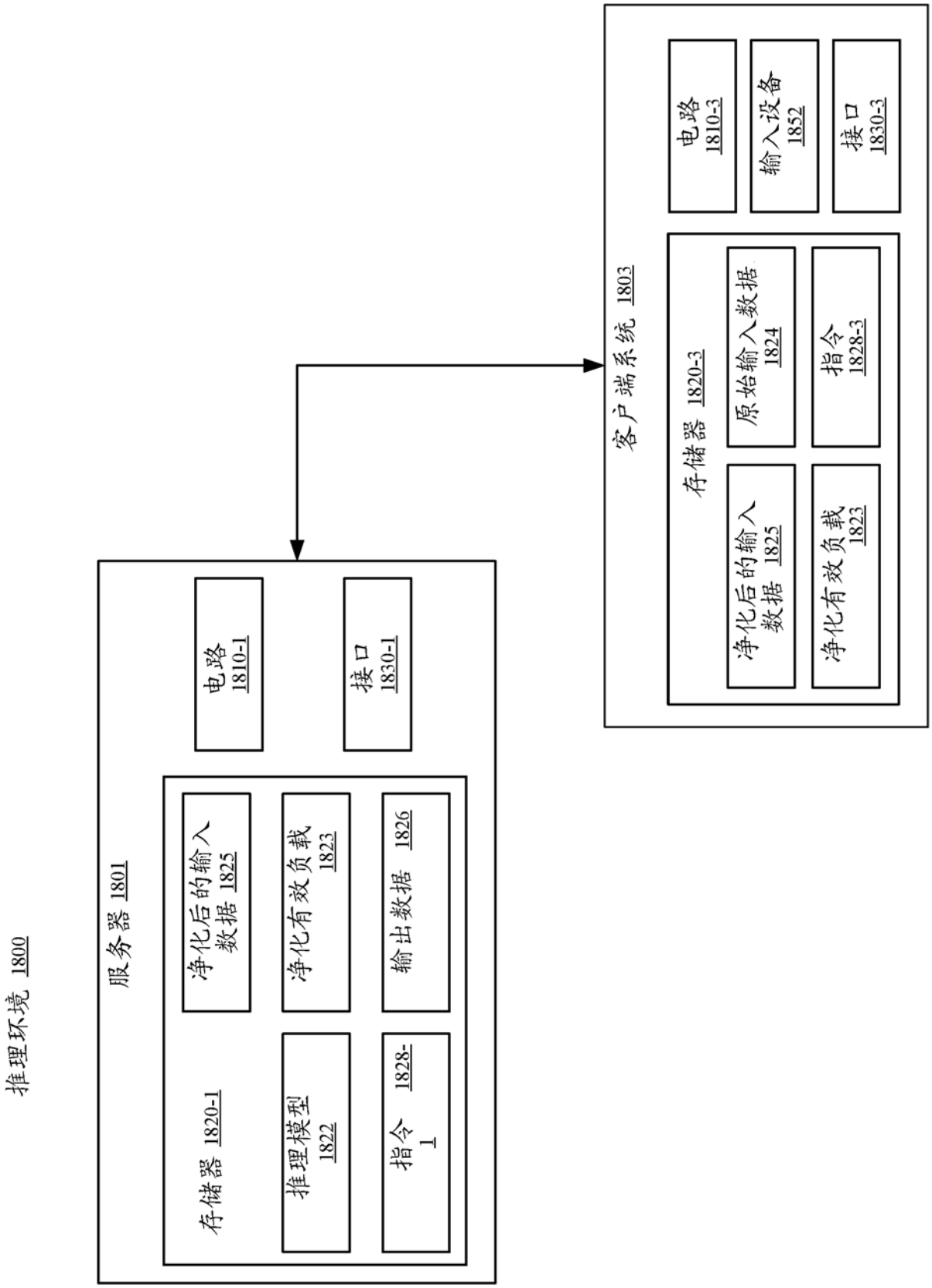

推理环境 1800
服务器 1801
存储器 1820-1
推理模型 1822
指令 1828-1
净化后的输入数据 1825
净化有效负载 1823
输出数据 1826
电路 1810-1
接口 1830-1
客户端系统 1803
存储器 1820-3
净化后的输入数据 1825
净化有效负载 1823
原始输入数据 1824
指令 1828-3
电路 1810-3
输入设备 1852
接口 1830-3

图 18

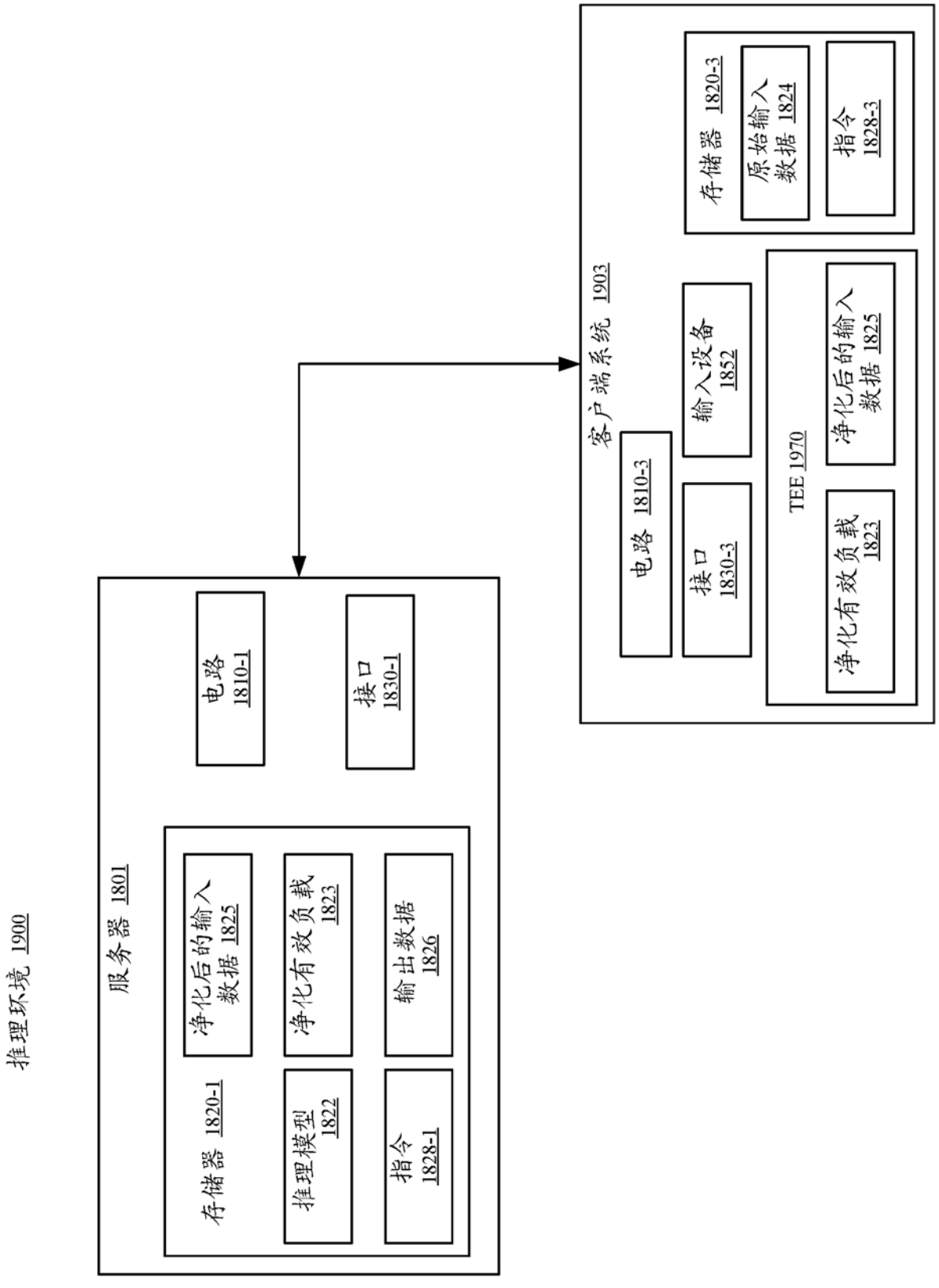

推理环境 1900
服务器 1801
存储器 1820-1
净化后的输入数据 1825
推理模型 1822
净化有效负载 1823
指令 1828-1
输出数据 1826
电路 1810-1
接口 1830-1
客户端系统 1903
电路 1810-3
接口 1830-3
输入设备 1852
TEE 1970
净化有效负载 1823
净化后的输入数据 1825
存储器 1820-3
原始输入数据 1824
指令 1828-3

图 19

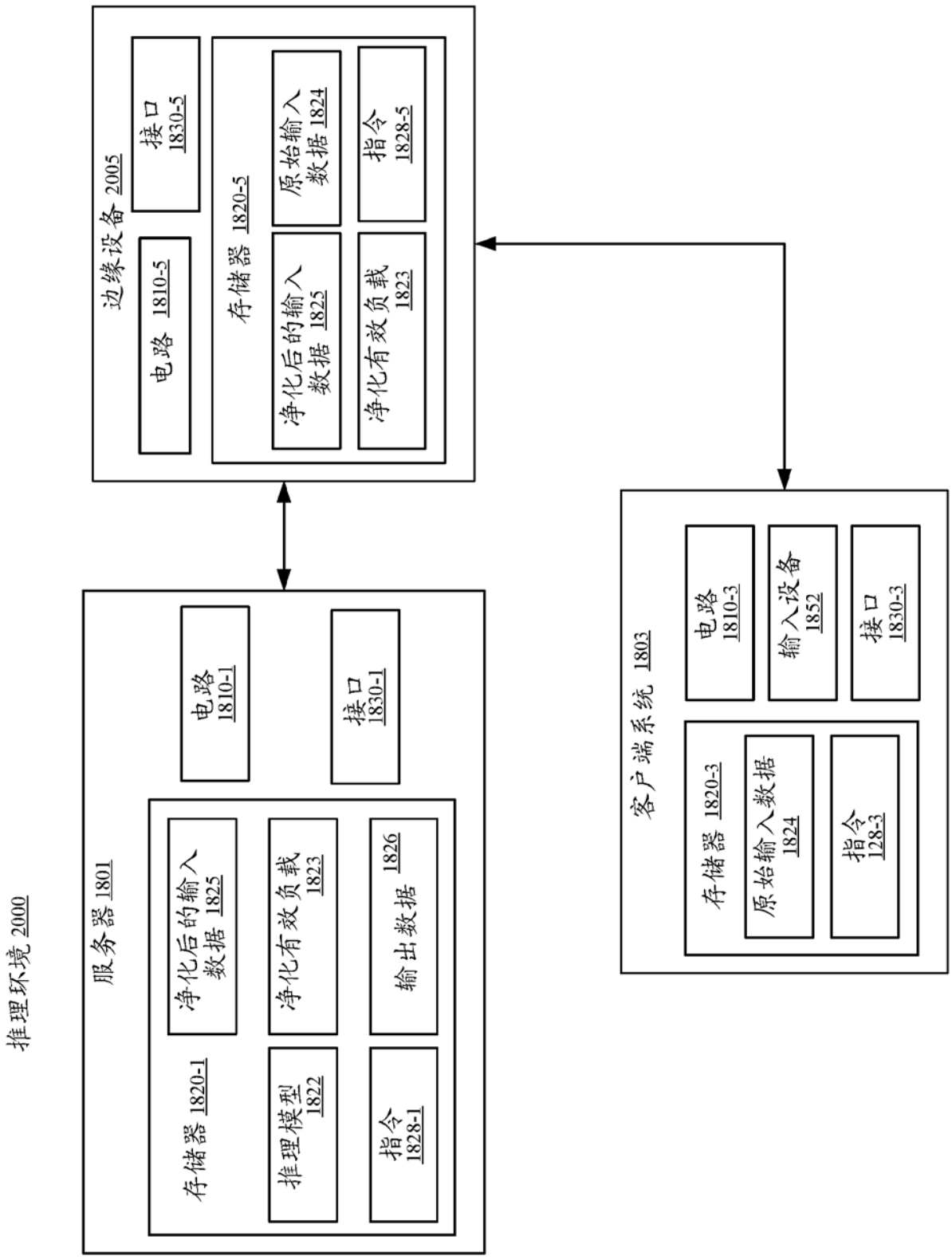

推理环境 2000
服务器 1801
存储器 1820-1
净化后的输入数据 1825
净化有效负载 1823
输出数据 1826
推理模型 1822
指令 1828-1
电路 1810-1
接口 1830-1
边缘设备 2005
电路 1810-5
存储器 1820-5
净化后的输入数据 1825
净化有效负载 1823
原始输入数据 1824
指令 1828-5
接口 1830-5
客户端系统 1803
存储器 1820-3
原始输入数据 1824
指令 128-3
电路 1810-3
输入设备 1852
接口 1830-3

图 20

技术 2100

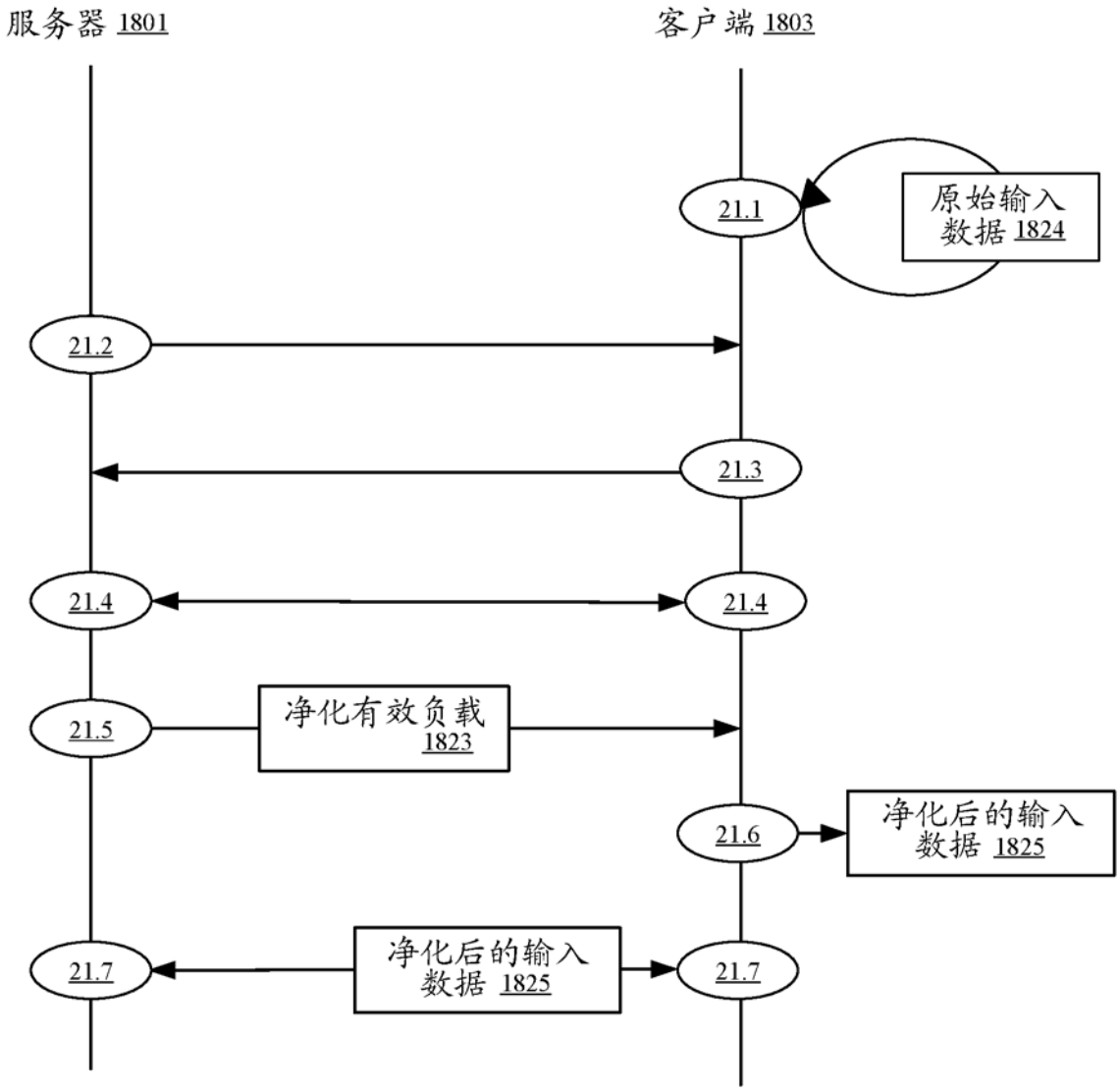

服务器 1801
客户端 1803
21.1
原始输入数据 1824
21.2
21.3
21.4
21.4
21.5
净化有效负载 1823
21.6
净化后的输入数据 1825
21.7
净化后的输入数据 1825
21.7

图 21

图形 2200

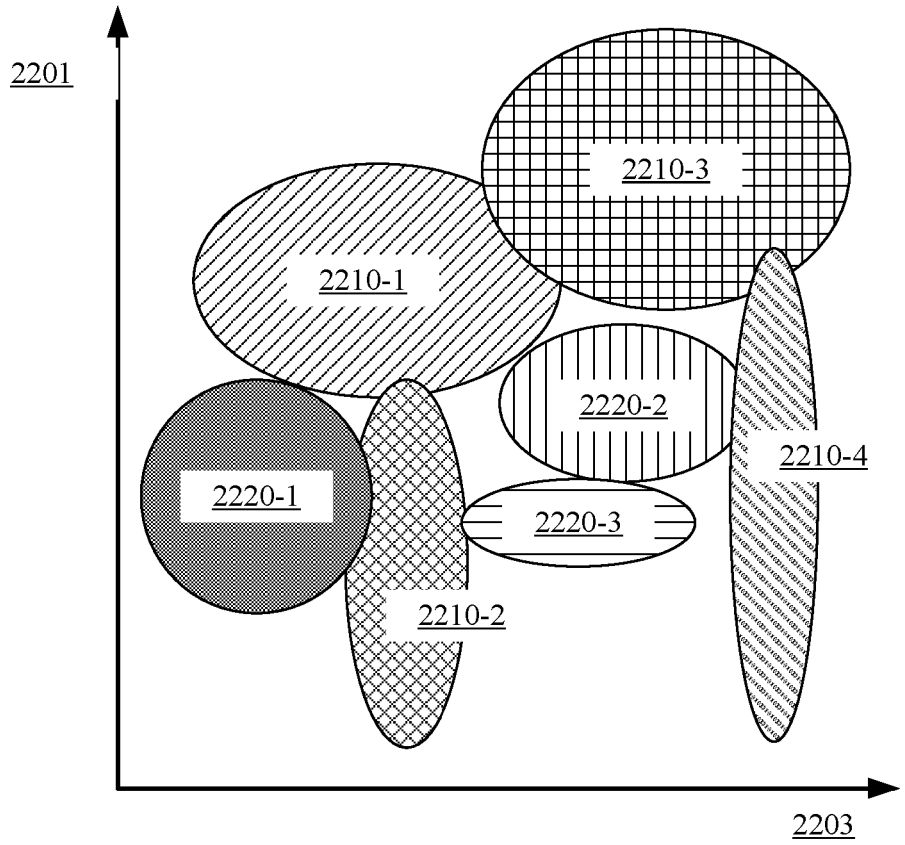



图 22

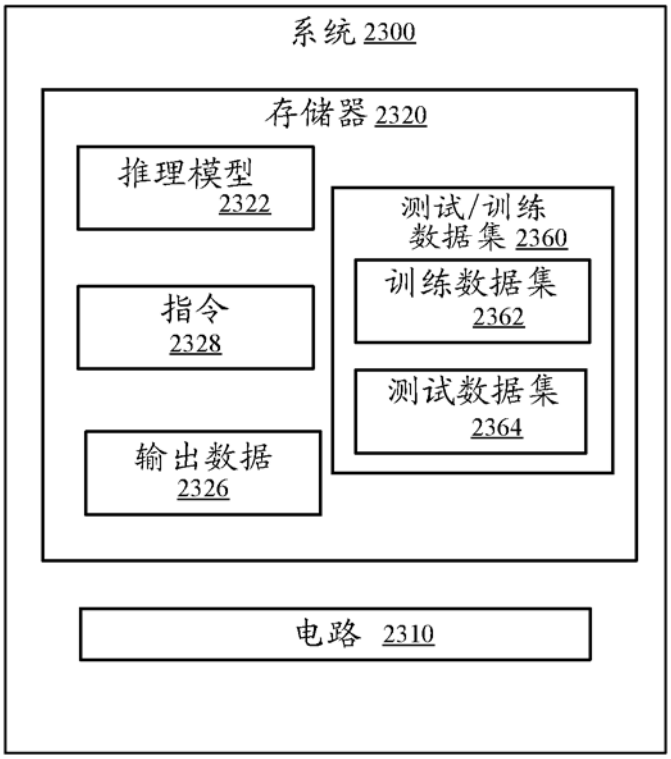

系统 2300
存储器 2320
推理模型 2322
测试/训练数据集 2360
训练数据集 2362
指令 2328
测试数据集 2364
输出数据 2326
电路 2310

图 23A

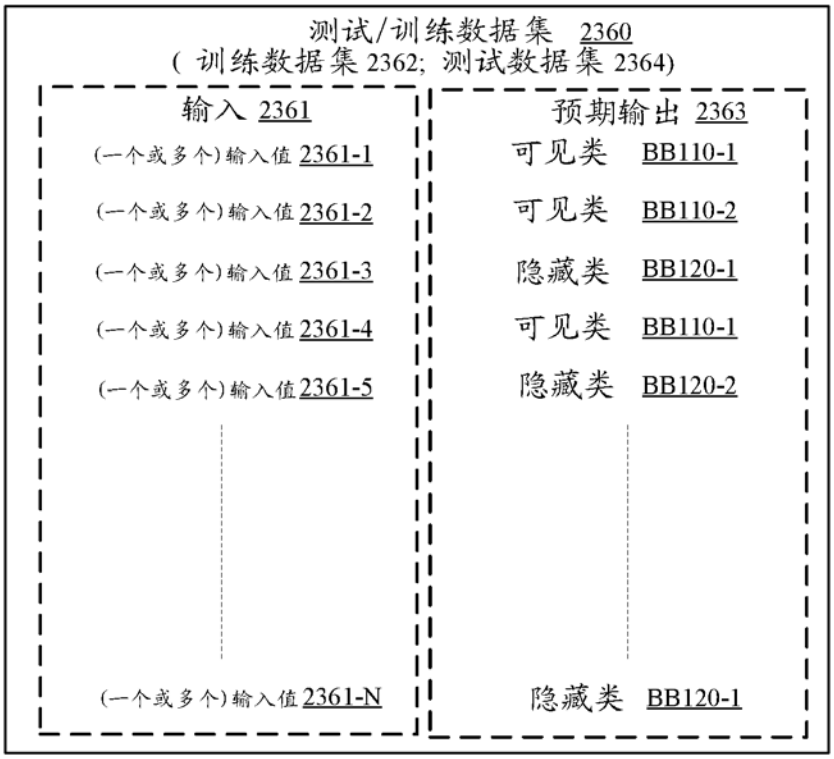

测试/训练数据集 2360
( 训练数据集 2362; 测试数据集 2364)
输入 2361
预期输出 2363
(一个或多个)输入值 2361-1
可见类 BB110-1
(一个或多个)输入值 2361-2
可见类 BB110-2
(一个或多个)输入值 2361-3
隐藏类 BB120-1
(一个或多个)输入值 2361-4
可见类 BB110-1
(一个或多个)输入值 2361-5
隐藏类 BB120-2
(一个或多个)输入值 2361-N
隐藏类 BB120-1

图 23B

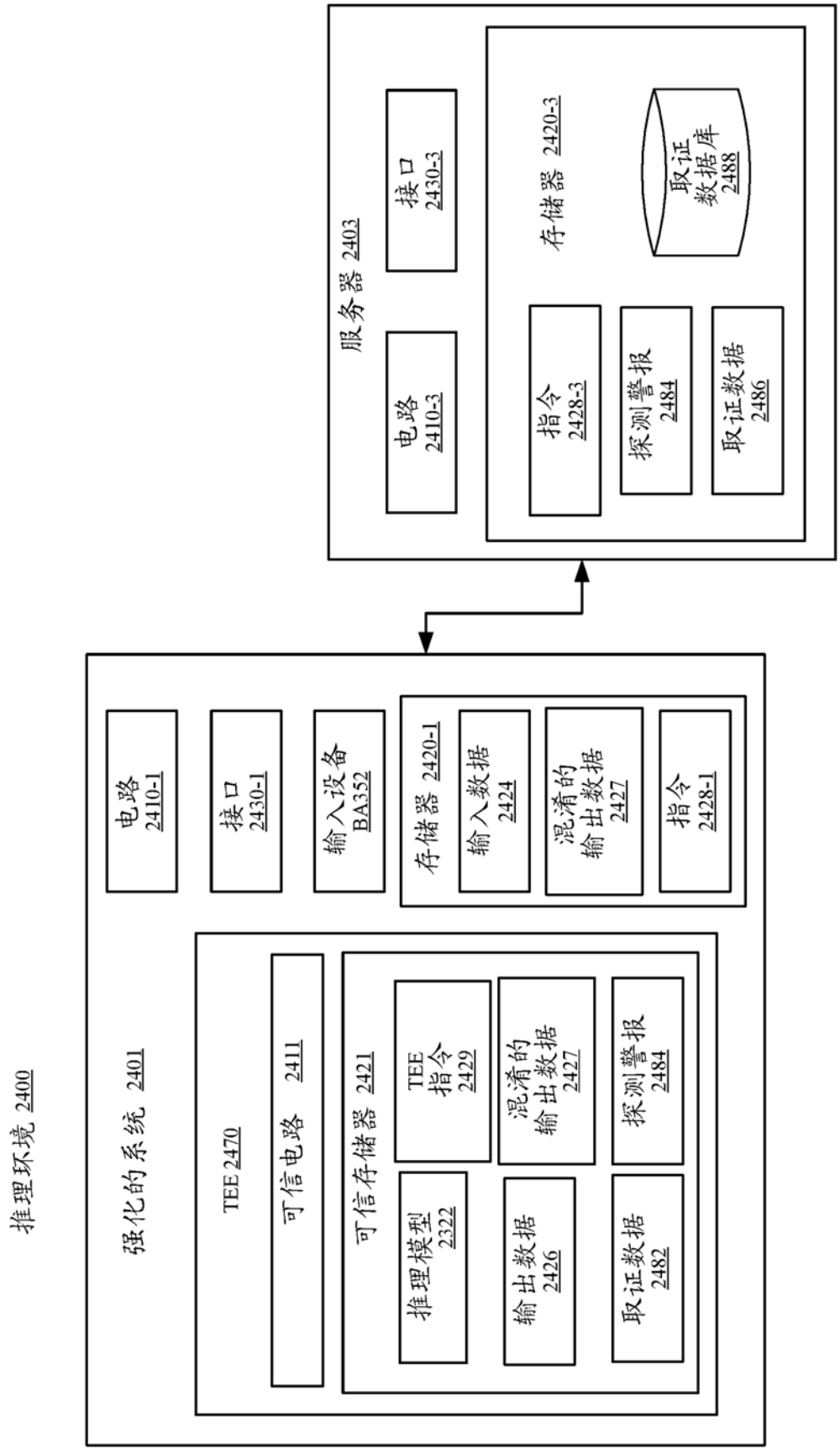

推理环境 2400
强化的系统 2401
TEE 2470
可信电路 2411
可信存储器 2421
推理模型 2322
TEE 指令 2429
输出数据 2426
混淆的输出数据 2427
取证数据 2482
探测警报 2484
电路 2410-1
接口 2430-1
输入设备 BA352
存储器 2420-1
输入数据 2424
混淆的输出数据 2427
指令 2428-1
服务器 2403
电路 2410-3
接口 2430-3
存储器 2420-3
指令 2428-3
探测警报 2484
取证数据 2486
取证数据库 2488

图 24

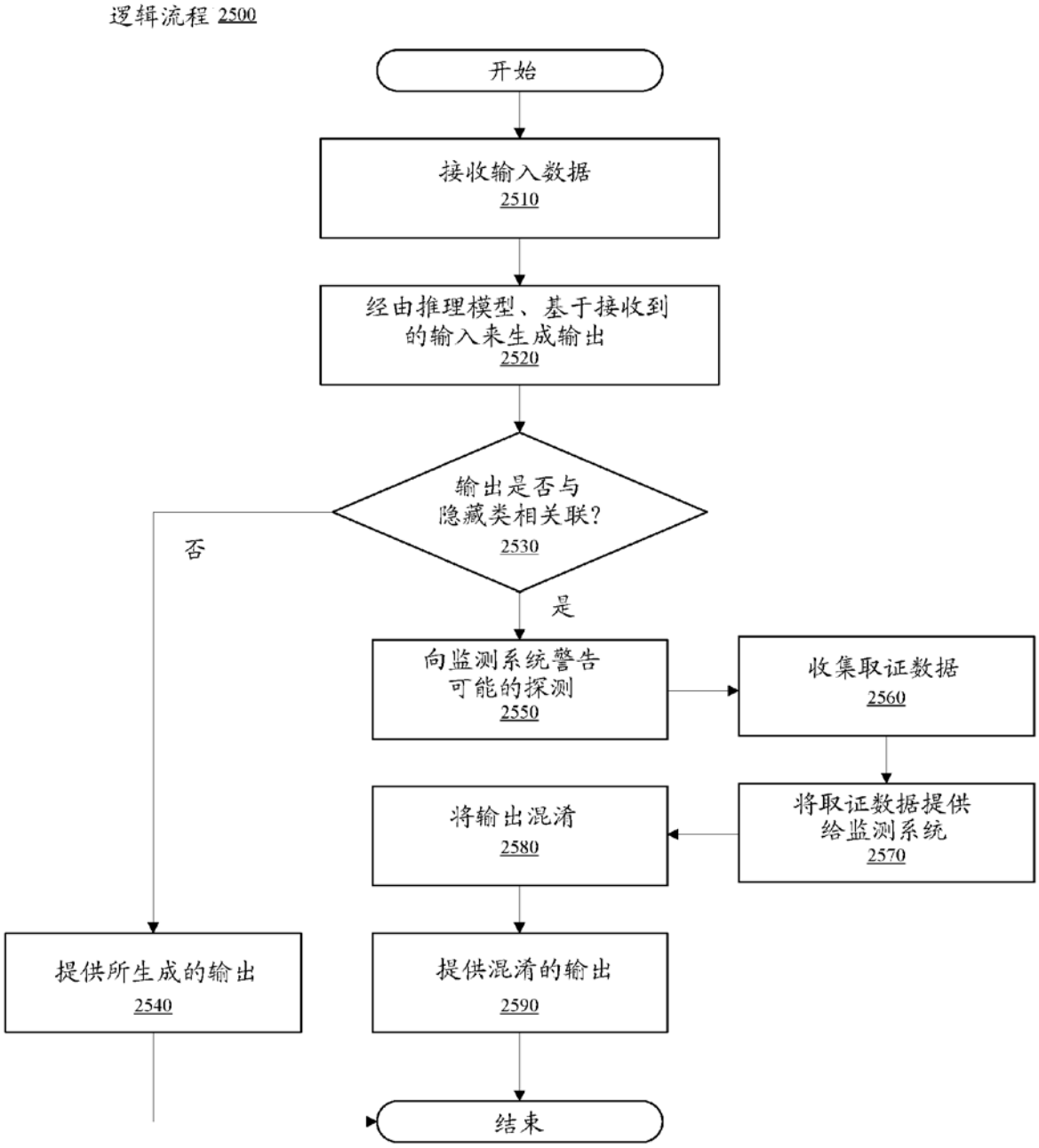

逻辑流程 2500
开始
接收输入数据
2510
经由推理模型、基于接收到的输入来生成输出
2520
输出是否与隐藏类相关联?
2530
否
是
向监测系统警告可能的探测
2550
收集取证数据
2560
将取证数据提供给监测系统
2570
将输出混淆
2580
提供所生成的输出
2540
提供混淆的输出
2590
结束

图 25

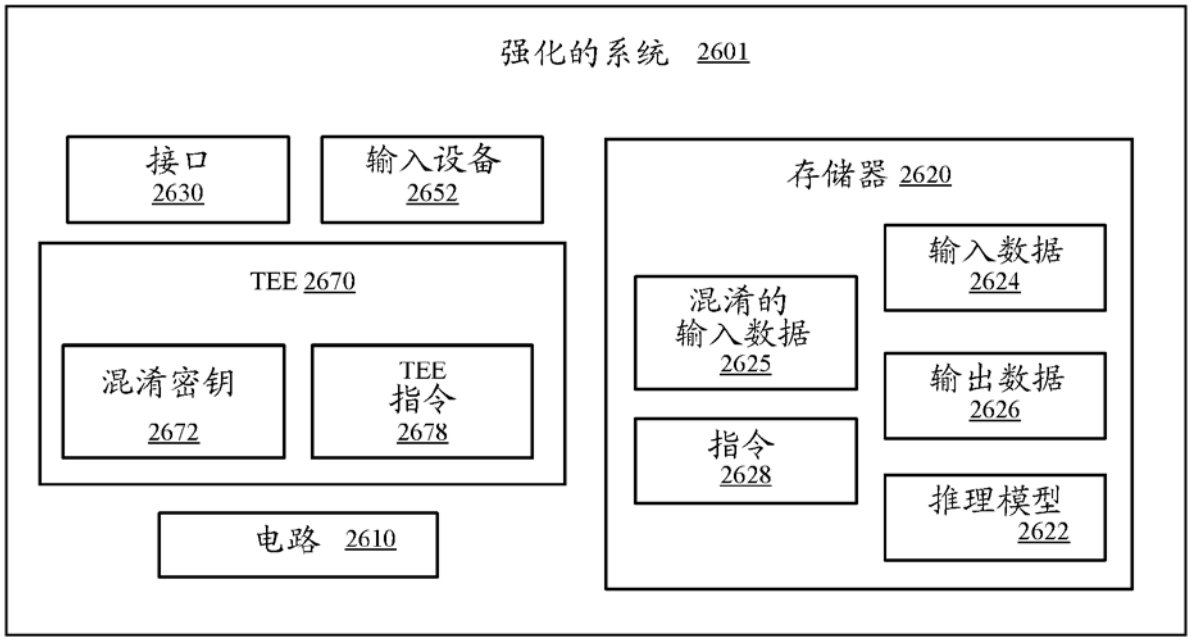

强化的系统 2601
接口 2630
输入设备 2652
TEE 2670
混淆密钥 2672
TEE 指令 2678
电路 2610
存储器 2620
输入数据 2624
混淆的 输入数据 2625
输出数据 2626
指令 2628
推理模型 2622

图 26

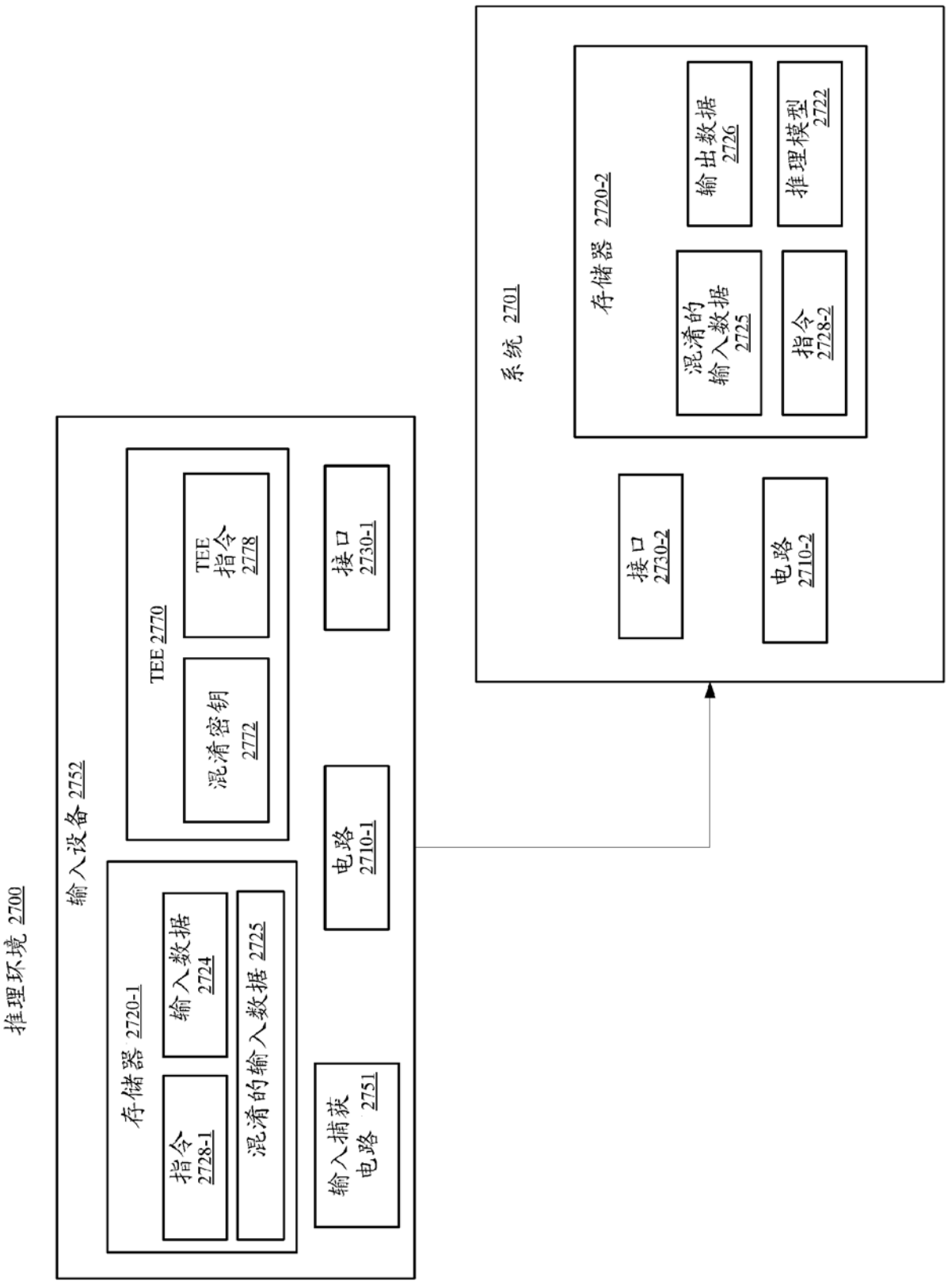

推理环境 2700
输入设备 2752
存储器 2720-1
指令 2728-1
输入数据 2724
混淆的输入数据 2725
输入捕获电路 2751
电路 2710-1
TEE 2770
混淆密钥 2772
TEE 指令 2778
接口 2730-1
系统 2701
存储器 2720-2
混淆的输入数据 2725
指令 2728-2
输出数据 2726
推理模型 2722
接口 2730-2
电路 2710-2

图 27

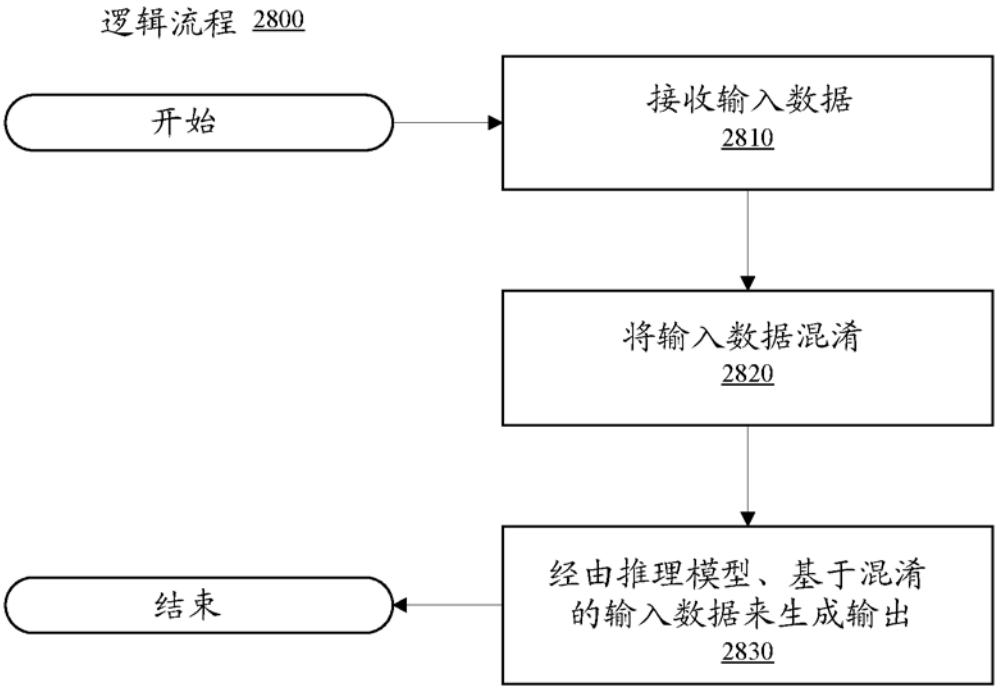

逻辑流程 2800
开始
接收输入数据
2810
将输入数据混淆
2820
经由推理模型、基于混淆的输入数据来生成输出
2830
结束

图 28

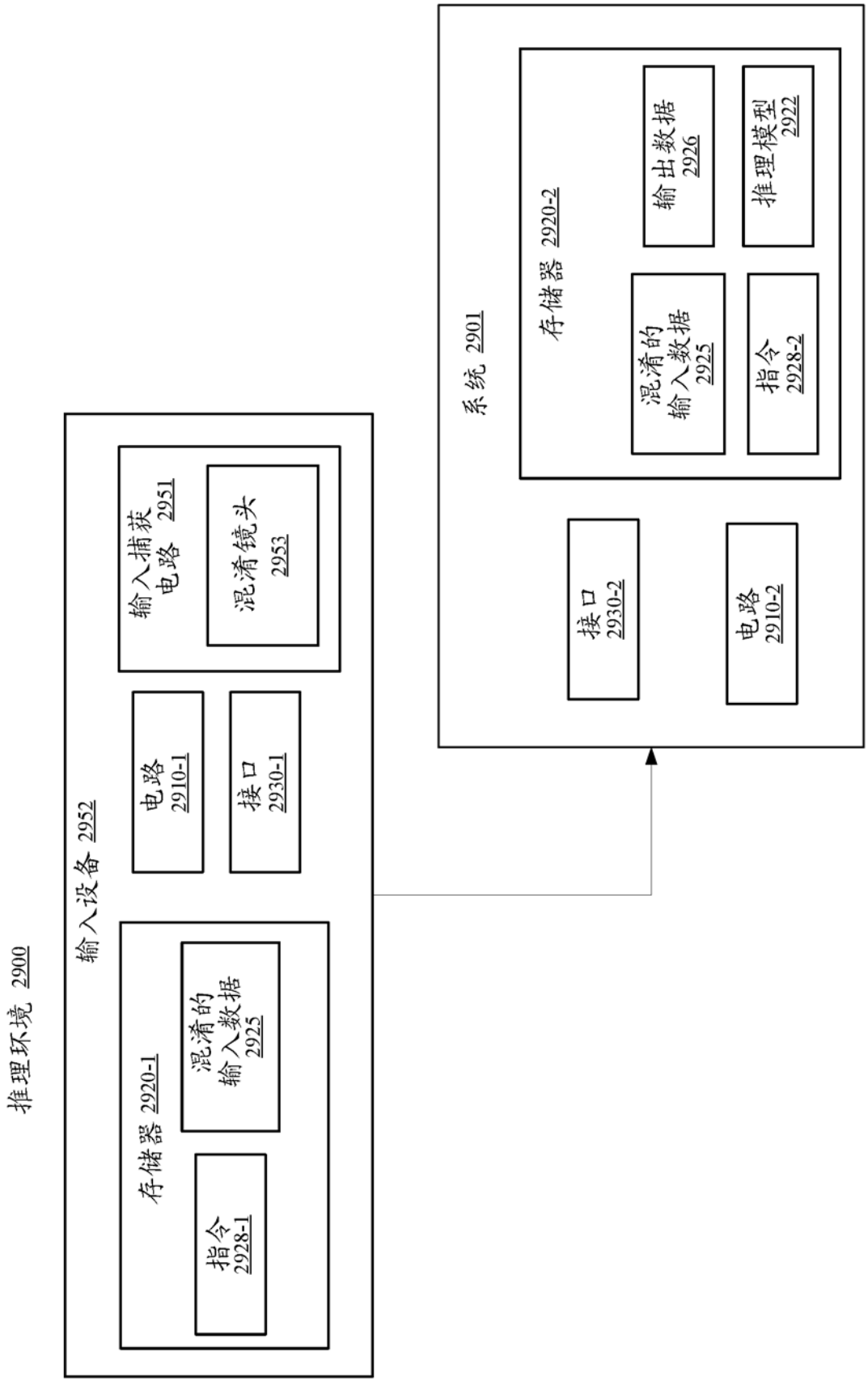

推理环境 2900
输入设备 2952
存储器 2920-1
指令 2928-1
混淆的输入数据 2925
电路 2910-1
接口 2930-1
输入捕获电路 2951
混淆镜头 2953
系统 2901
存储器 2920-2
输出数据 2926
推理模型 2922
混淆的输入数据 2925
指令 2928-2
接口 2930-2
电路 2910-2

图 29

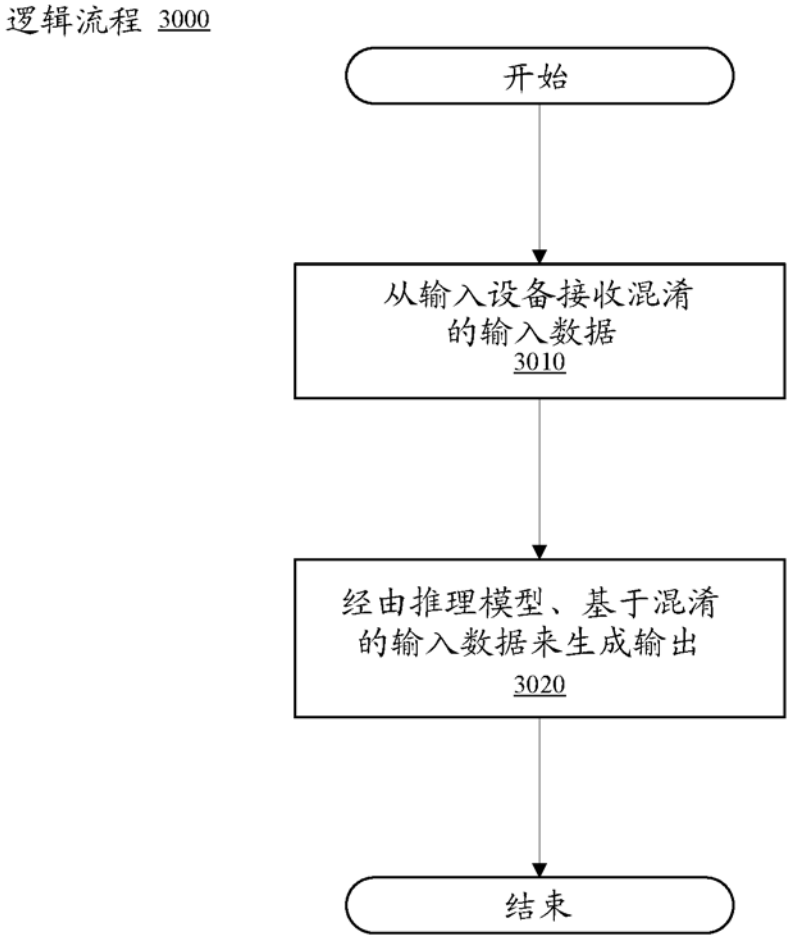

逻辑流程 3000
开始
从输入设备接收混淆的输入数据
3010
经由推理模型、基于混淆的输入数据来生成输出
3020
结束

图 30

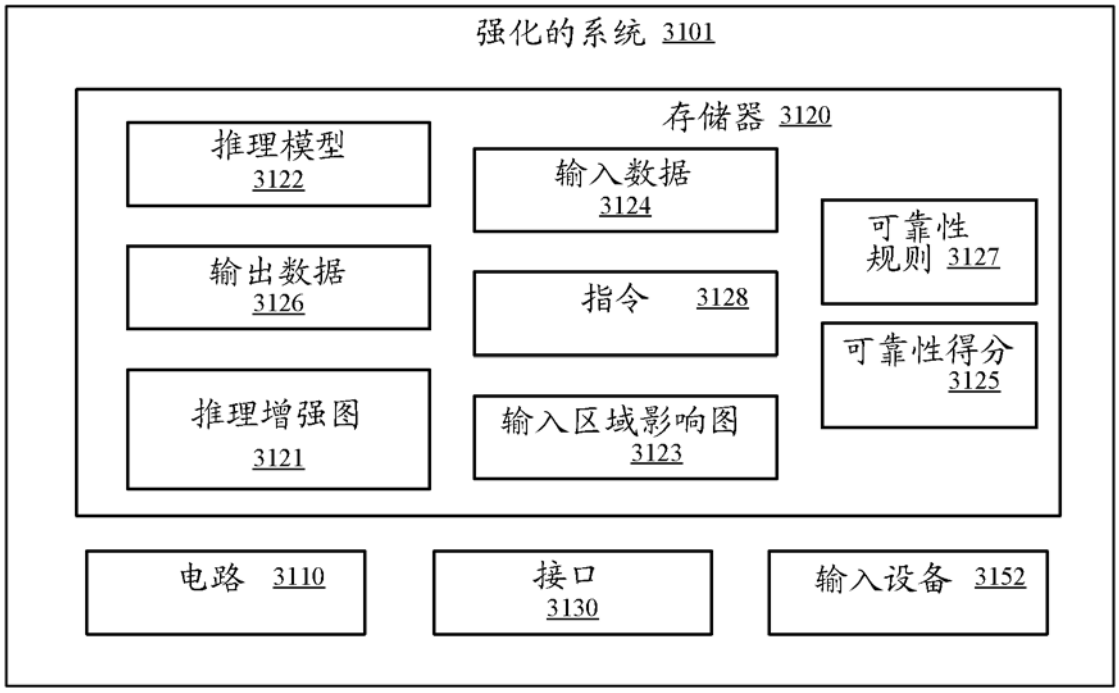

强化的系统 3101
存储器 3120
推理模型 3122
输入数据 3124
可靠性规则 3127
输出数据 3126
指令 3128
可靠性得分 3125
推理增强图 3121
输入区域影响图 3123
电路 3110
接口 3130
输入设备 3152

图 31

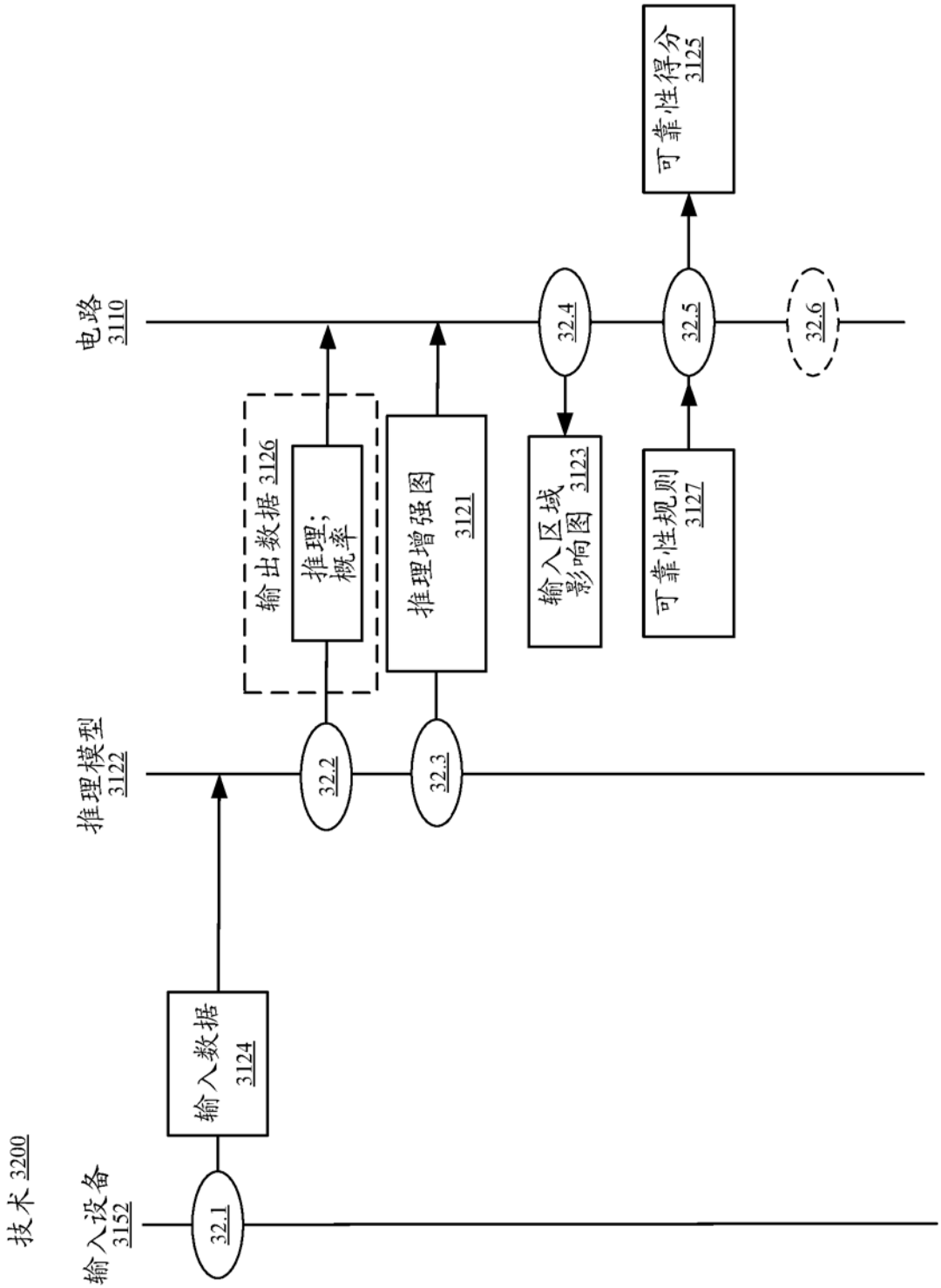

技术 3200
输入设备 3152
32.1
输入数据 3124
推理模型 3122
32.2
32.3
输出数据 3126
推理; 概率
推理增强图 3121
电路 3110
32.4
输入区域影响图 3123
32.5
可靠性规则 3127
32.6
可靠性得分 3125

图 32

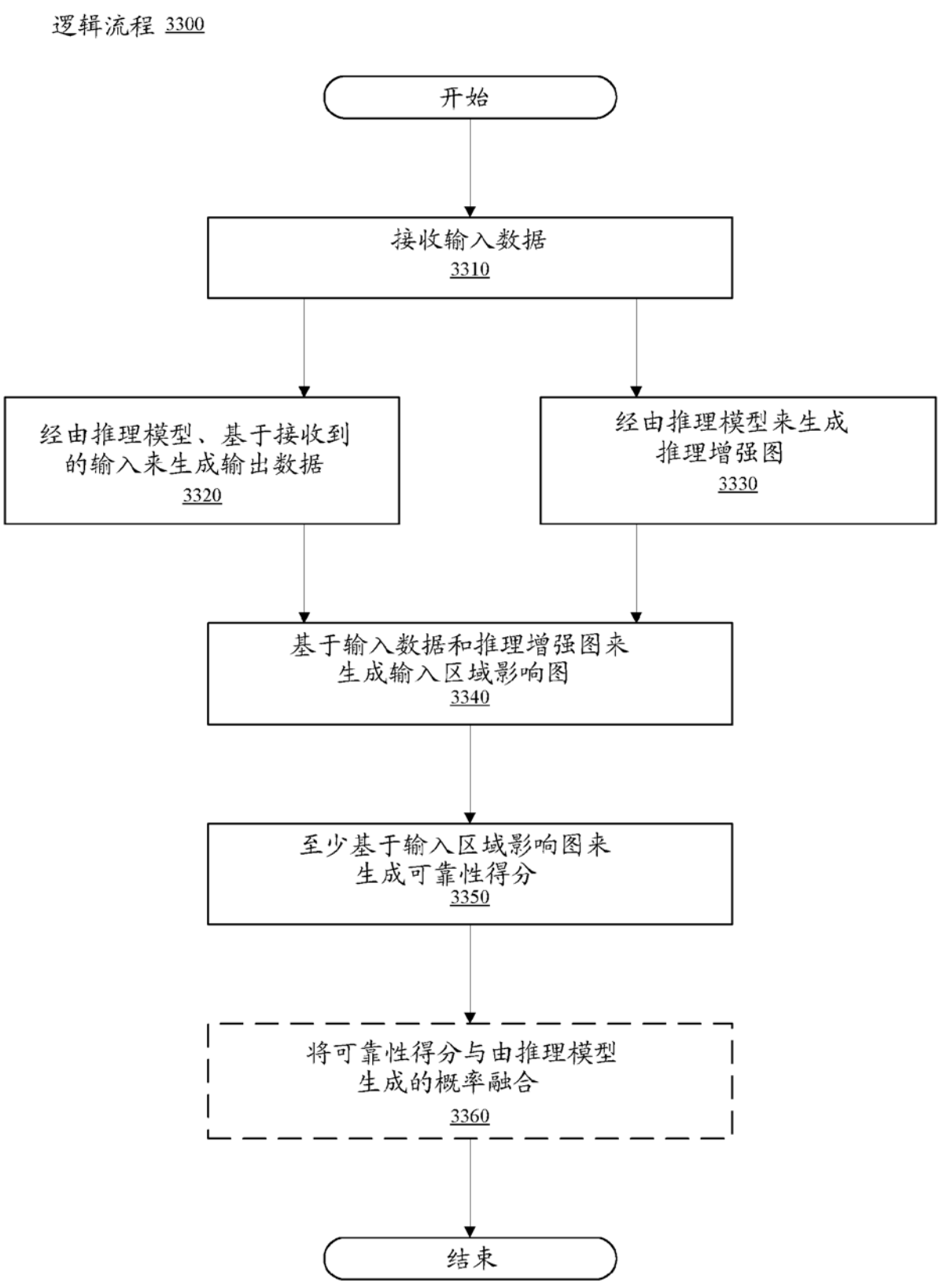

逻辑流程 3300
开始
接收输入数据
3310
经由推理模型、基于接收到的输入来生成输出数据
3320
经由推理模型来生成推理增强图
3330
基于输入数据和推理增强图来生成输入区域影响图
3340
至少基于输入区域影响图来生成可靠性得分
3350
将可靠性得分与由推理模型生成的概率融合
3360
结束

图 33

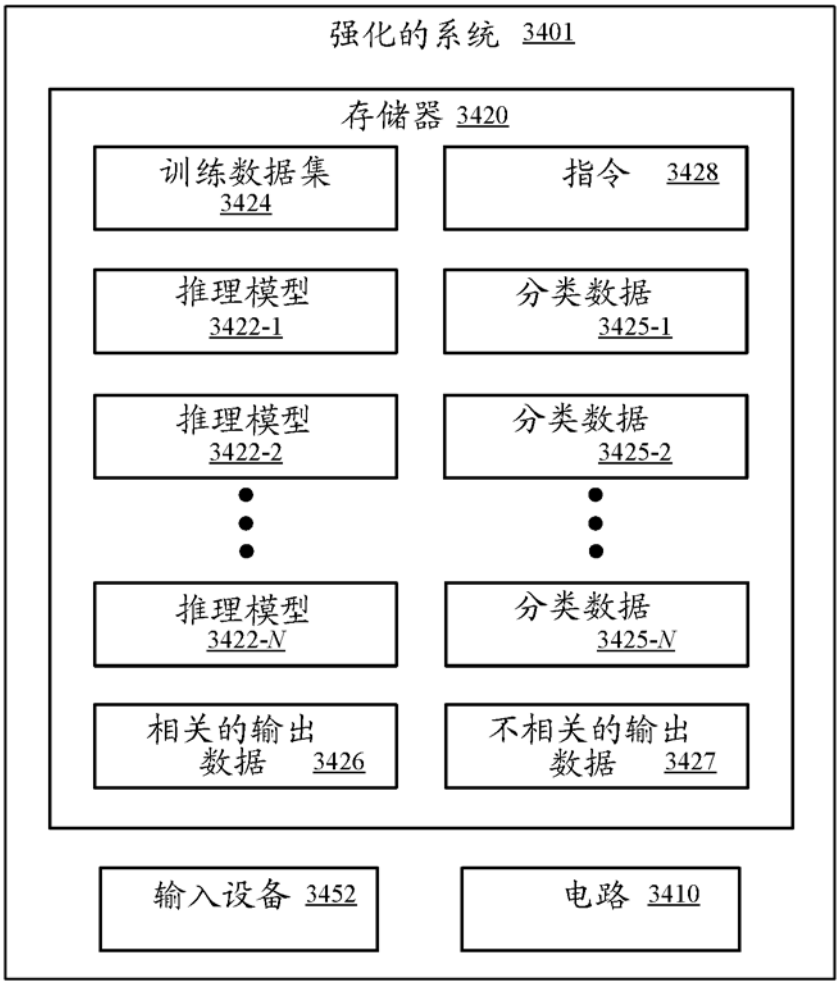

强化的系统 3401
存储器 3420
训练数据集 3424
指令 3428
推理模型 3422-1
分类数据 3425-1
推理模型 3422-2
分类数据 3425-2
推理模型 3422-N
分类数据 3425-N
相关的输出数据 3426
不相关的输出数据 3427
输入设备 3452
电路 3410

图 34

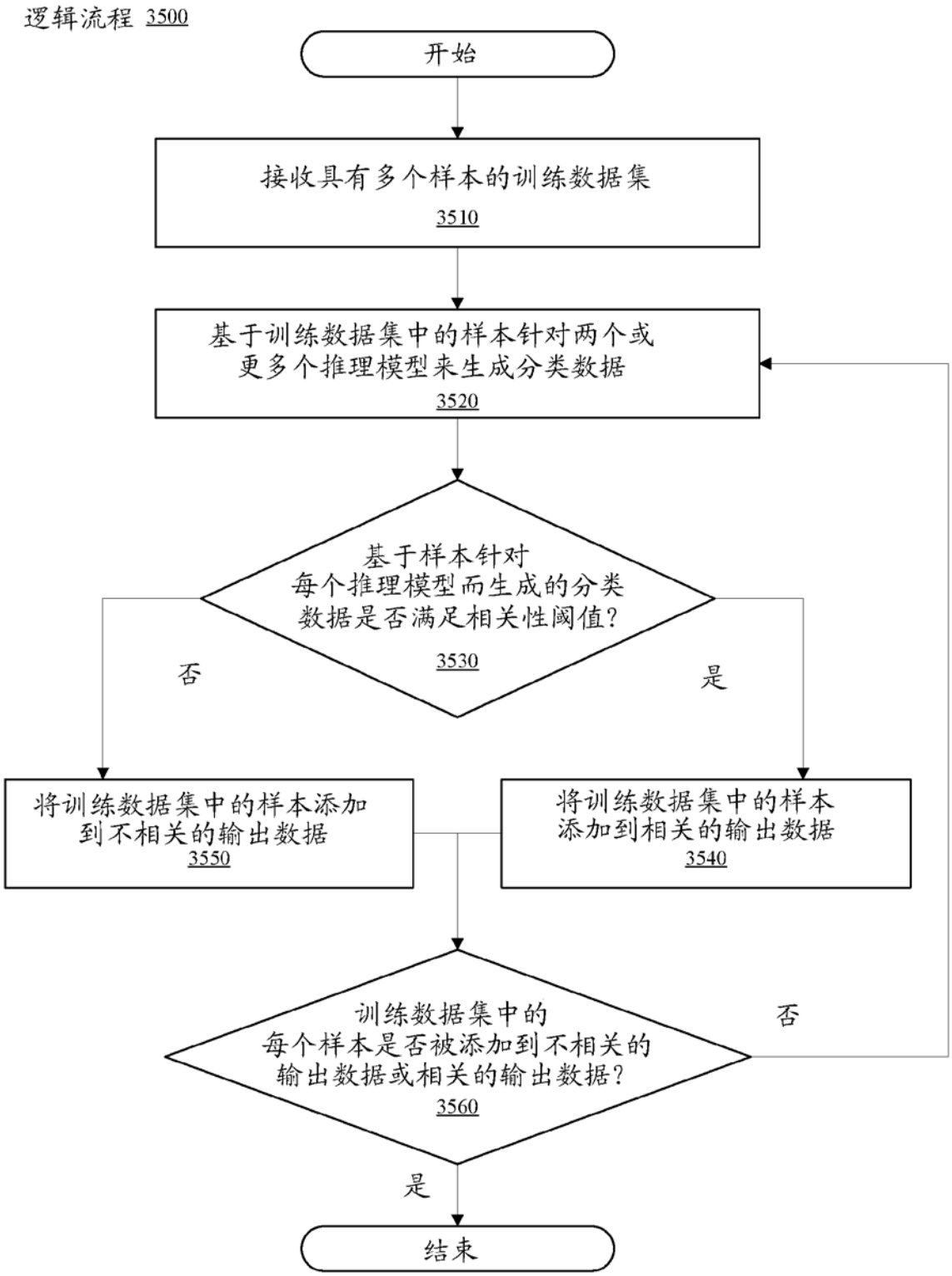

逻辑流程 3500
开始
接收具有多个样本的训练数据集
3510
基于训练数据集中的样本针对两个或更多个推理模型来生成分类数据
3520
基于样本针对每个推理模型而生成的分类数据是否满足相关性阈值?
3530
否
是
将训练数据集中的样本添加到不相关的输出数据
3550
将训练数据集中的样本添加到相关的输出数据
3540
训练数据集中的每个样本是否被添加到不相关的输出数据或相关的输出数据?
3560
否
是
结束

图 35

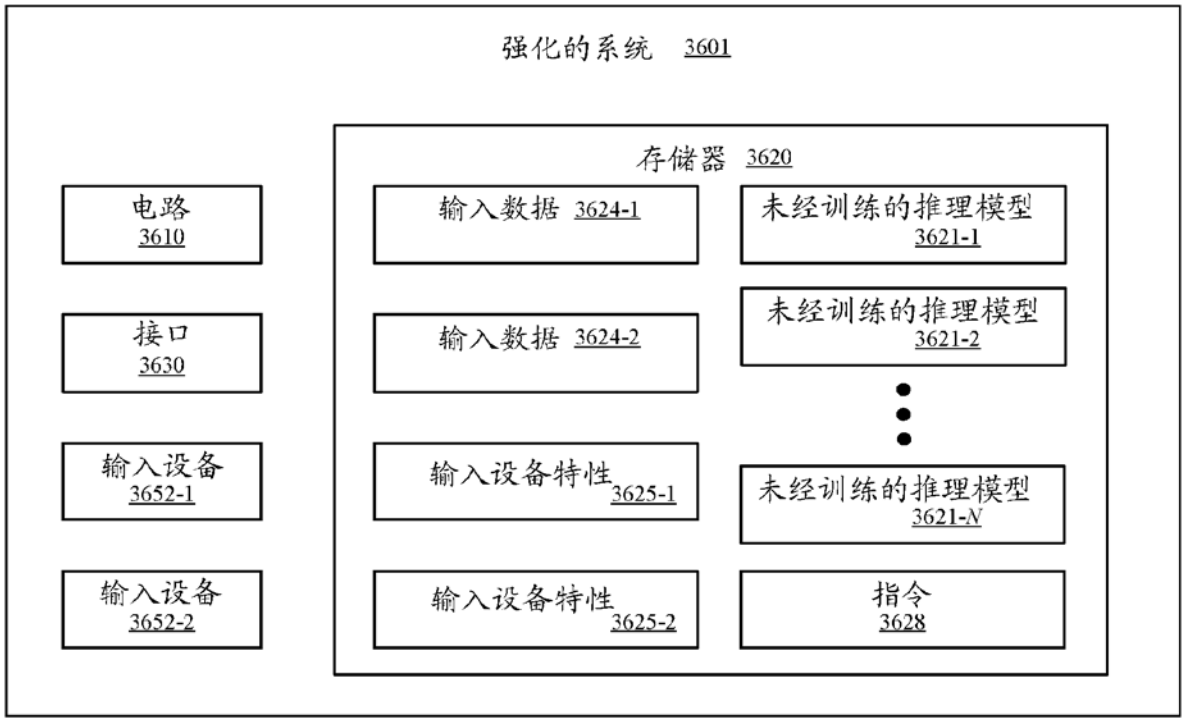

强化的系统 3601
存储器 3620
电路 3610
接口 3630
输入设备 3652-1
输入设备 3652-2
输入数据 3624-1
输入数据 3624-2
输入设备特性 3625-1
输入设备特性 3625-2
未经训练的推理模型 3621-1
未经训练的推理模型 3621-2
未经训练的推理模型 3621-N
指令 3628

图 36

逻辑流程 3700

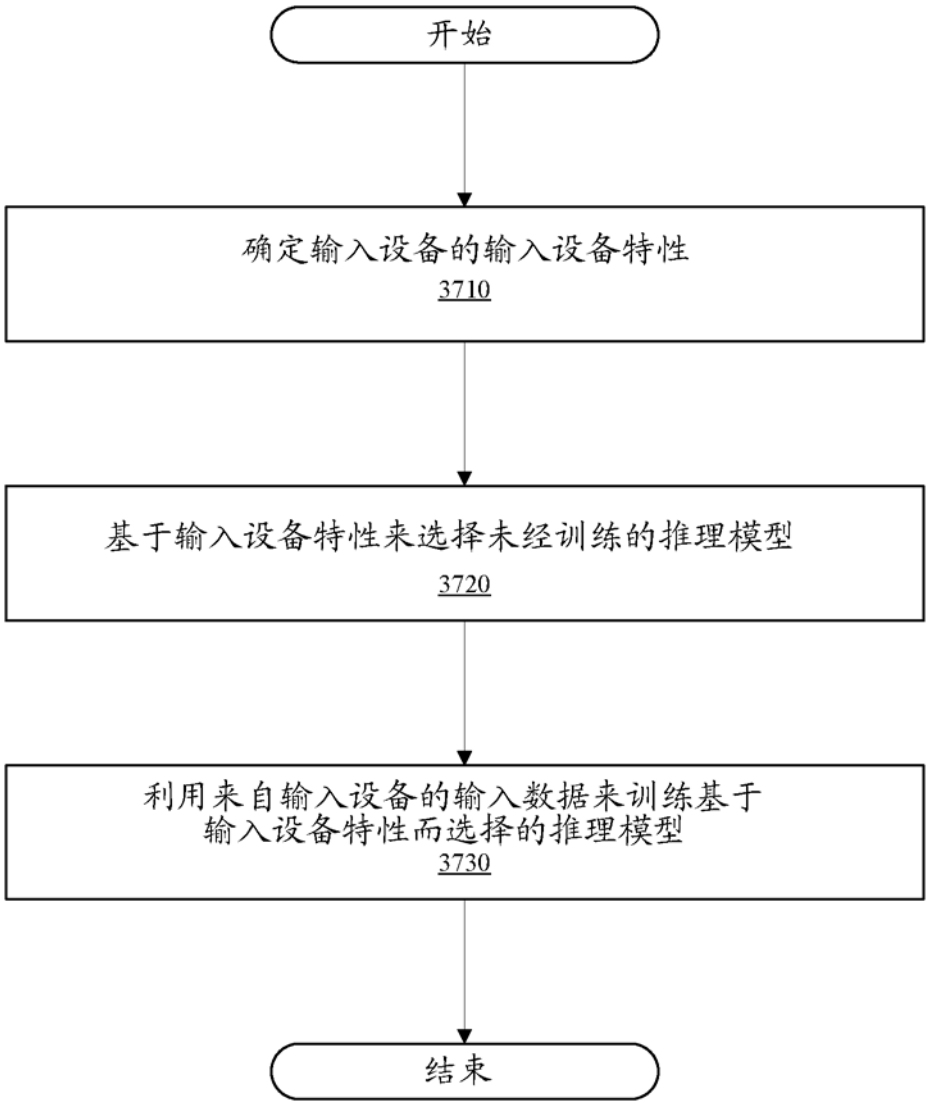



图 37

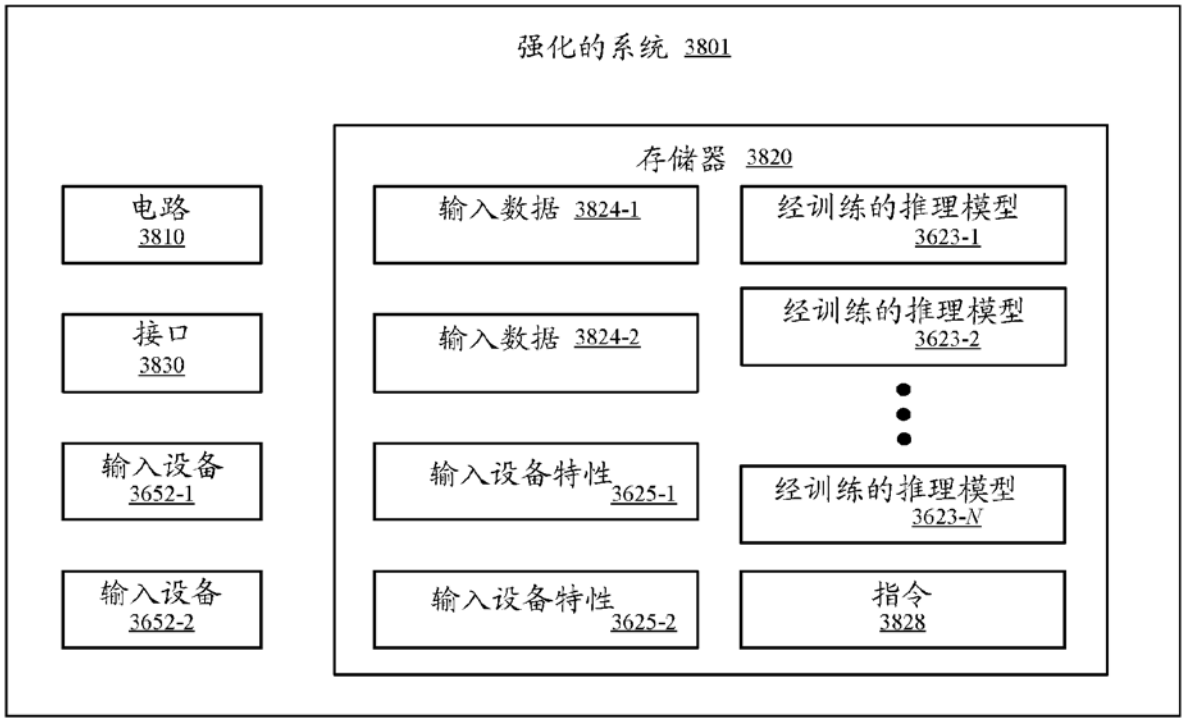

强化的系统 3801
存储器 3820
电路 3810
接口 3830
输入设备 3652-1
输入设备 3652-2
输入数据 3824-1
输入数据 3824-2
输入设备特性 3625-1
输入设备特性 3625-2
经训练的推理模型 3623-1
经训练的推理模型 3623-2
经训练的推理模型 3623-N
指令 3828

图 38

逻辑流程 3900

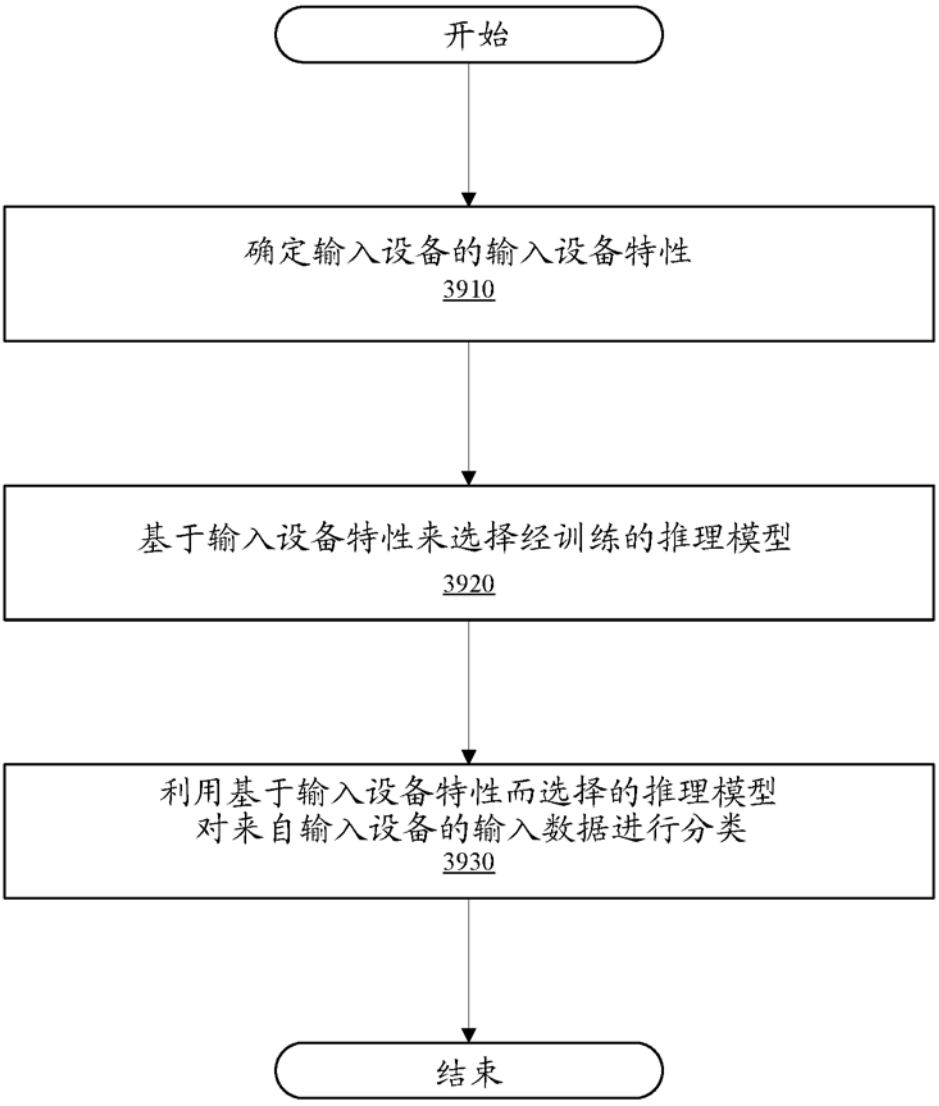



图 39

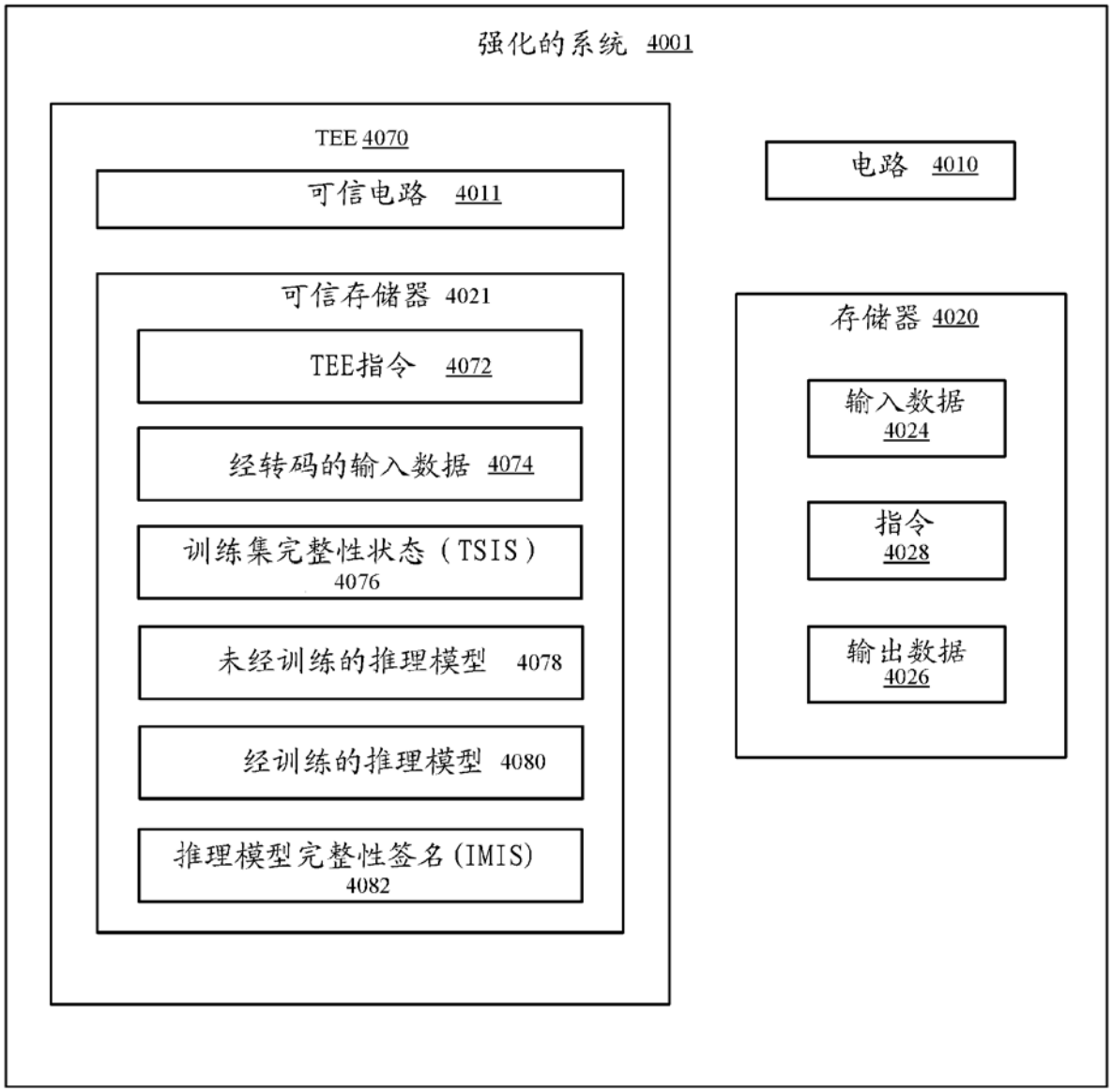

强化的系统 4001
TEE 4070
可信电路 4011
可信存储器 4021
TEE指令 4072
经转码的输入数据 4074
训练集完整性状态（TSIS）
4076
未经训练的推理模型 4078
经训练的推理模型 4080
推理模型完整性签名(IMIS)
4082
电路 4010
存储器 4020
输入数据
4024
指令
4028
输出数据
4026

图 40

逻辑流程 4100

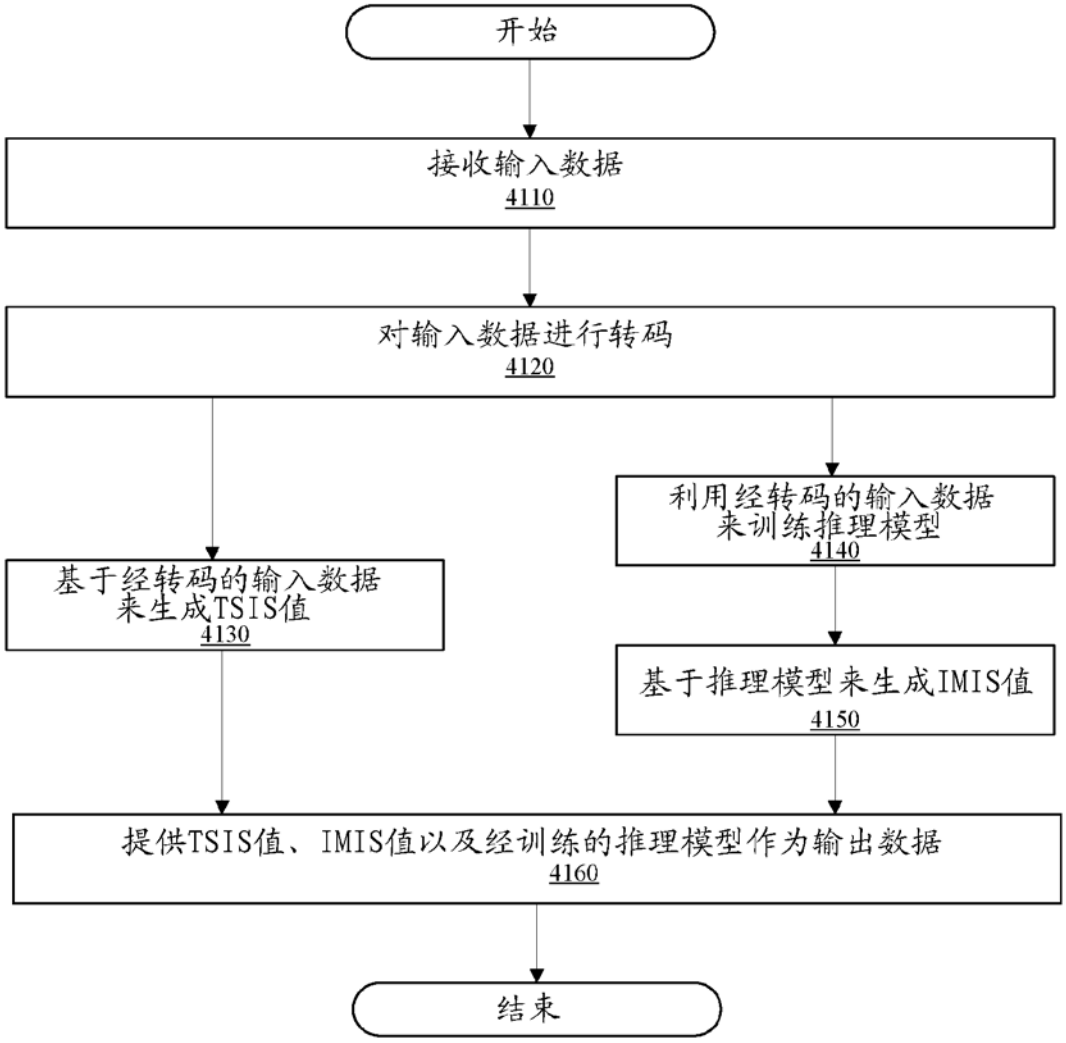



图 41

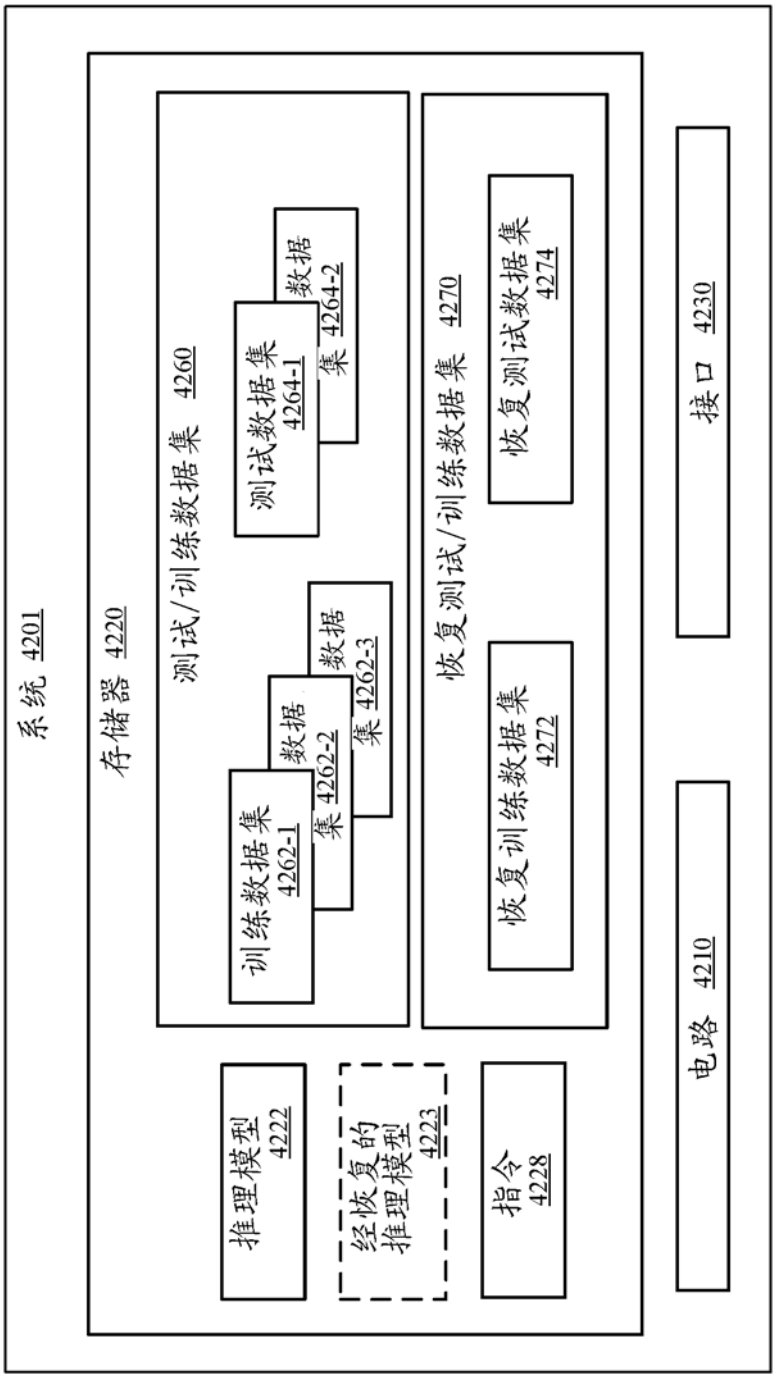

系统 4201
存储器 4220
测试/训练数据集 4260
训练数据集 4262-1
数据集 4262-2
数据集 4262-3
测试数据集 4264-1
数据集 4264-2
恢复测试/训练数据集 4270
恢复训练数据集 4272
恢复测试数据集 4274
推理模型 4222
经恢复的推理模型 4223
指令 4228
电路 4210
接口 4230

图 42

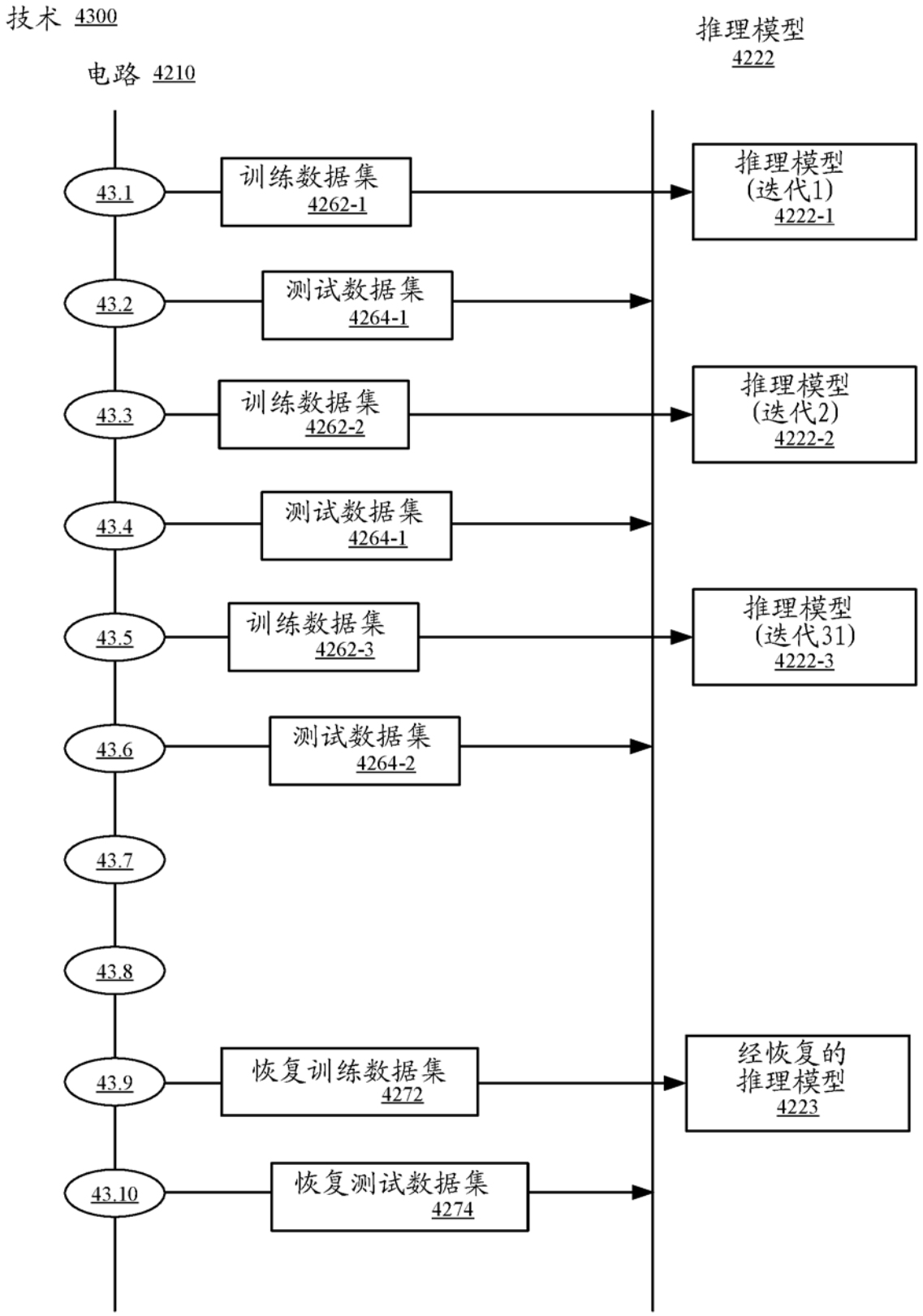

技术 4300
电路 4210
推理模型
4222
43.1
训练数据集
4262-1
推理模型
(迭代1)
4222-1
43.2
测试数据集
4264-1
43.3
训练数据集
4262-2
推理模型
(迭代2)
4222-2
43.4
测试数据集
4264-1
43.5
训练数据集
4262-3
推理模型
(迭代31)
4222-3
43.6
测试数据集
4264-2
43.7
43.8
43.9
恢复训练数据集
4272
经恢复的
推理模型
4223
43.10
恢复测试数据集
4274

图 43

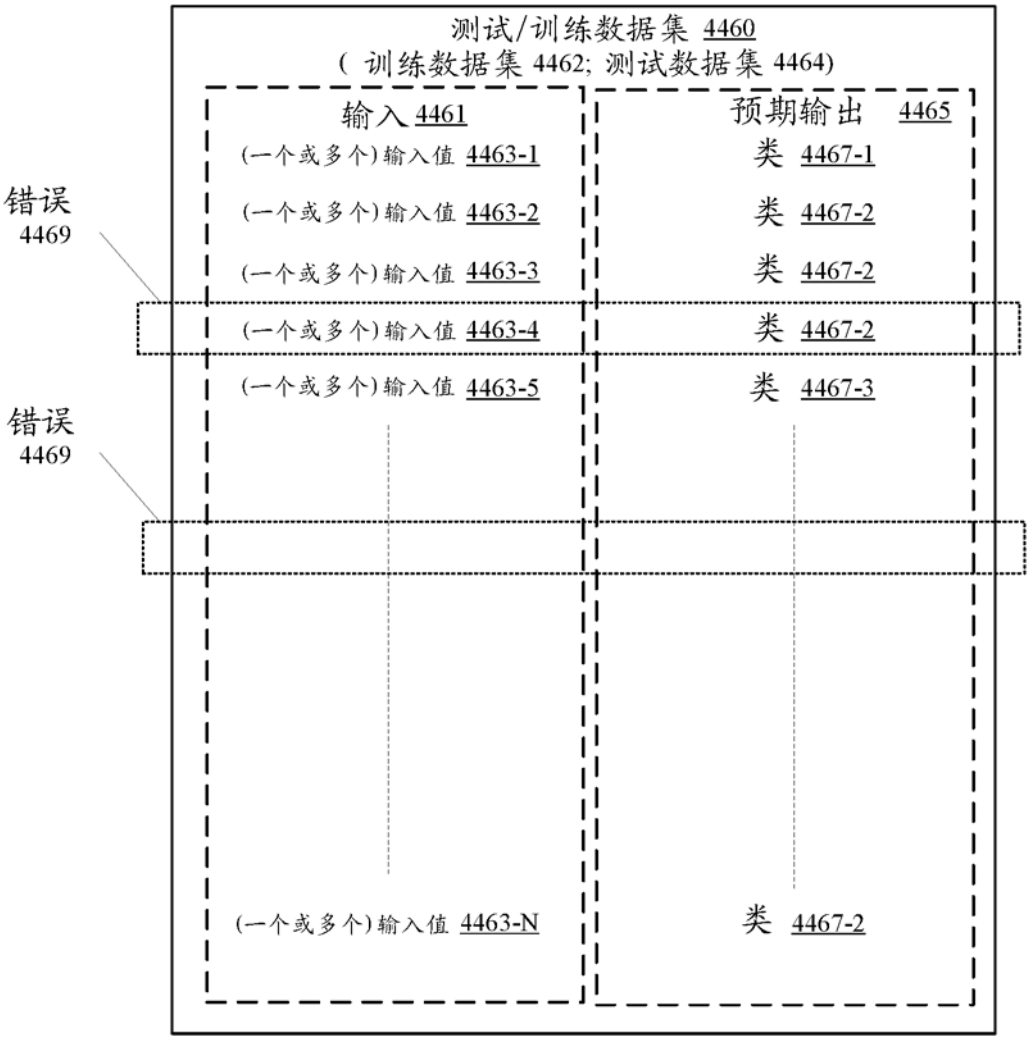

测试/训练数据集 4460
( 训练数据集 4462; 测试数据集 4464)
输入 4461
预期输出 4465
(一个或多个)输入值 4463-1
类 4467-1
(一个或多个)输入值 4463-2
类 4467-2
(一个或多个)输入值 4463-3
类 4467-2
(一个或多个)输入值 4463-4
类 4467-2
(一个或多个)输入值 4463-5
类 4467-3
(一个或多个)输入值 4463-N
类 4467-2
错误 4469
错误 4469

图 44

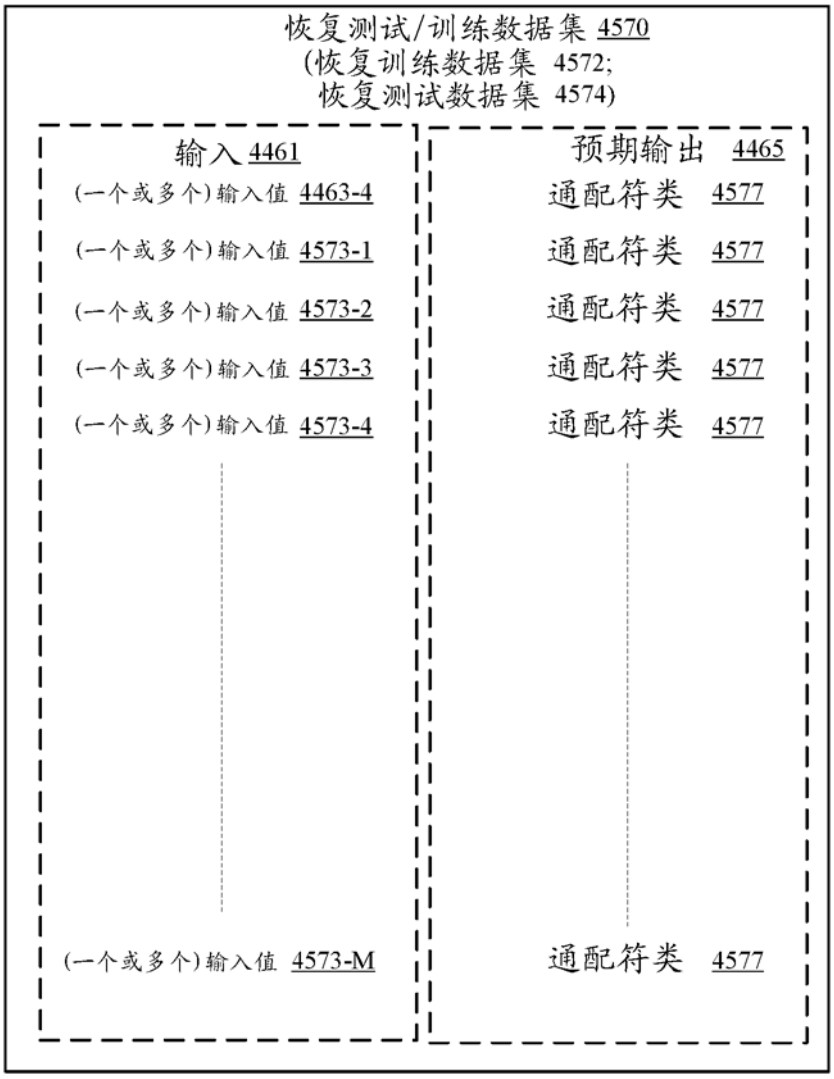

恢复测试/训练数据集 4570
(恢复训练数据集 4572;
恢复测试数据集 4574)
输入 4461
(一个或多个)输入值 4463-4
(一个或多个)输入值 4573-1
(一个或多个)输入值 4573-2
(一个或多个)输入值 4573-3
(一个或多个)输入值 4573-4
(一个或多个)输入值 4573-M
预期输出 4465
通配符类 4577
通配符类 4577
通配符类 4577
通配符类 4577
通配符类 4577
通配符类 4577

图 45

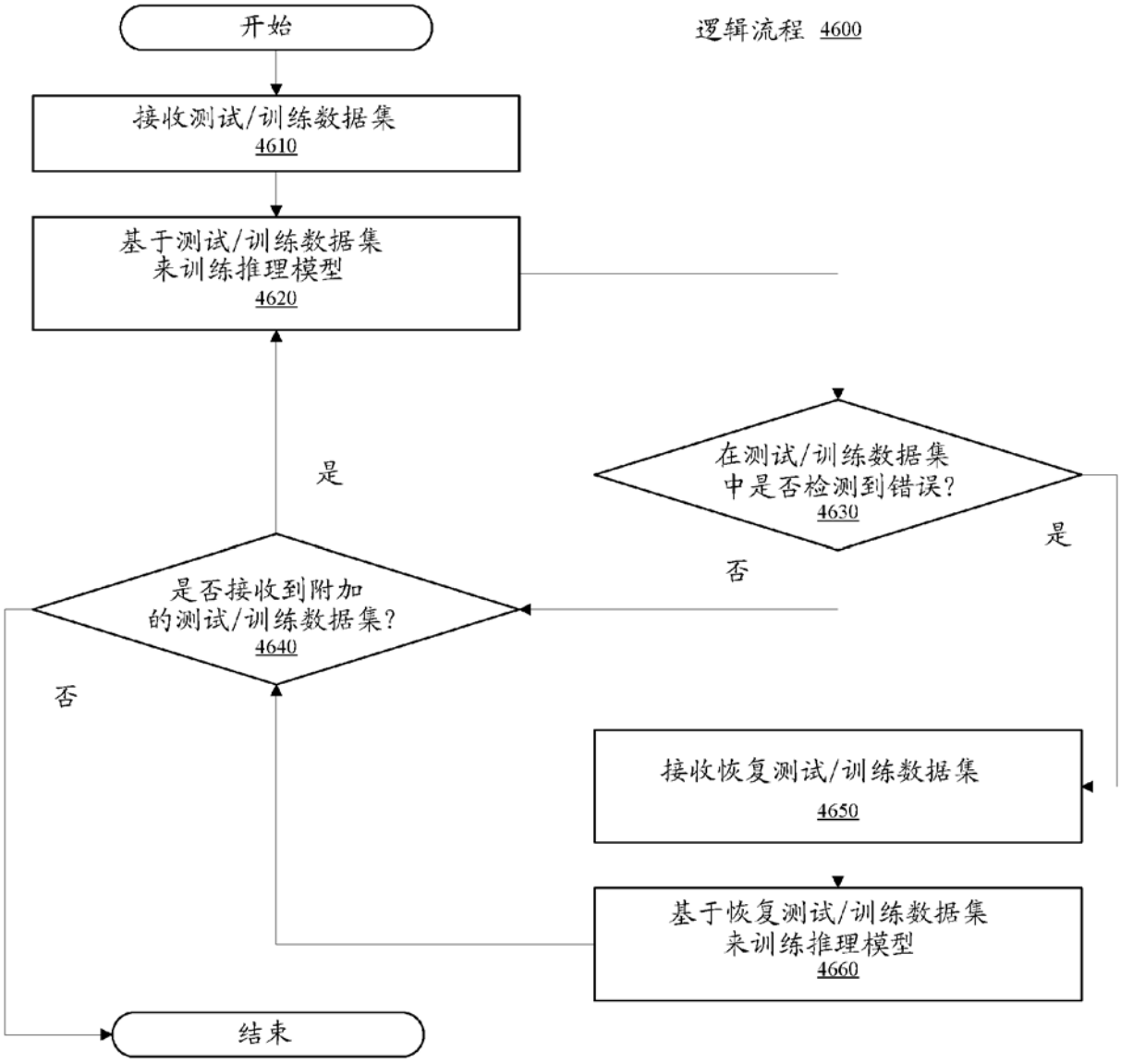

开始
逻辑流程 4600
接收测试/训练数据集
4610
基于测试/训练数据集
来训练推理模型
4620
在测试/训练数据集
中是否检测到错误?
4630
是
否
是
是否接收到附加
的测试/训练数据集?
4640
否
接收恢复测试/训练数据集
4650
基于恢复测试/训练数据集
来训练推理模型
4660
结束

图 46

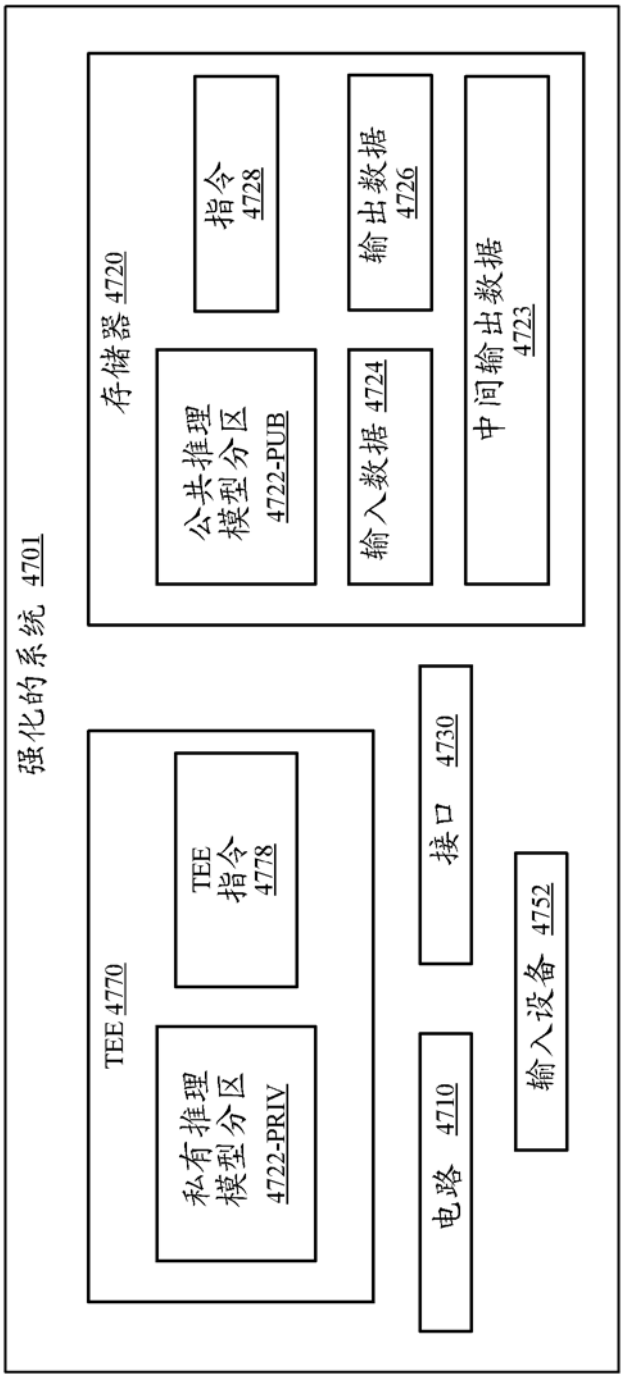

强化的系统 4701
存储器 4720
公共推理模型分区 4722-PUB
指令 4728
输入数据 4724
输出数据 4726
中间输出数据 4723
TEE 4770
私有推理模型分区 4722-PRIV
TEE 指令 4778
接口 4730
电路 4710
输入设备 4752

图 47

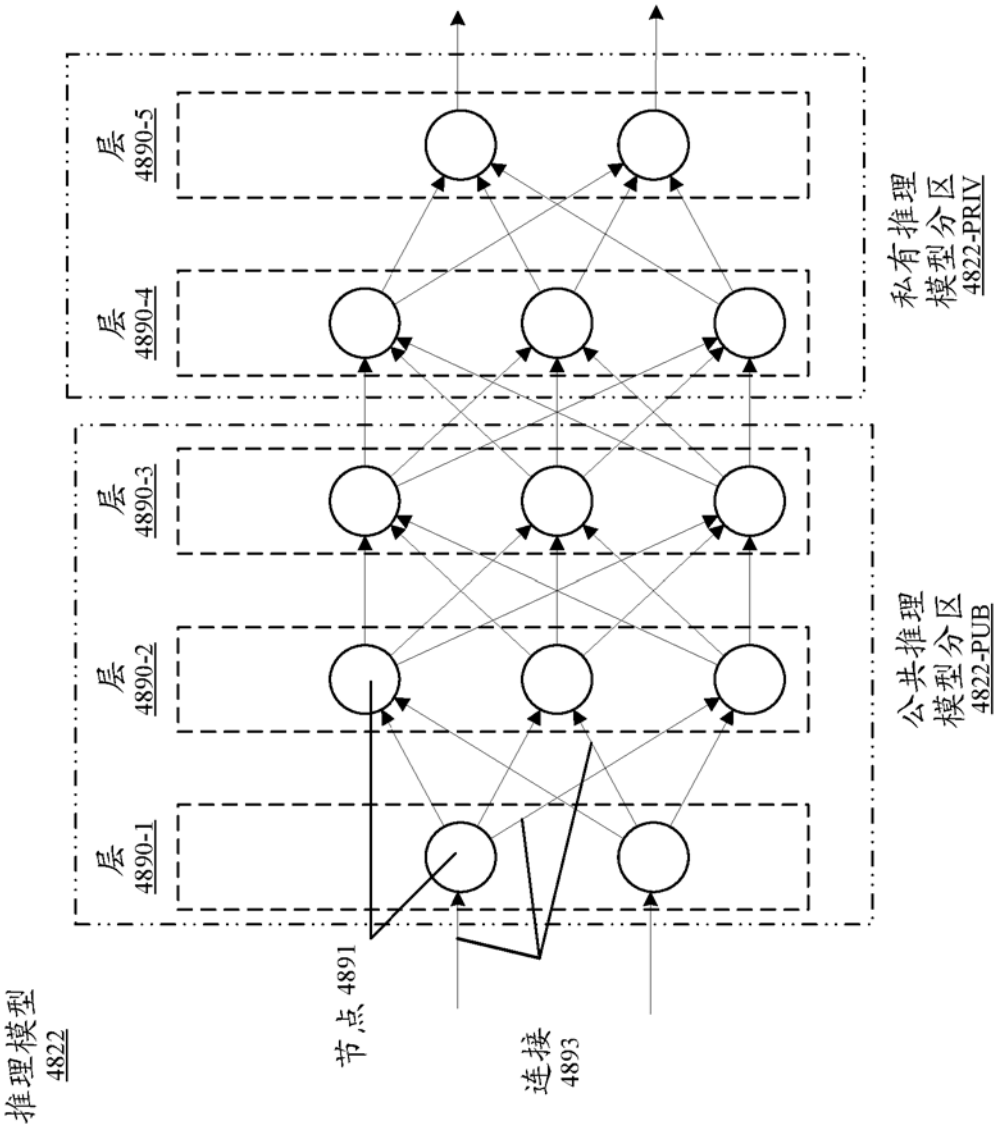

推理模型
4822
层
4890-1
层
4890-2
层
4890-3
层
4890-4
层
4890-5
公共推理
模型分区
4822-PUB
私有推理
模型分区
4822-PRIV
节点 4891
连接
4893

图 48

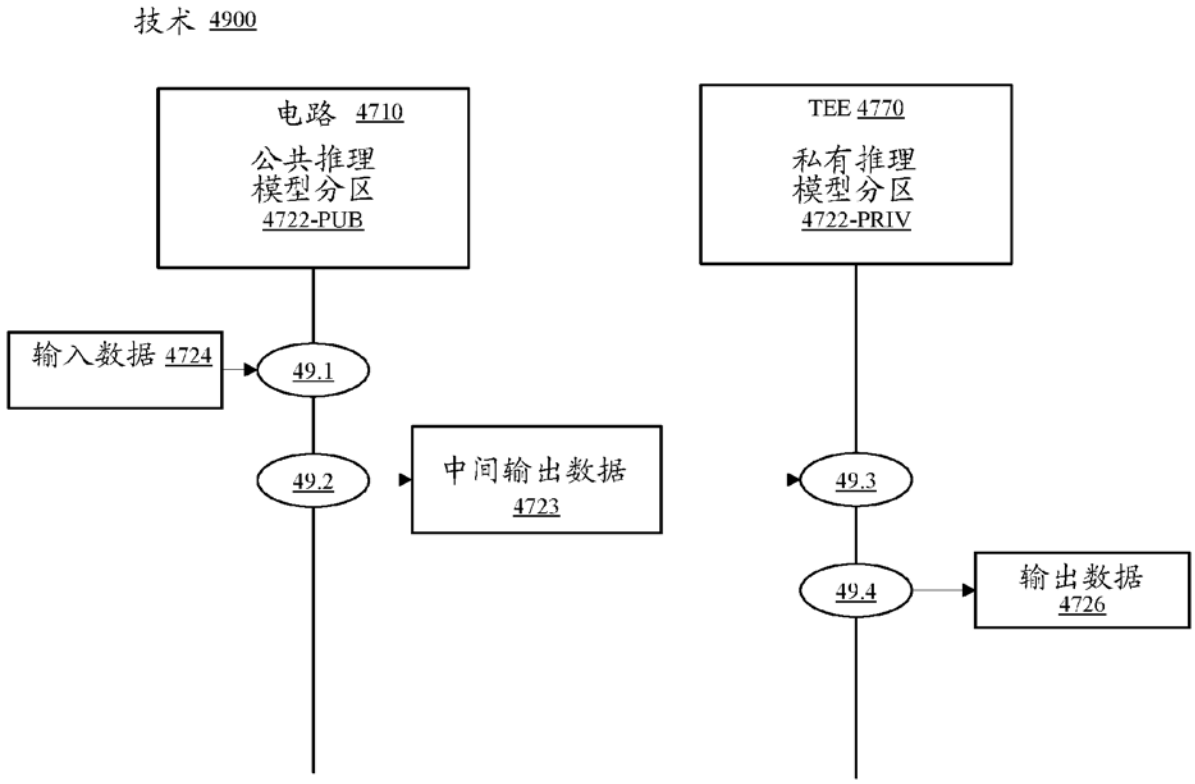

技术 4900
电路 4710
公共推理
模型分区
4722-PUB
TEE 4770
私有推理
模型分区
4722-PRIV
输入数据 4724
49.1
49.2
中间输出数据
4723
49.3
49.4
输出数据
4726

图 49

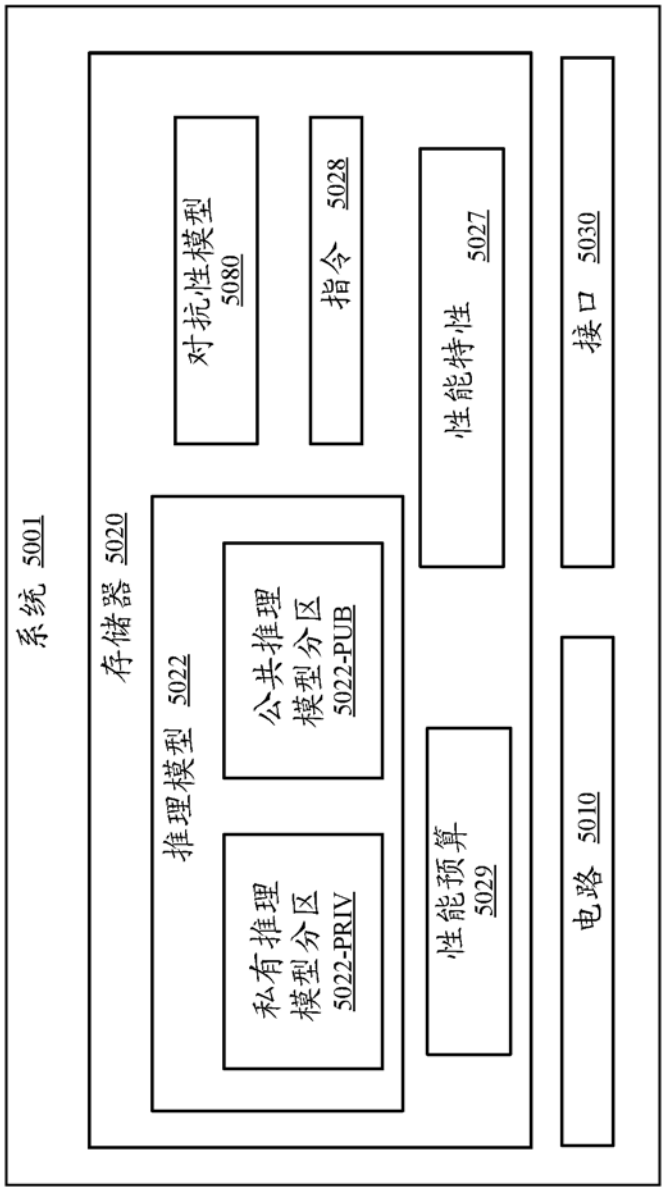

系统 5001
存储器 5020
推理模型 5022
私有推理模型分区 5022-PRIV
公共推理模型分区 5022-PUB
性能预算 5029
对抗性模型 5080
指令 5028
性能特性 5027
电路 5010
接口 5030

图 50

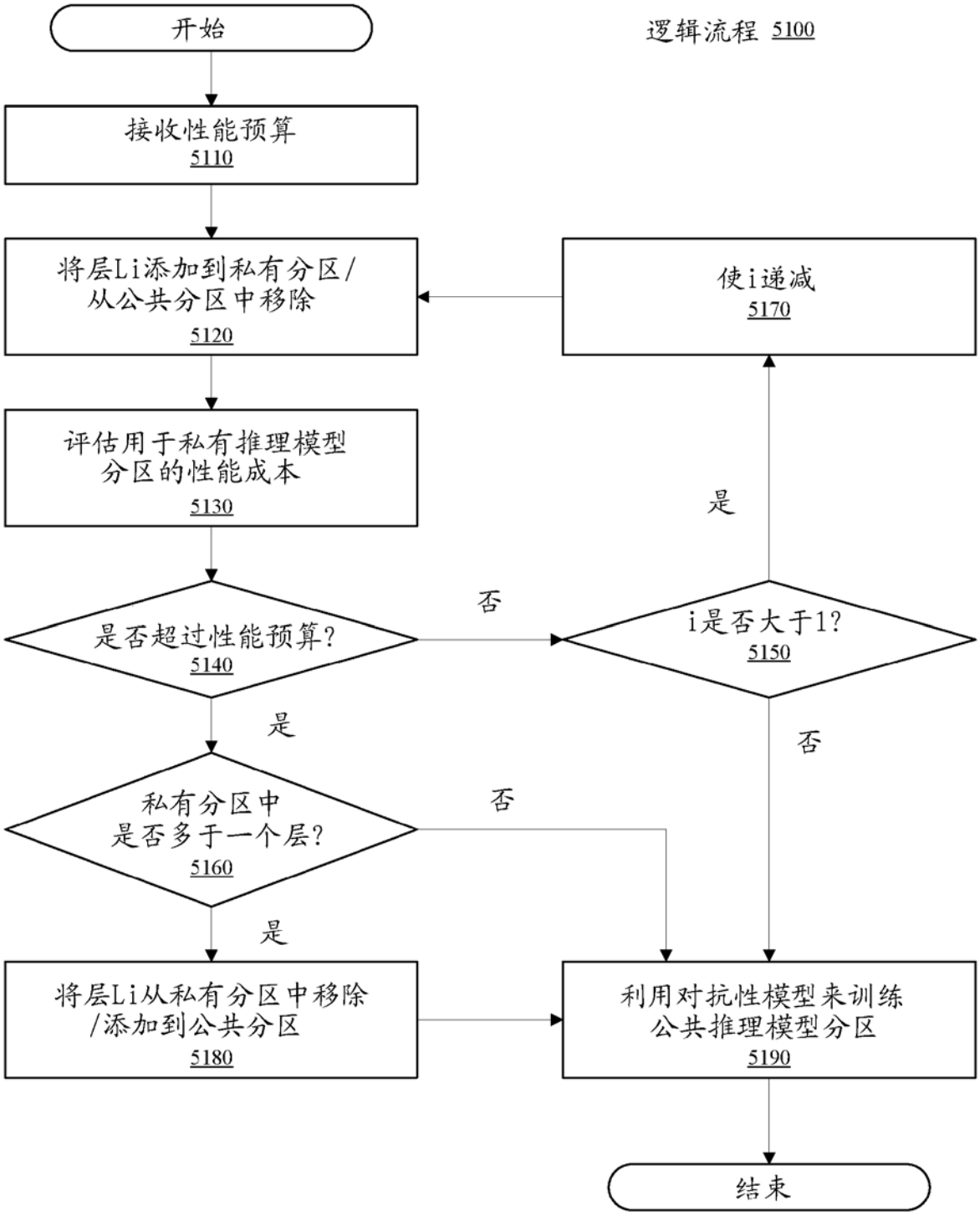

开始
逻辑流程 5100
接收性能预算
5110
将层Li添加到私有分区/
从公共分区中移除
5120
使i递减
5170
评估用于私有推理模型
分区的性能成本
5130
是
是否超过性能预算?
5140
否
i是否大于1?
5150
是
否
私有分区中
是否多于一个层?
5160
否
是
将层Li从私有分区中移除
/添加到公共分区
5180
利用对抗性模型来训练
公共推理模型分区
5190
结束

图 51

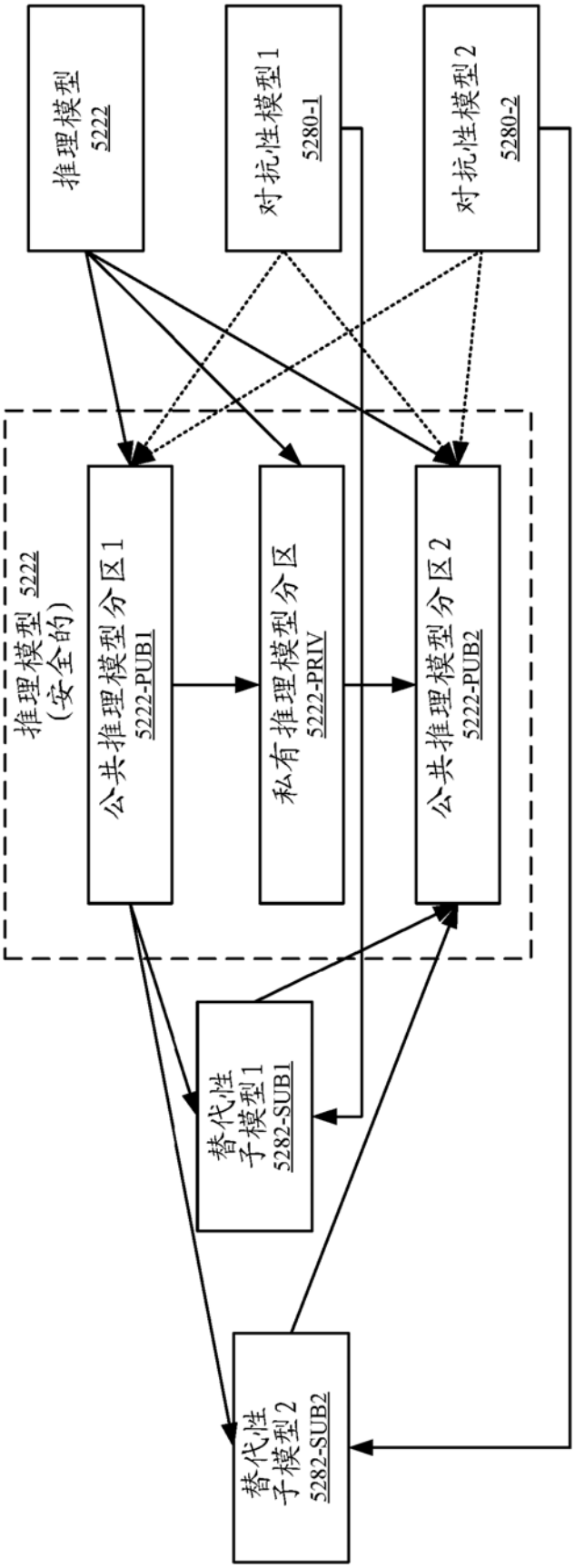

推理模型
5222
对抗性模型1
5280-1
对抗性模型2
5280-2
推理模型 5222
(安全的)
公共推理模型分区1
5222-PUB1
私有推理模型分区
5222-PRIV
公共推理模型分区2
5222-PUB2
替代性
子模型1
5282-SUB1
替代性
子模型2
5282-SUB2

图 52

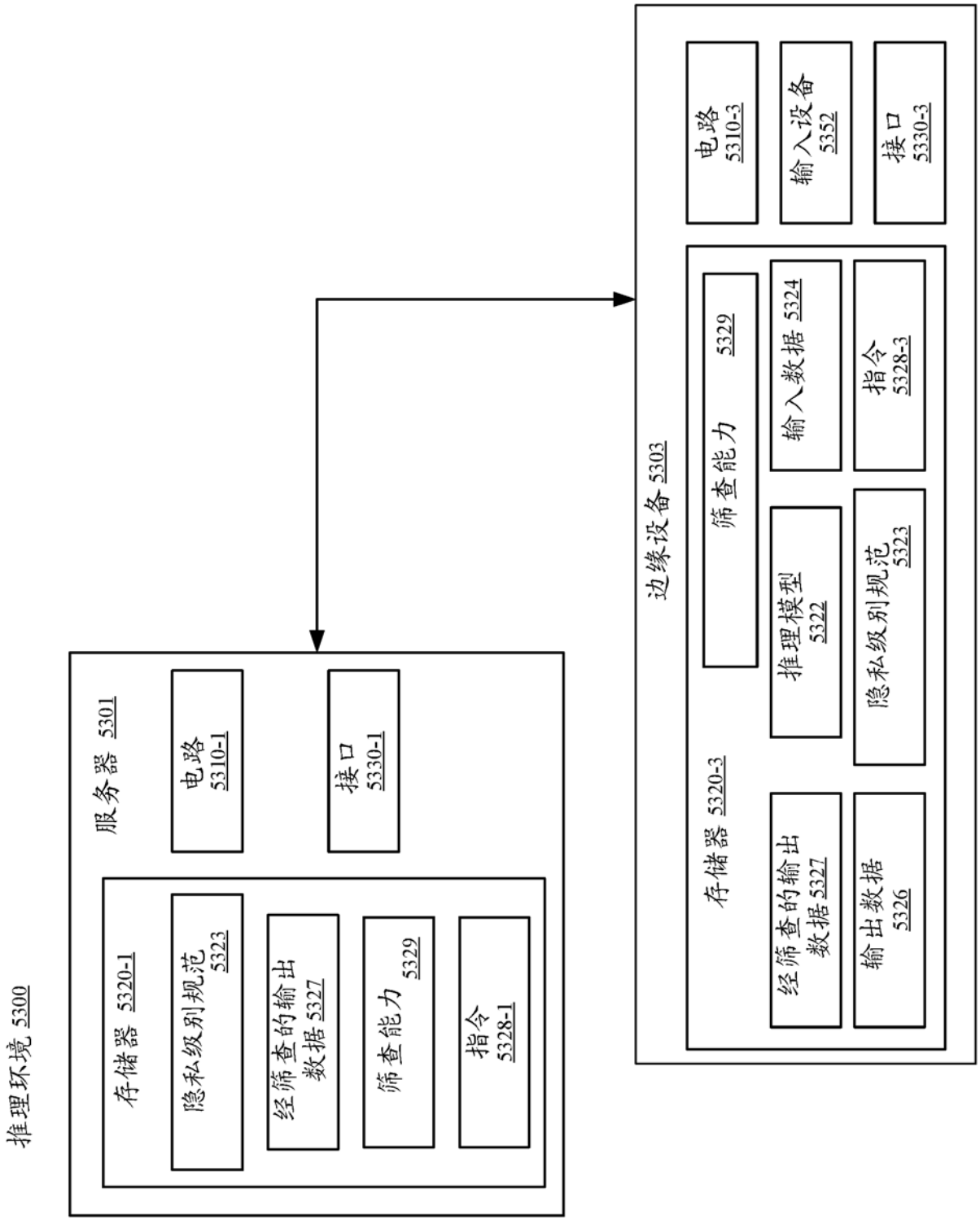

推理环境 5300
服务器 5301
电路 5310-1
接口 5330-1
存储器 5320-1
隐私级别规范 5323
经筛查的输出数据 5327
筛查能力 5329
指令 5328-1
边缘设备 5303
电路 5310-3
输入设备 5352
接口 5330-3
存储器 5320-3
筛查能力 5329
输入数据 5324
推理模型 5322
指令 5328-3
隐私级别规范 5323
经筛查的输出数据 5327
输出数据 5326

图 53

技术 5400

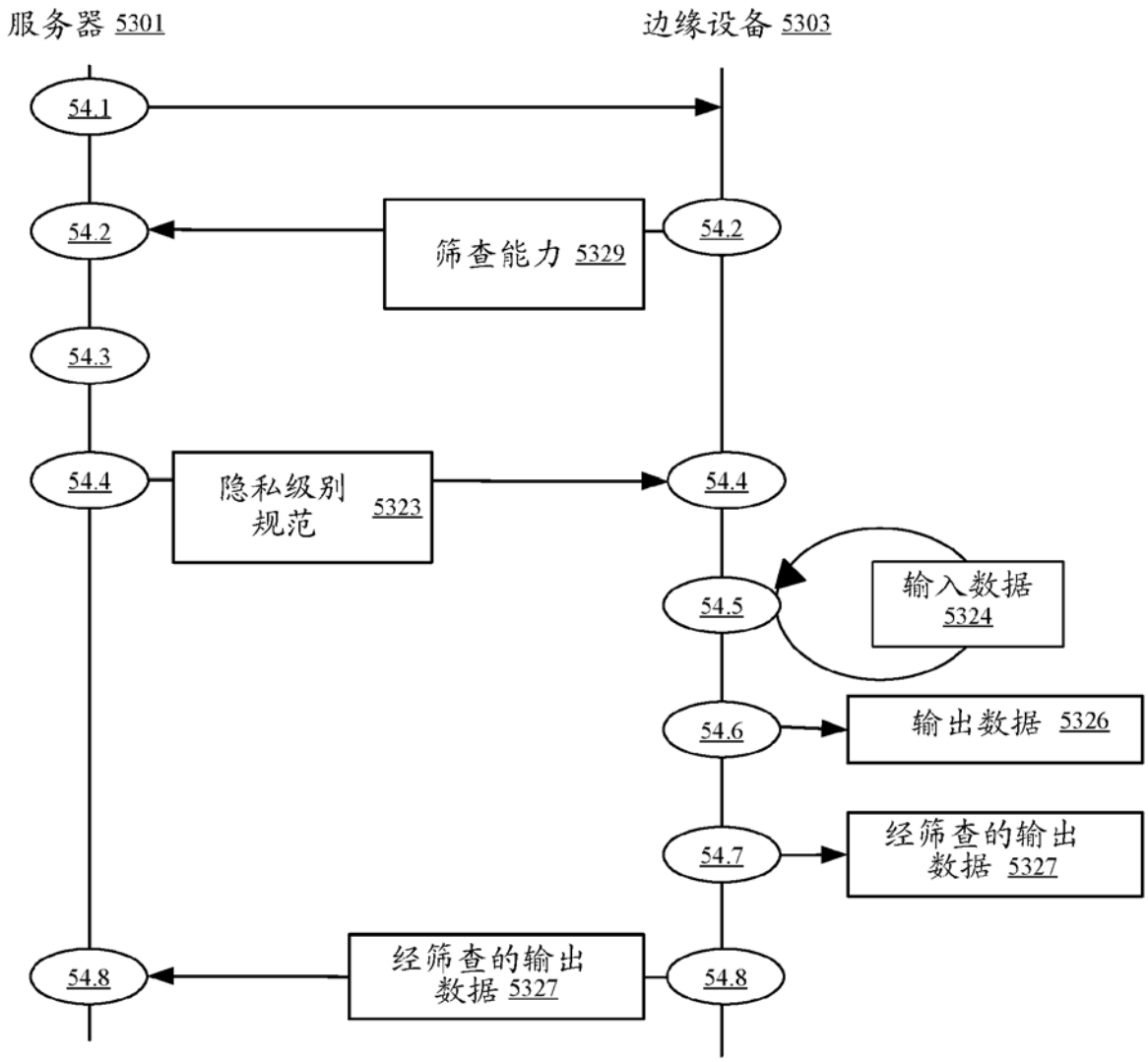

服务器 5301
边缘设备 5303
54.1
54.2
筛查能力 5329
54.2
54.3
54.4
隐私级别规范 5323
54.4
54.5
输入数据 5324
54.6
输出数据 5326
54.7
经筛查的输出数据 5327
54.8
经筛查的输出数据 5327
54.8

图 54

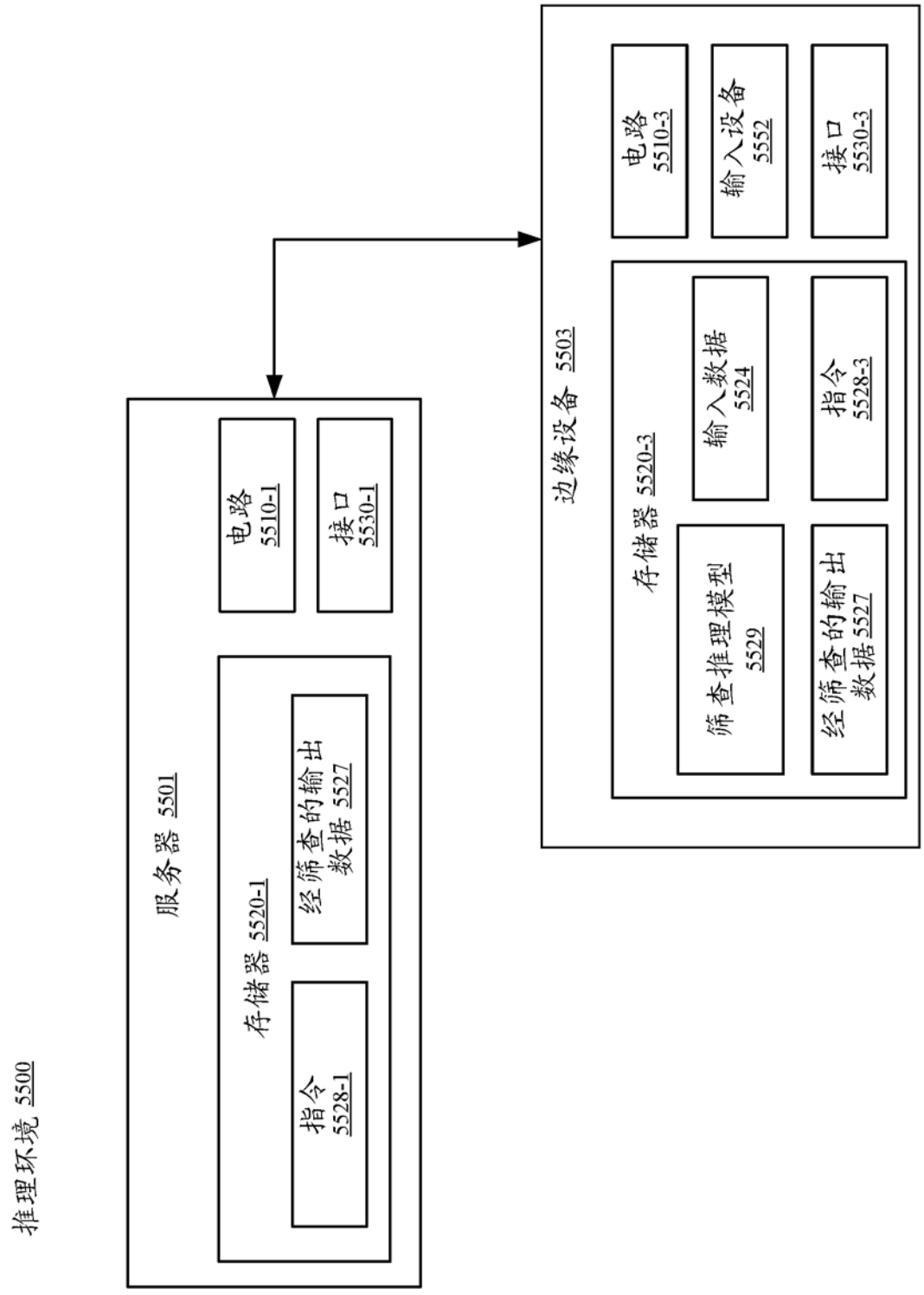

推理环境 5500
服务器 5501
存储器 5520-1
指令 5528-1
经筛查的输出数据 5527
电路 5510-1
接口 5530-1
边缘设备 5503
存储器 5520-3
筛查推理模型 5529
输入数据 5524
经筛查的输出数据 5527
指令 5528-3
电路 5510-3
输入设备 5552
接口 5530-3

图 55

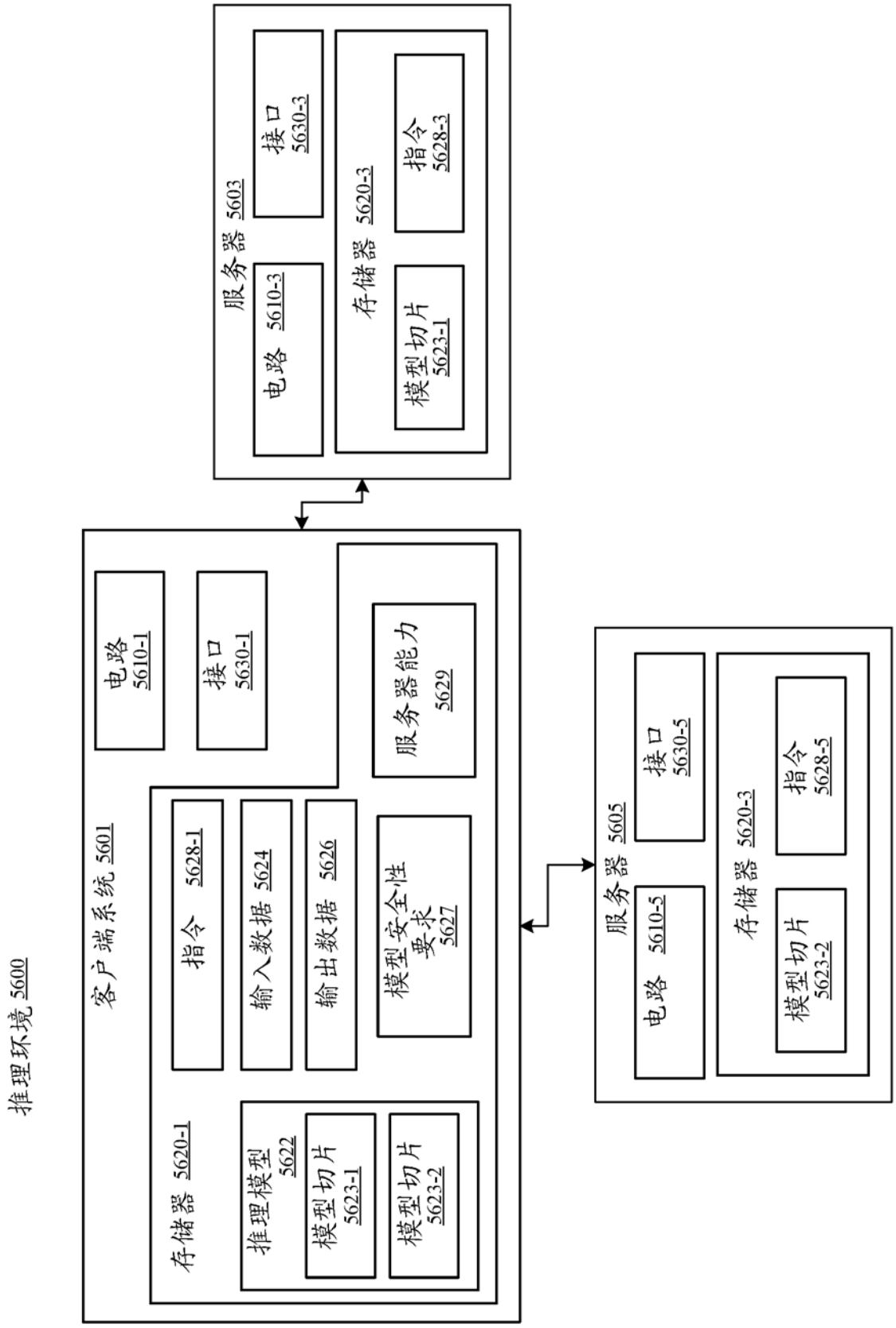

推理环境 5600
客户端系统 5601
电路 5610-1
接口 5630-1
服务器能力 5629
指令 5628-1
输入数据 5624
输出数据 5626
模型安全性要求 5627
存储器 5620-1
推理模型 5622
模型切片 5623-1
模型切片 5623-2
服务器 5603
电路 5610-3
接口 5630-3
存储器 5620-3
模型切片 5623-1
指令 5628-3
服务器 5605
电路 5610-5
接口 5630-5
存储器 5620-3
模型切片 5623-2
指令 5628-5

图 56

技术 5700

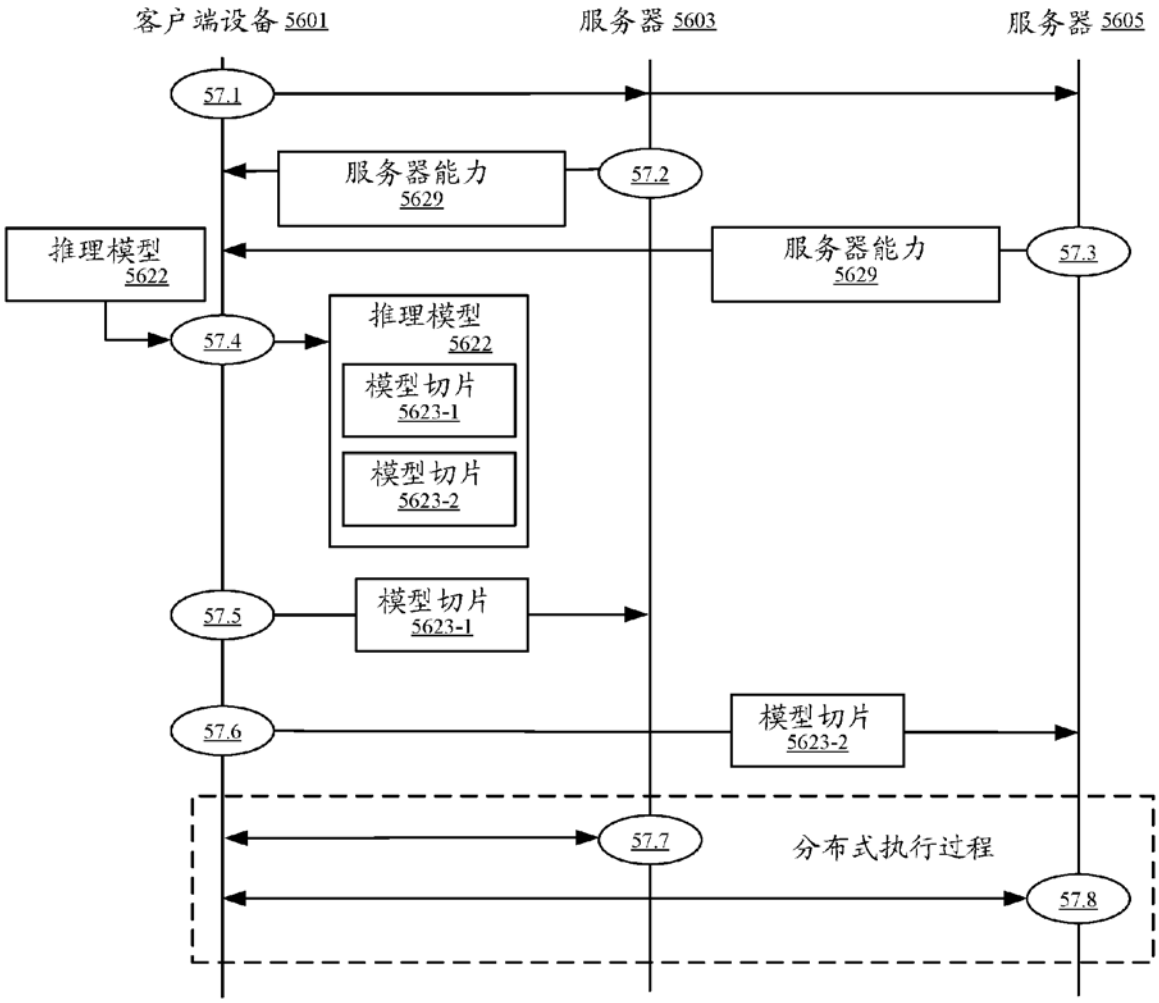

客户端设备 5601
服务器 5603
服务器 5605
57.1
服务器能力
5629
57.2
推理模型
5622
服务器能力
5629
57.3
57.4
推理模型
5622
模型切片
5623-1
模型切片
5623-2
57.5
模型切片
5623-1
57.6
模型切片
5623-2
57.7
分布式执行过程
57.8

图 57

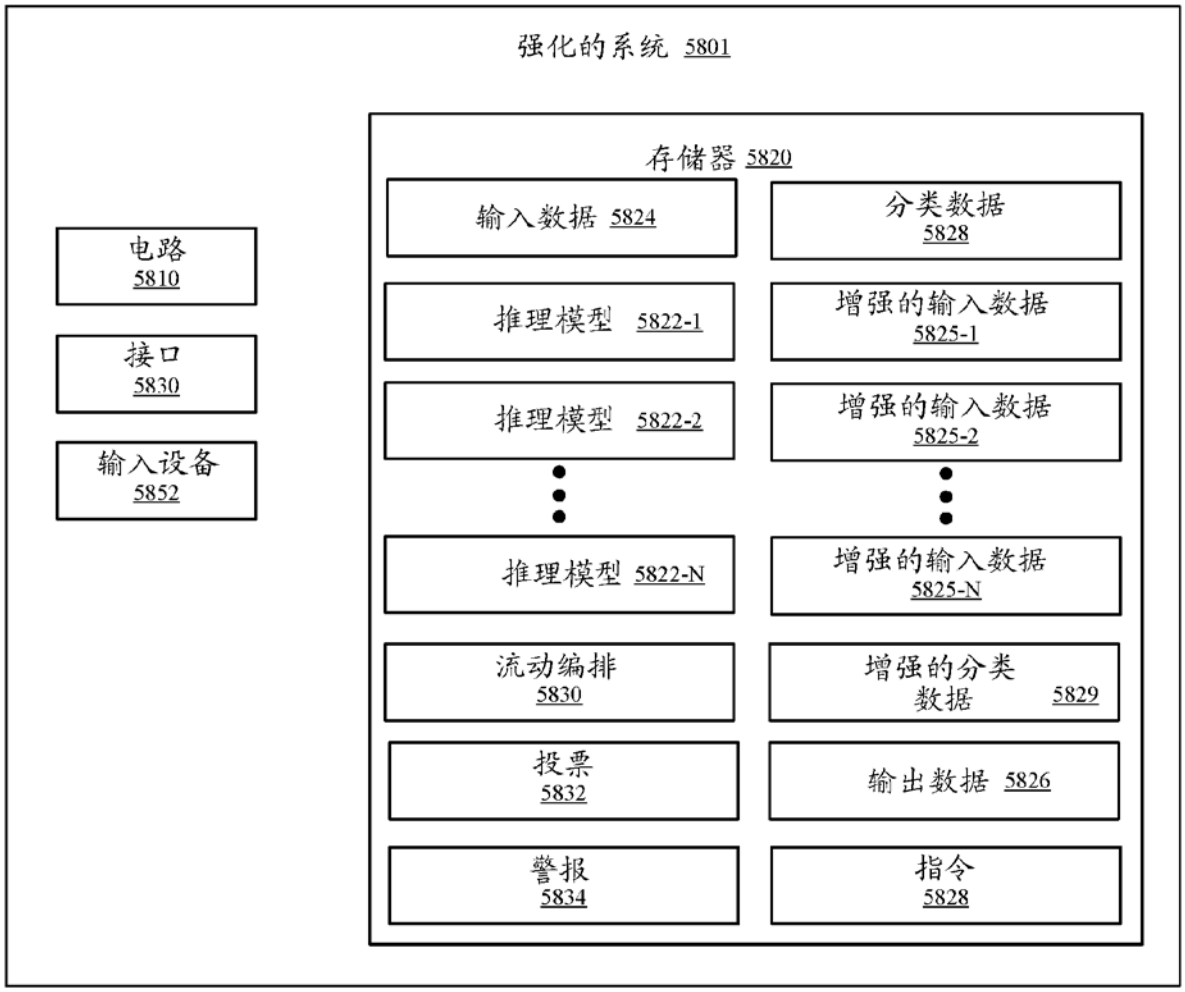

强化的系统 5801
电路 5810
接口 5830
输入设备 5852
存储器 5820
输入数据 5824
分类数据 5828
推理模型 5822-1
增强的输入数据 5825-1
推理模型 5822-2
增强的输入数据 5825-2
推理模型 5822-N
增强的输入数据 5825-N
流动编排 5830
增强的分类数据 5829
投票 5832
输出数据 5826
警报 5834
指令 5828

图 58

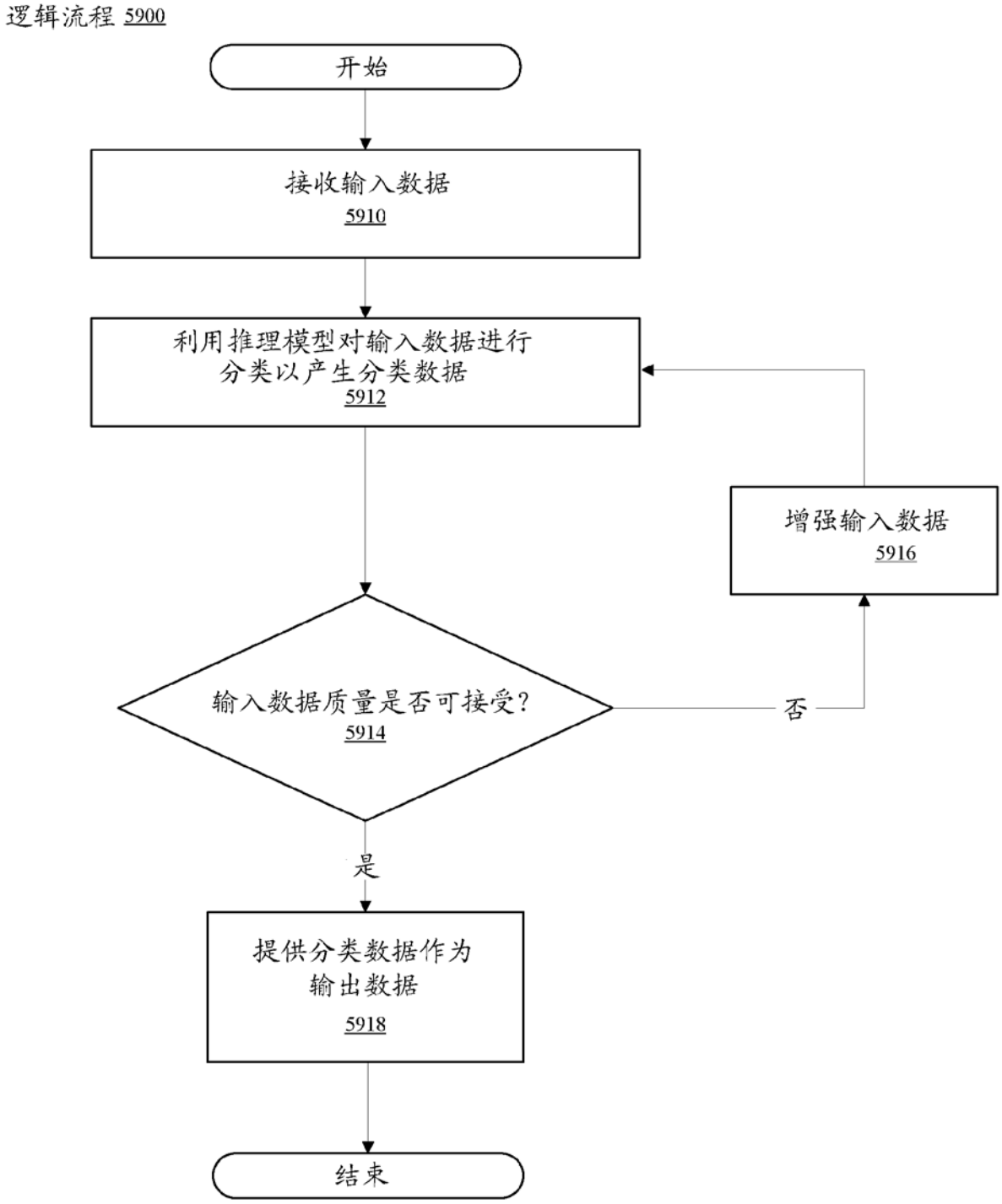

逻辑流程 5900
开始
接收输入数据
5910
利用推理模型对输入数据进行
分类以产生分类数据
5912
增强输入数据
5916
输入数据质量是否可接受?
5914
否
是
提供分类数据作为
输出数据
5918
结束

图 59

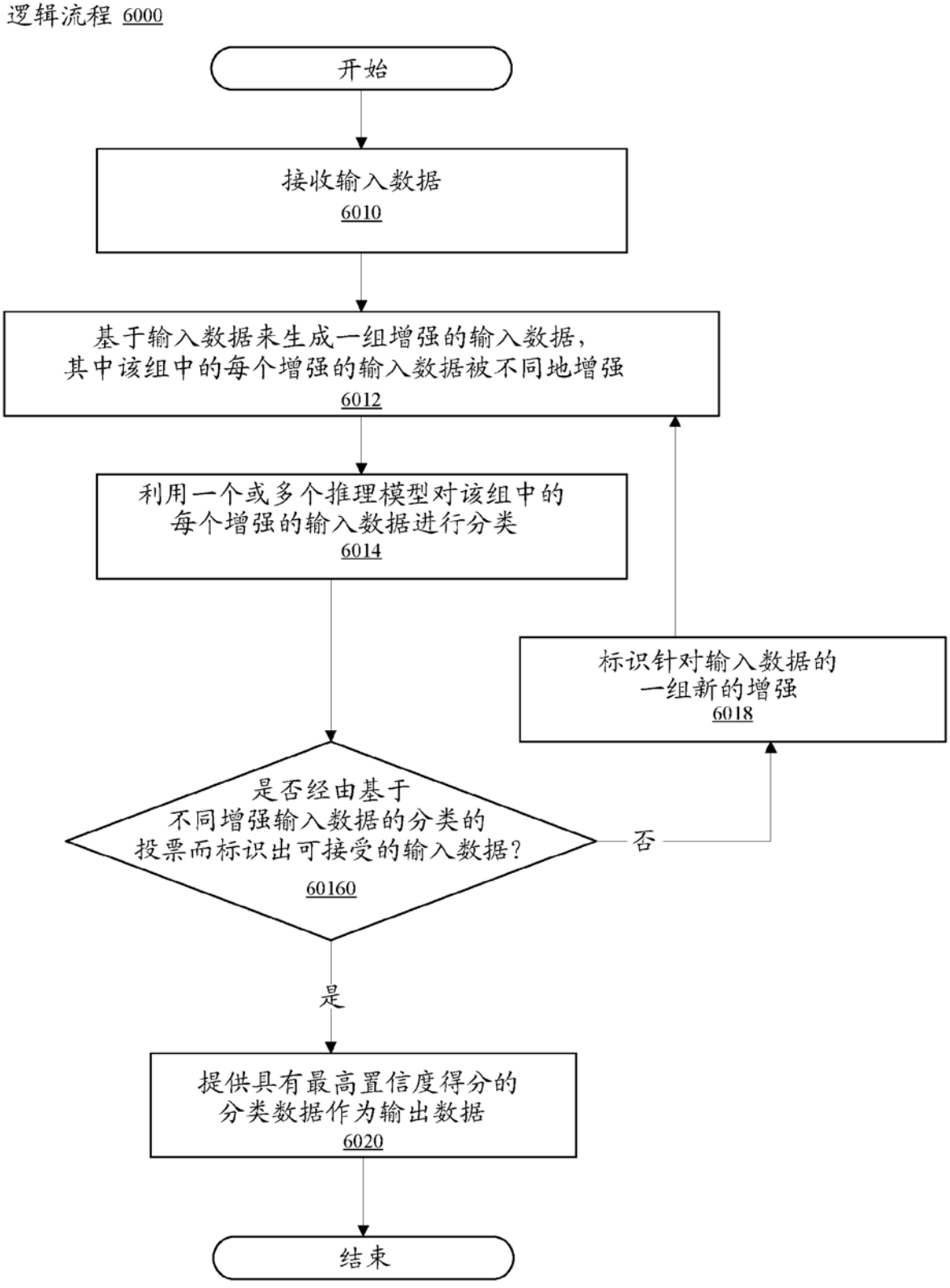

逻辑流程 6000
开始
接收输入数据
6010
基于输入数据来生成一组增强的输入数据，
其中该组中的每个增强的输入数据被不同地增强
6012
利用一个或多个推理模型对该组中的
每个增强的输入数据进行分类
6014
标识针对输入数据的
一组新的增强
6018
是否经由基于
不同增强输入数据的分类的
投票而标识出可接受的输入数据？
60160
否
是
提供具有最高置信度得分的
分类数据作为输出数据
6020
结束

图 60

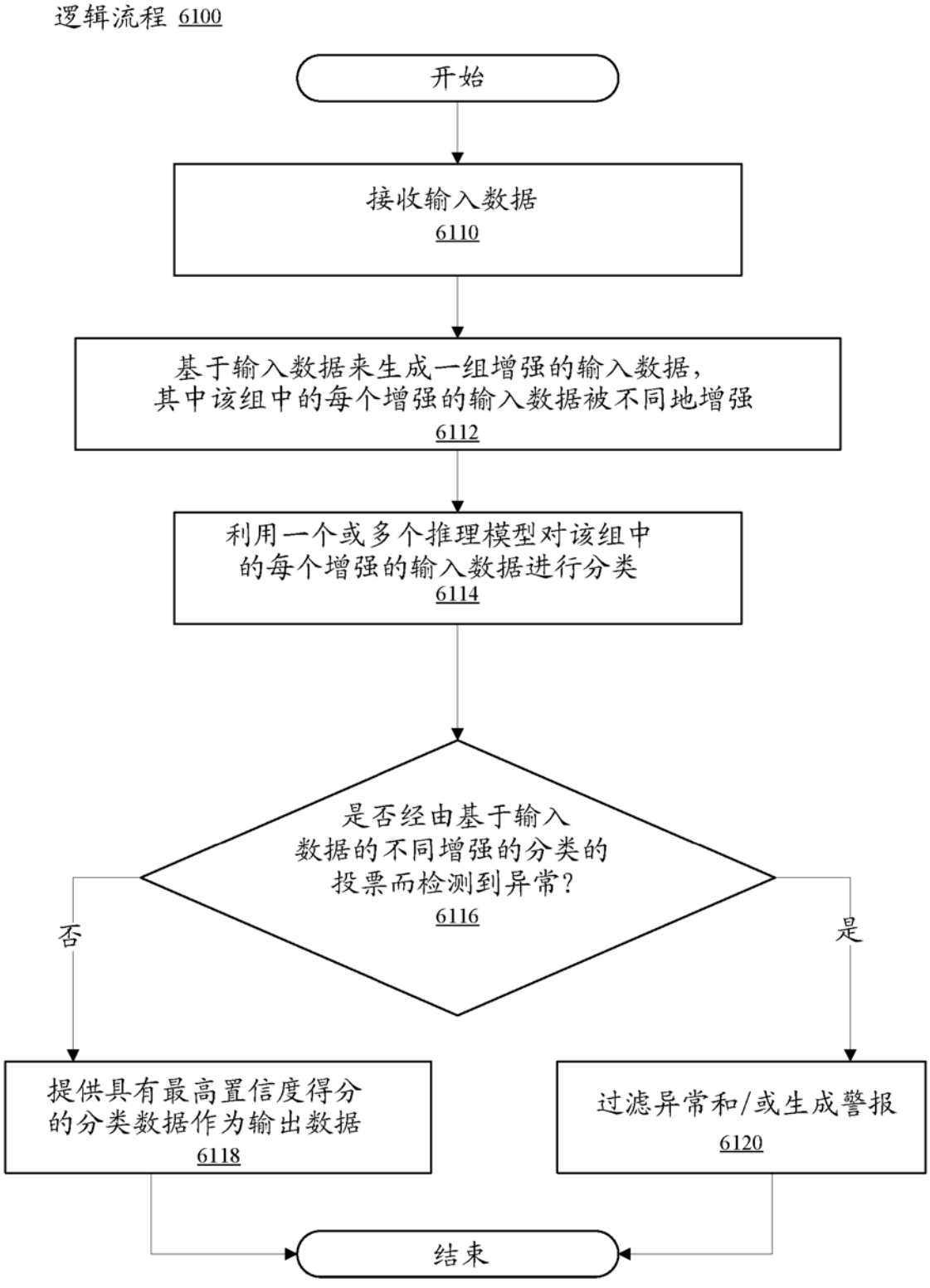

逻辑流程 6100
开始
接收输入数据
6110
基于输入数据来生成一组增强的输入数据，
其中该组中的每个增强的输入数据被不同地增强
6112
利用一个或多个推理模型对该组中
的每个增强的输入数据进行分类
6114
是否经由基于输入
数据的不同增强的分类的
投票而检测到异常？
6116
否
是
提供具有最高置信度得分
的分类数据作为输出数据
6118
过滤异常和/或生成警报
6120
结束

图 61

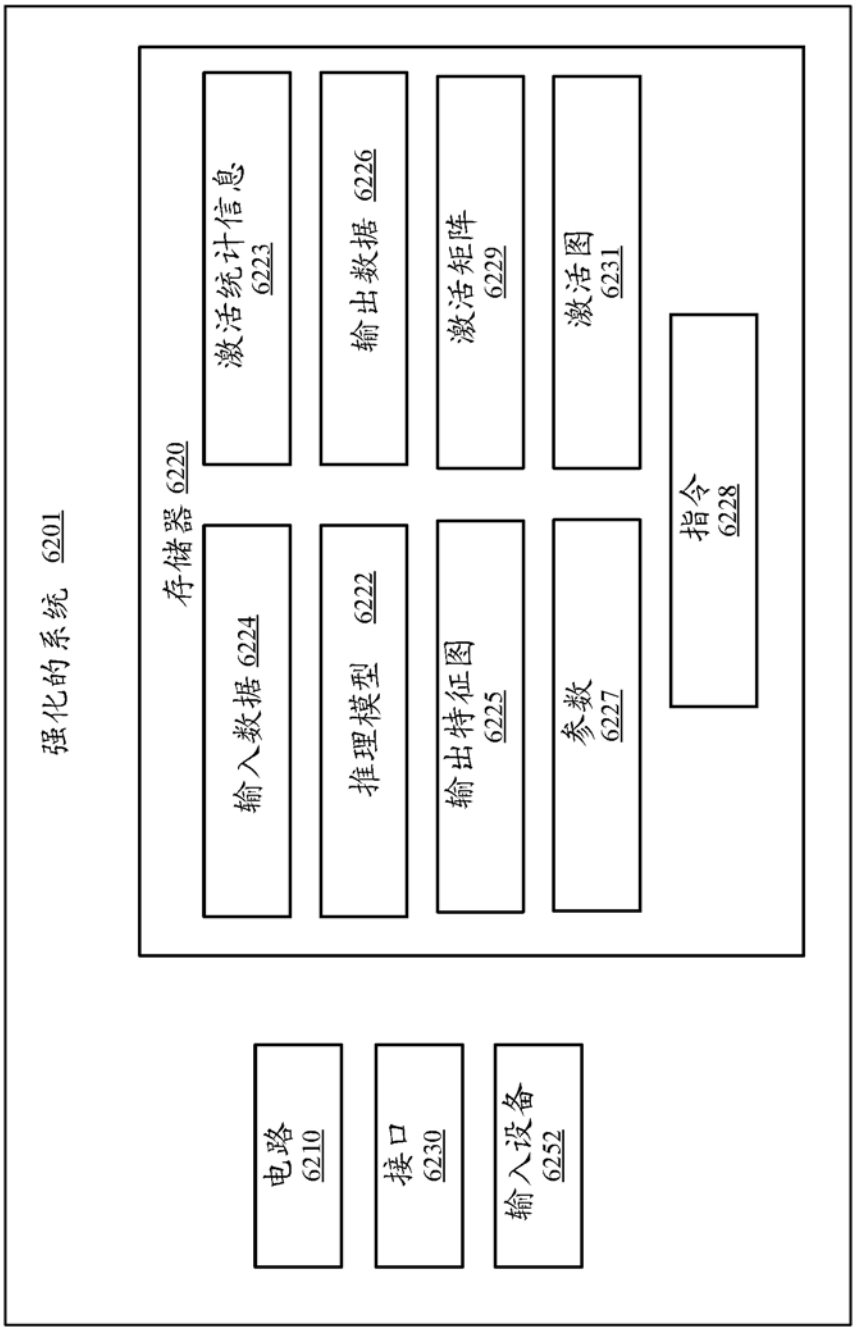

强化的系统 6201
存储器 6220
输入数据 6224
推理模型 6222
输出特征图 6225
参数 6227
指令 6228
激活统计信息 6223
输出数据 6226
激活矩阵 6229
激活图 6231
电路 6210
接口 6230
输入设备 6252

图 62

逻辑流程 6300

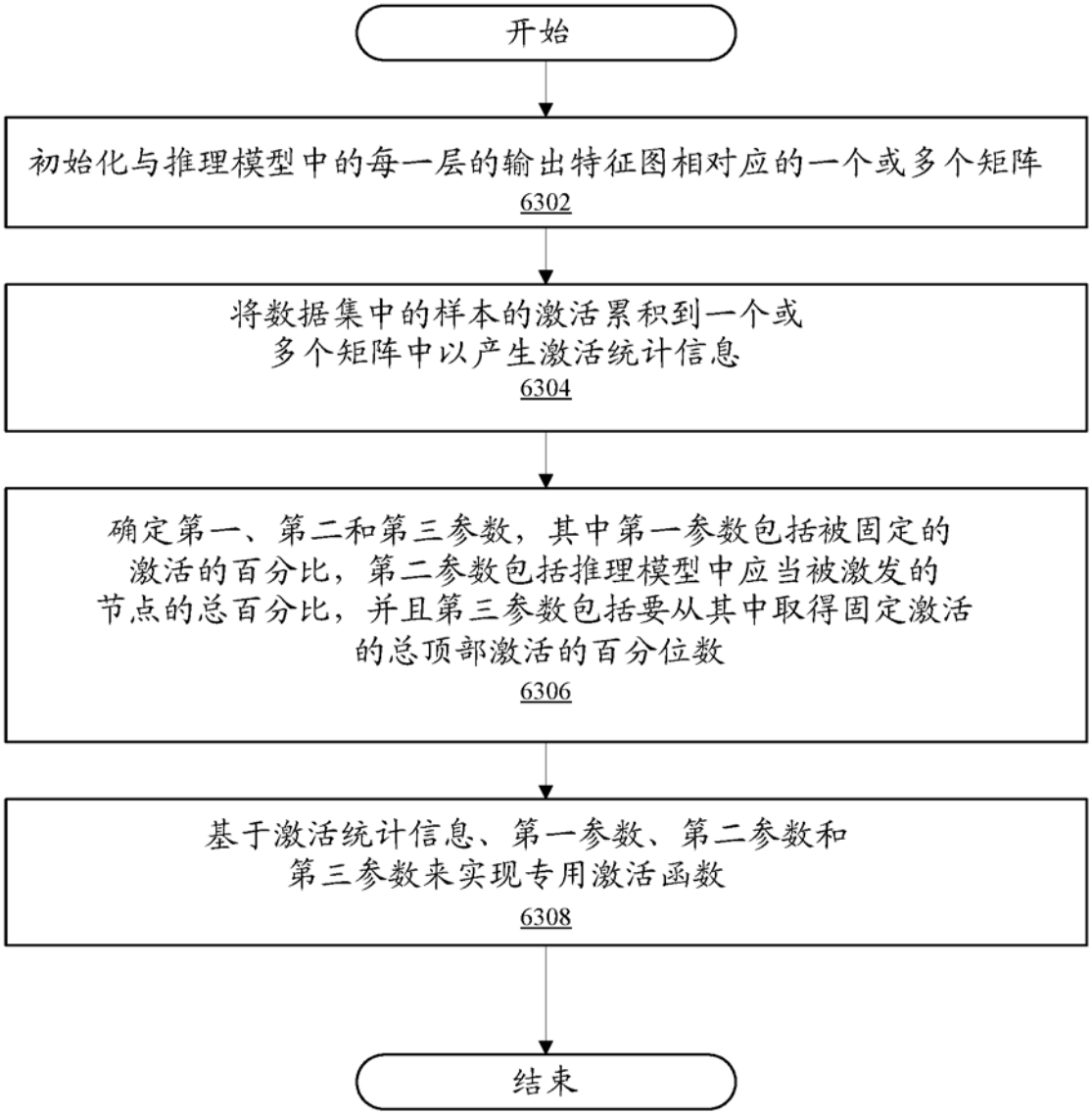



图 63

逻辑流程 6400

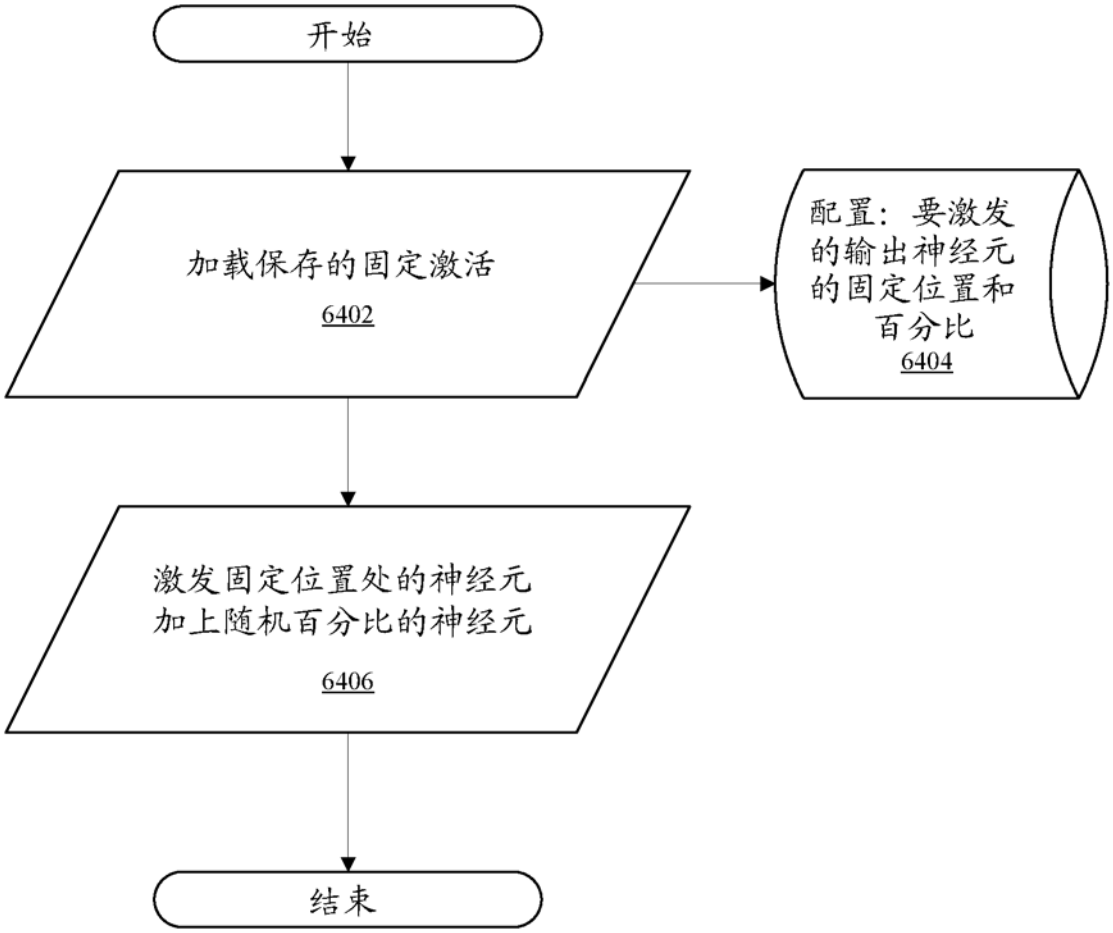



图 64

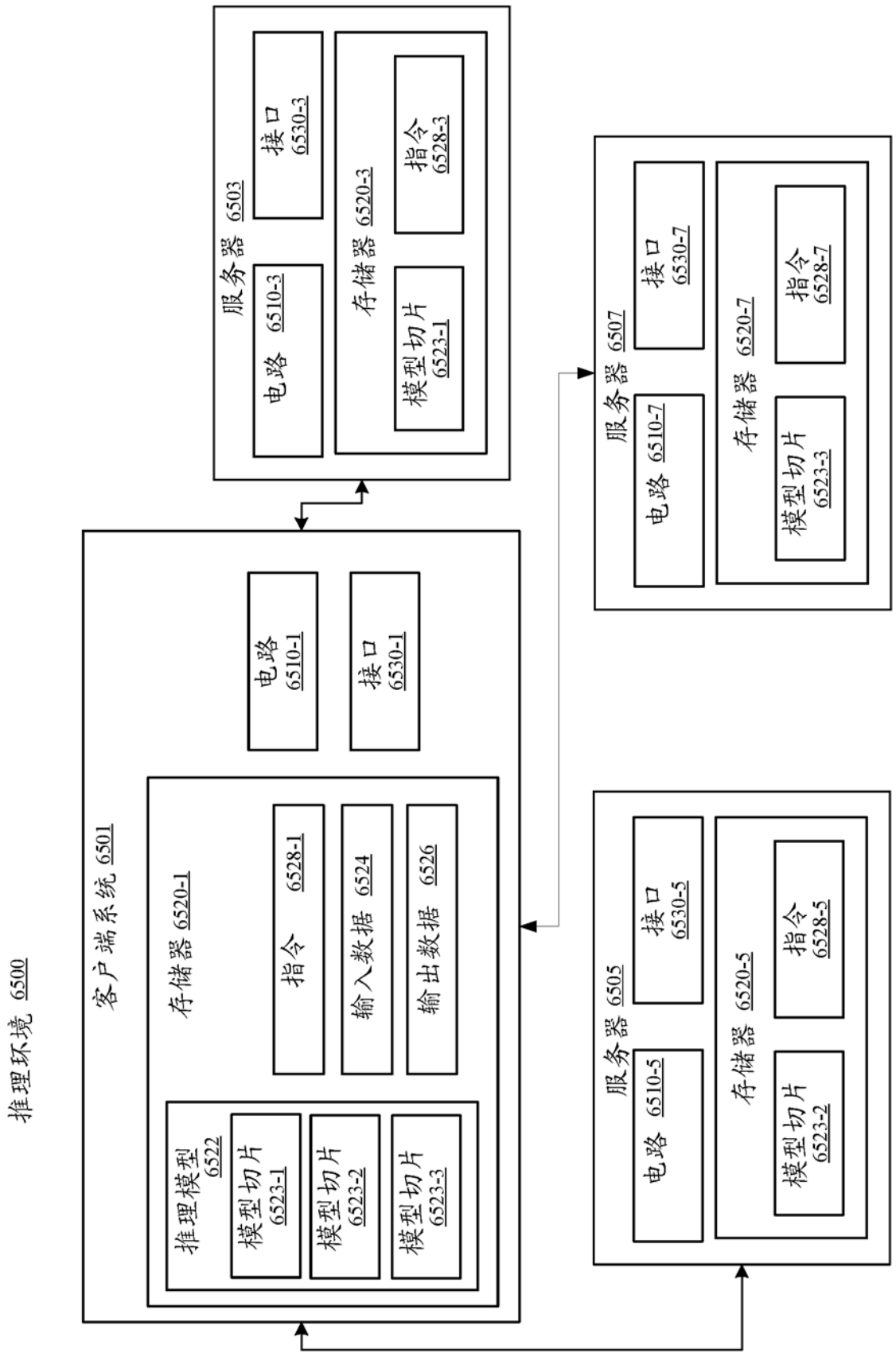

推理环境 6500
客户端系统 6501
存储器 6520-1
推理模型 6522
模型切片 6523-1
模型切片 6523-2
模型切片 6523-3
指令 6528-1
输入数据 6524
输出数据 6526
电路 6510-1
接口 6530-1
服务器 6503
电路 6510-3
接口 6530-3
存储器 6520-3
模型切片 6523-1
指令 6528-3
服务器 6507
电路 6510-7
接口 6530-7
存储器 6520-7
模型切片 6523-3
指令 6528-7
服务器 6505
电路 6510-5
接口 6530-5
存储器 6520-5
模型切片 6523-2
指令 6528-5

图 65

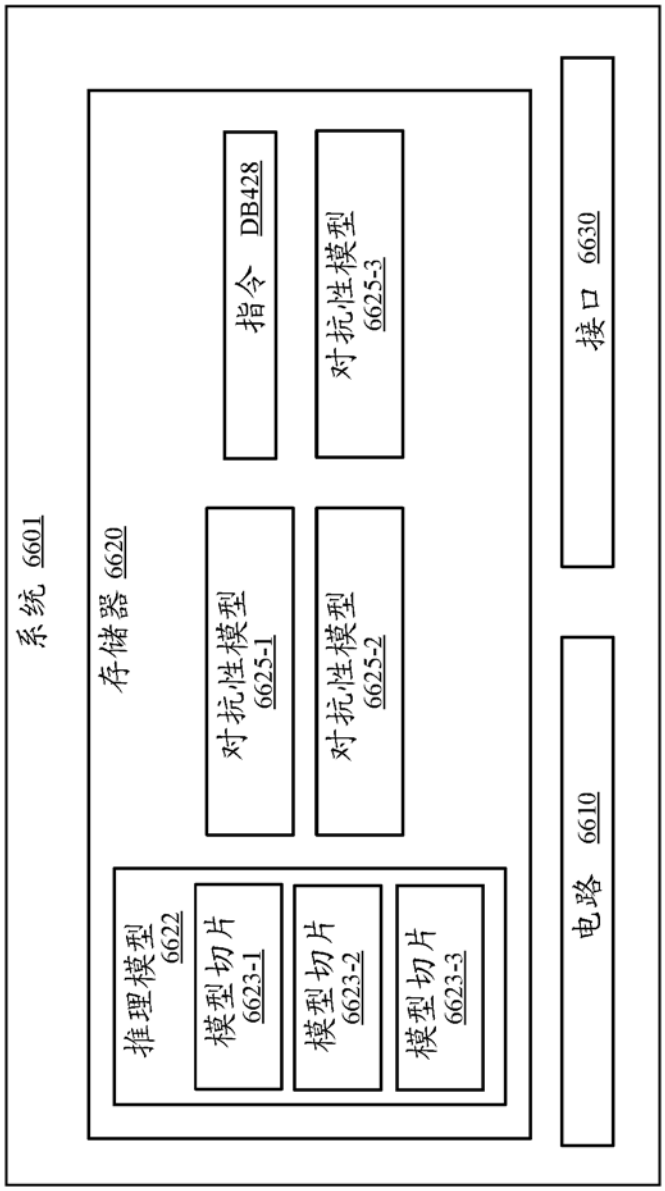

系统 6601
存储器 6620
推理模型 6622
模型切片 6623-1
模型切片 6623-2
模型切片 6623-3
对抗性模型 6625-1
对抗性模型 6625-2
指令 DB428
对抗性模型 6625-3
电路 6610
接口 6630

图 66

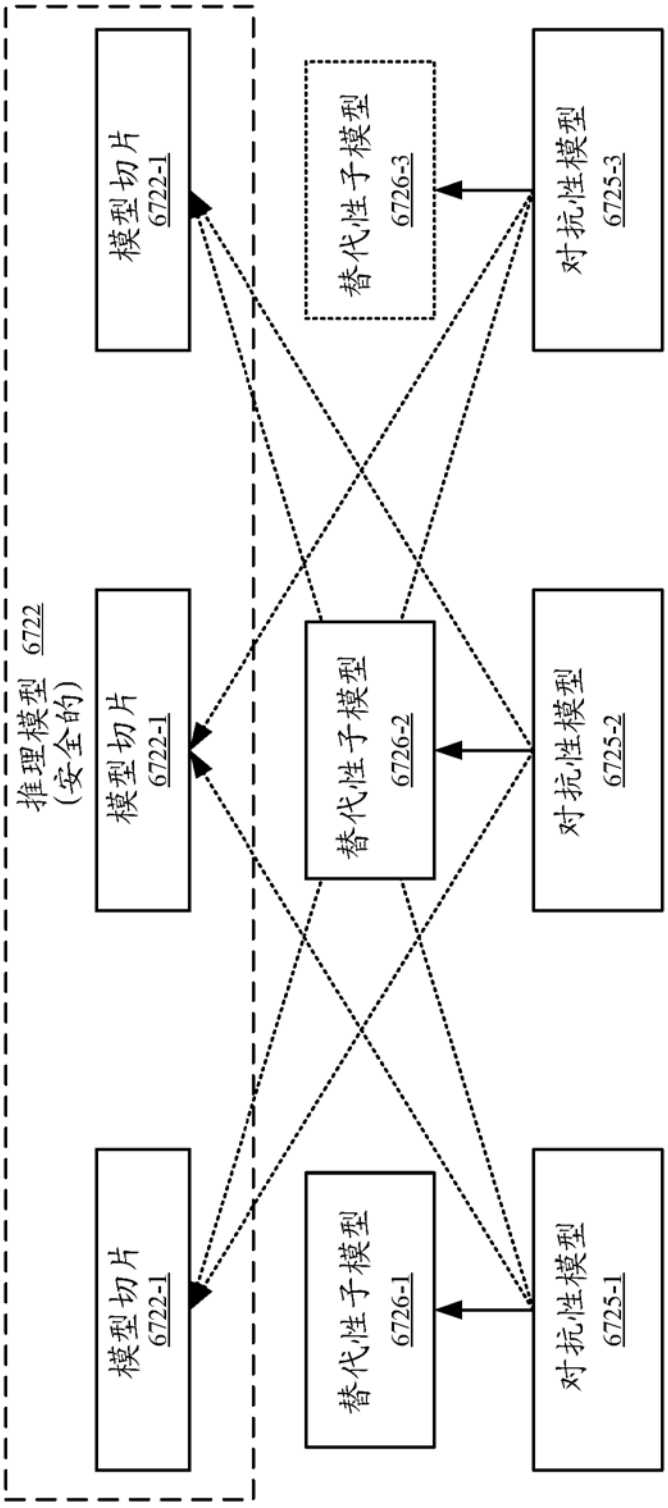

推理模型 6722
(安全的)
模型切片 6722-1
模型切片 6722-1
模型切片 6722-1
替代性子模型 6726-3
替代性子模型 6726-2
替代性子模型 6726-1
对抗性模型 6725-3
对抗性模型 6725-2
对抗性模型 6725-1

图 67

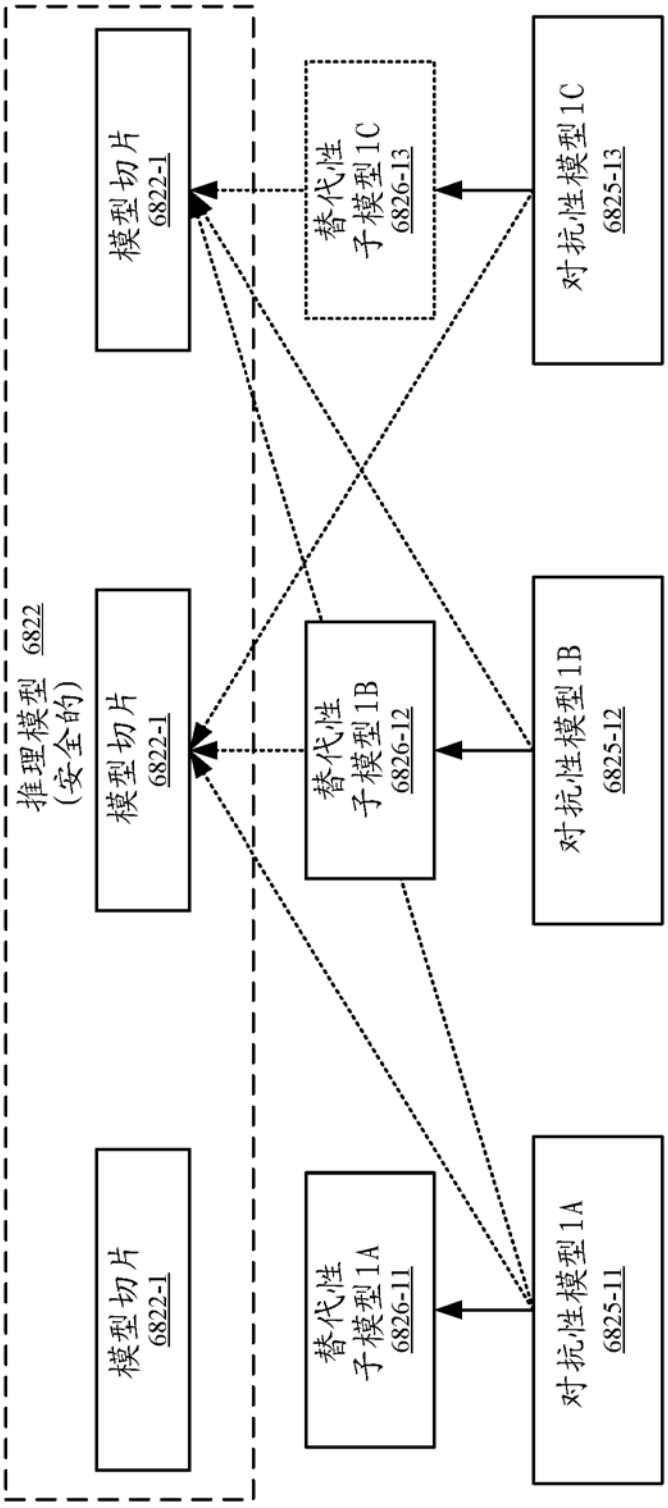

推理模型 6822
(安全的)
模型切片 6822-1
模型切片 6822-1
模型切片 6822-1
替代性子模型1C 6826-13
替代性子模型1B 6826-12
替代性子模型1A 6826-11
对抗性模型1C 6825-13
对抗性模型1B 6825-12
对抗性模型1A 6825-11

图 68A

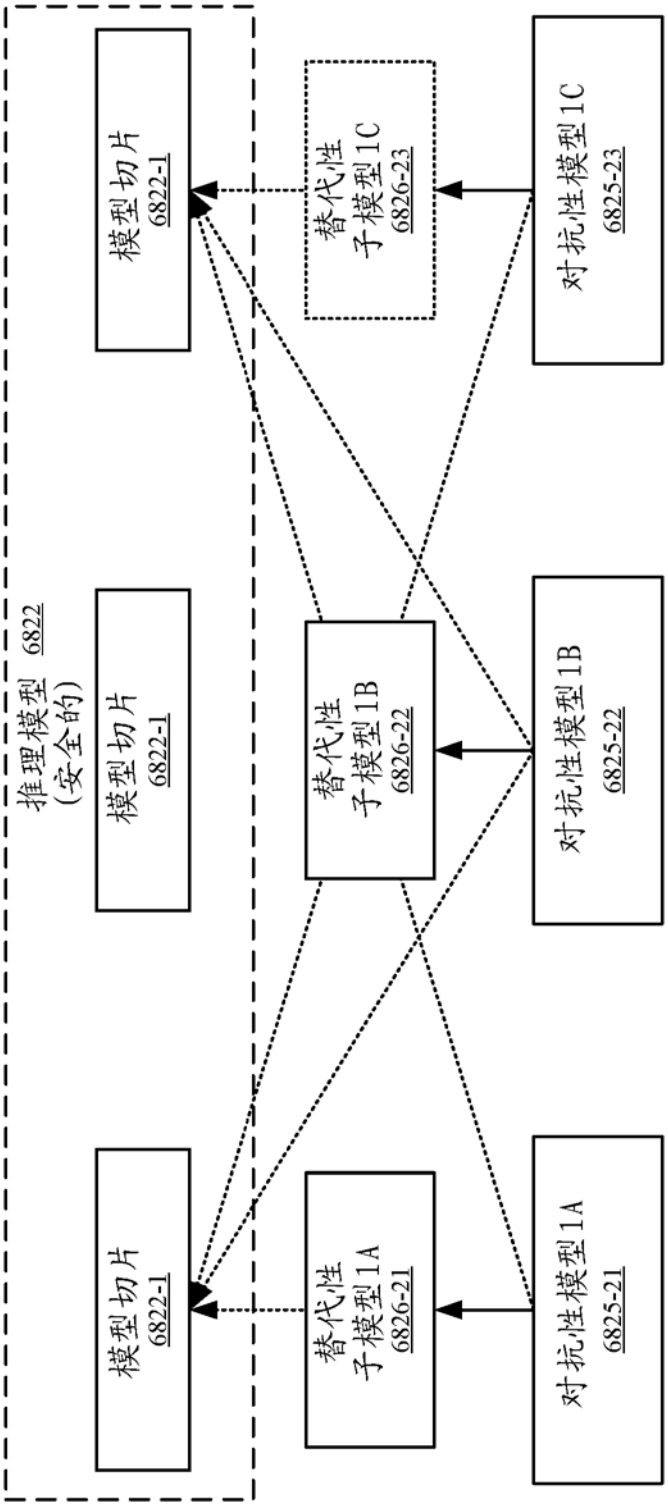

推理模型 6822
(安全的)
模型切片 6822-1
模型切片 6822-1
模型切片 6822-1
替代性子模型1C 6826-23
替代性子模型1B 6826-22
替代性子模型1A 6826-21
对抗性模型1C 6825-23
对抗性模型1B 6825-22
对抗性模型1A 6825-21

图 68B

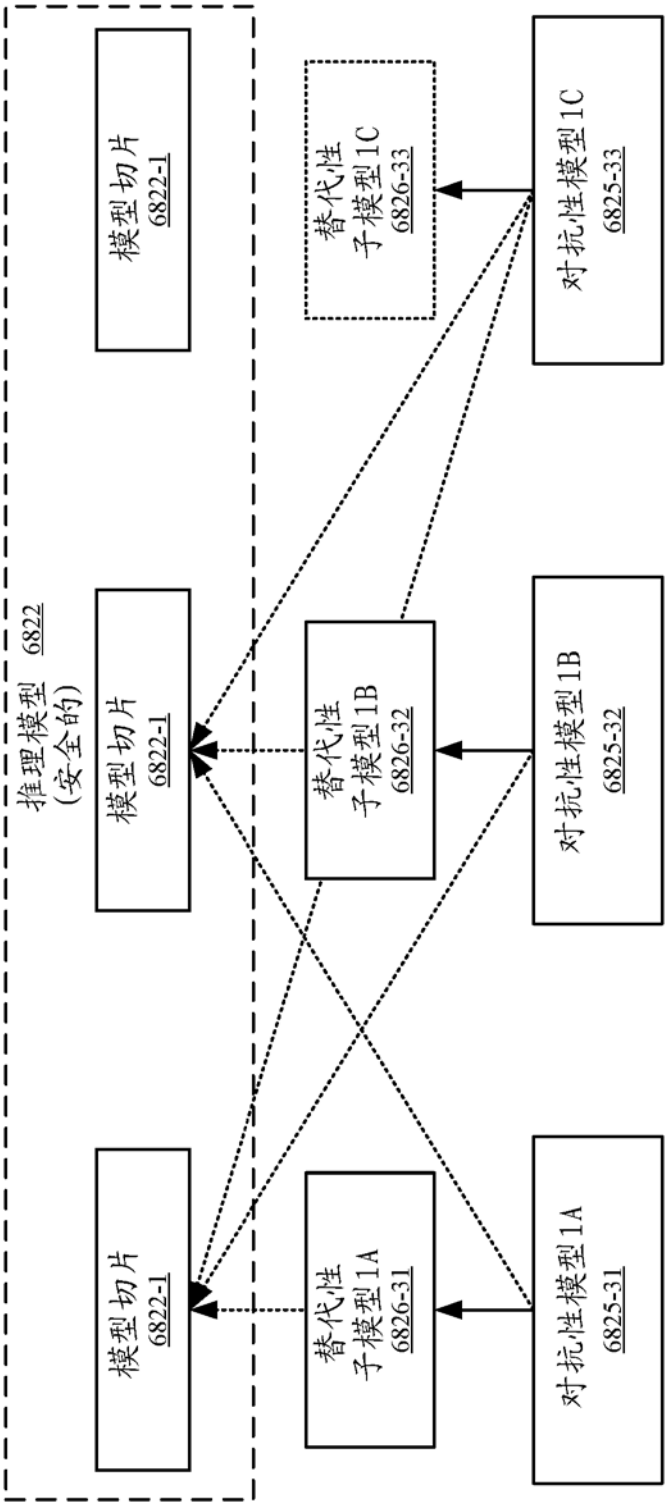

推理模型 6822
(安全的)
模型切片 6822-1
模型切片 6822-1
模型切片 6822-1
替代性子模型1C 6826-33
替代性子模型1B 6826-32
替代性子模型1A 6826-31
对抗性模型1C 6825-33
对抗性模型1B 6825-32
对抗性模型1A 6825-31

图 68C

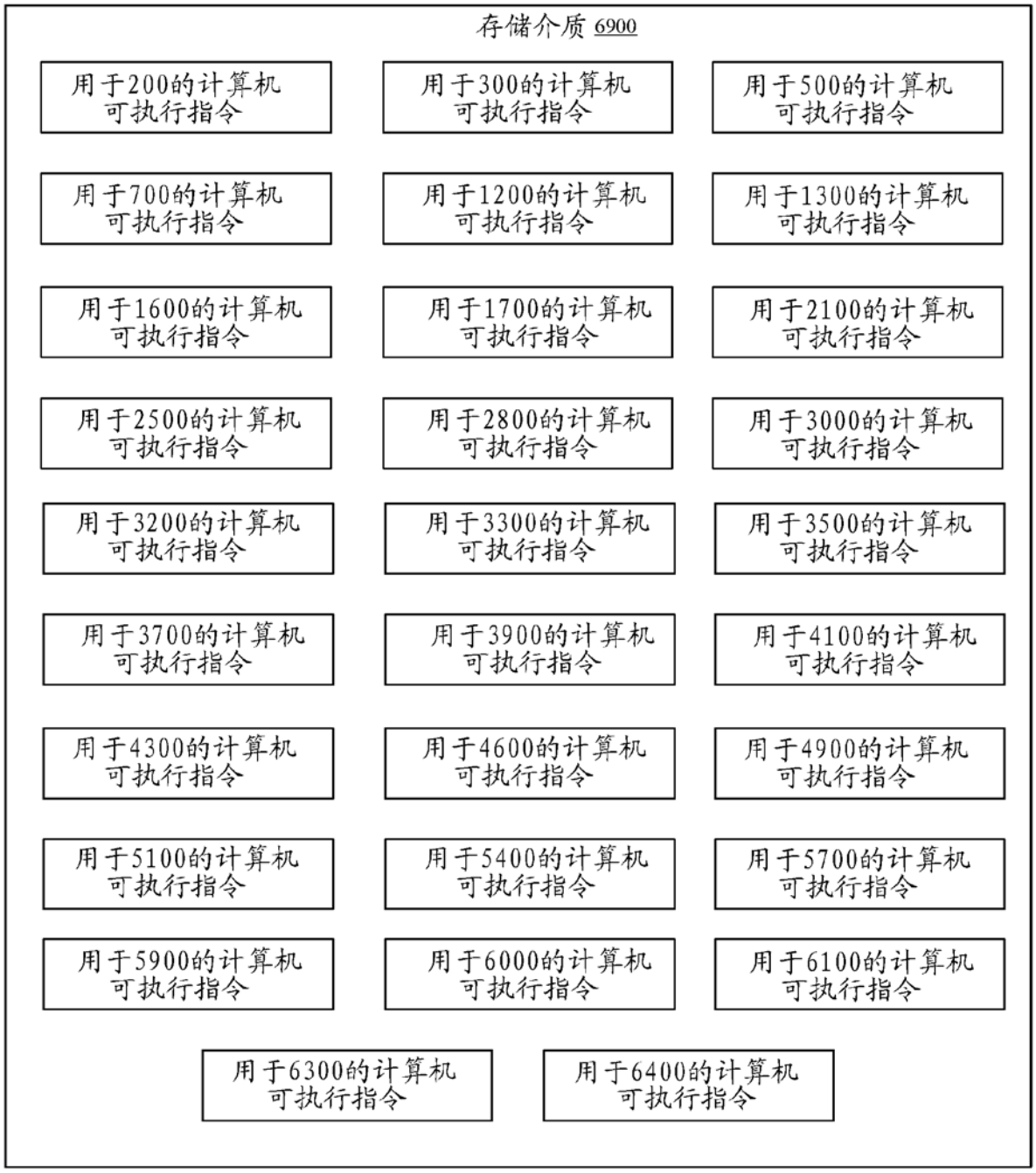

存储介质 6900
用于200的计算机可执行指令
用于300的计算机可执行指令
用于500的计算机可执行指令
用于700的计算机可执行指令
用于1200的计算机可执行指令
用于1300的计算机可执行指令
用于1600的计算机可执行指令
用于1700的计算机可执行指令
用于2100的计算机可执行指令
用于2500的计算机可执行指令
用于2800的计算机可执行指令
用于3000的计算机可执行指令
用于3200的计算机可执行指令
用于3300的计算机可执行指令
用于3500的计算机可执行指令
用于3700的计算机可执行指令
用于3900的计算机可执行指令
用于4100的计算机可执行指令
用于4300的计算机可执行指令
用于4600的计算机可执行指令
用于4900的计算机可执行指令
用于5100的计算机可执行指令
用于5400的计算机可执行指令
用于5700的计算机可执行指令
用于5900的计算机可执行指令
用于6000的计算机可执行指令
用于6100的计算机可执行指令
用于6300的计算机可执行指令
用于6400的计算机可执行指令

图 69

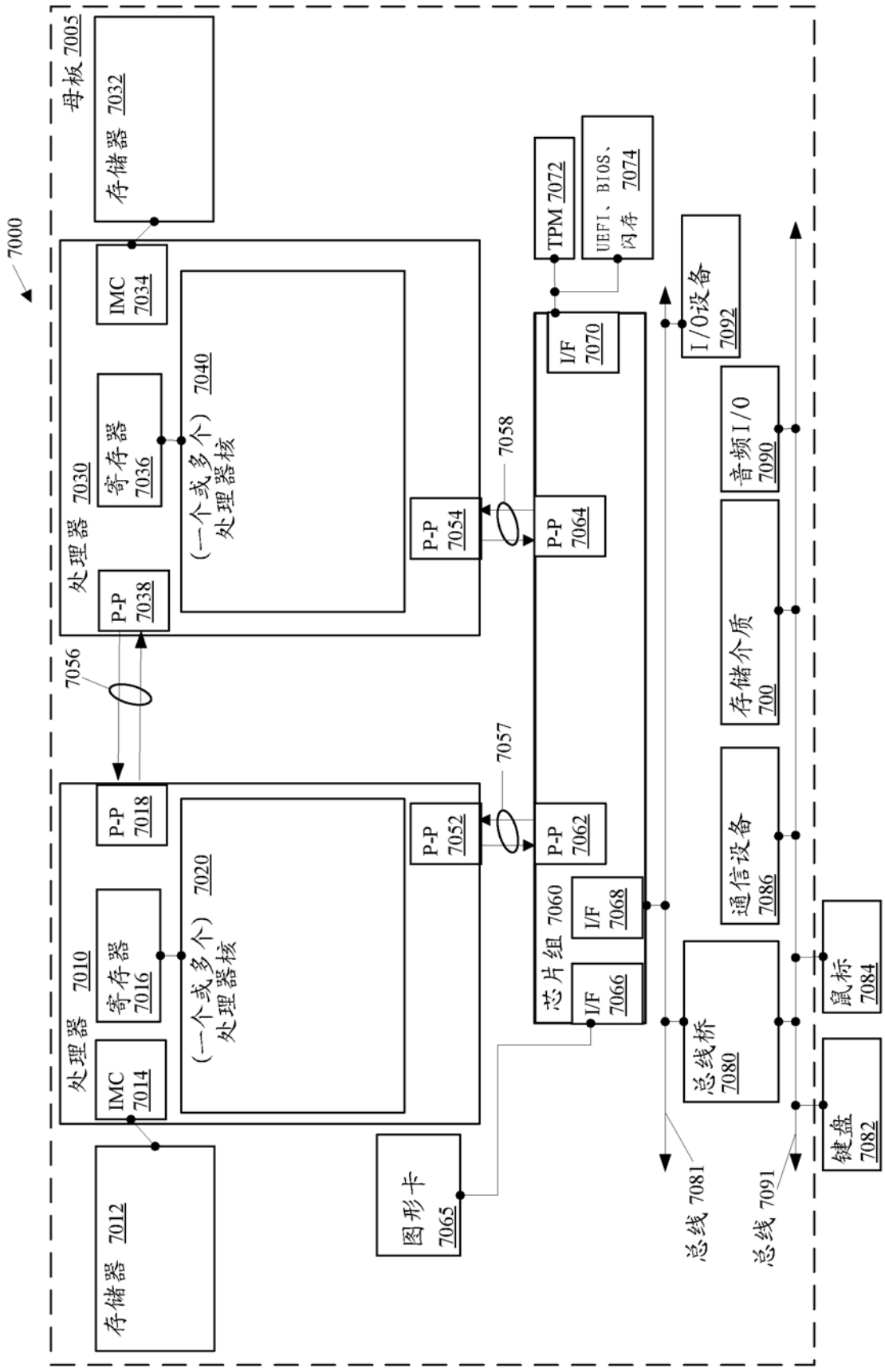

7000
母板 7005
存储器 7032
IMC 7034
处理器 7030
寄存器 7036
(一个或多个) 处理器核 7040
P-P 7038
P-P 7054
7058
P-P 7064
TPM 7072
UEFI、BIOS、闪存 7074
I/F 7070
I/O设备 7092
音频I/O 7090
存储介质 700
通信设备 7086
鼠标 7084
键盘 7082
总线桥 7080
总线 7081
总线 7091
芯片组 7060
I/F 7068
I/F 7066
P-P 7062
P-P 7052
7057
7056
P-P 7018
处理器 7010
寄存器 7016
(一个或多个) 处理器核 7020
IMC 7014
存储器 7012
图形卡 7065

图 70

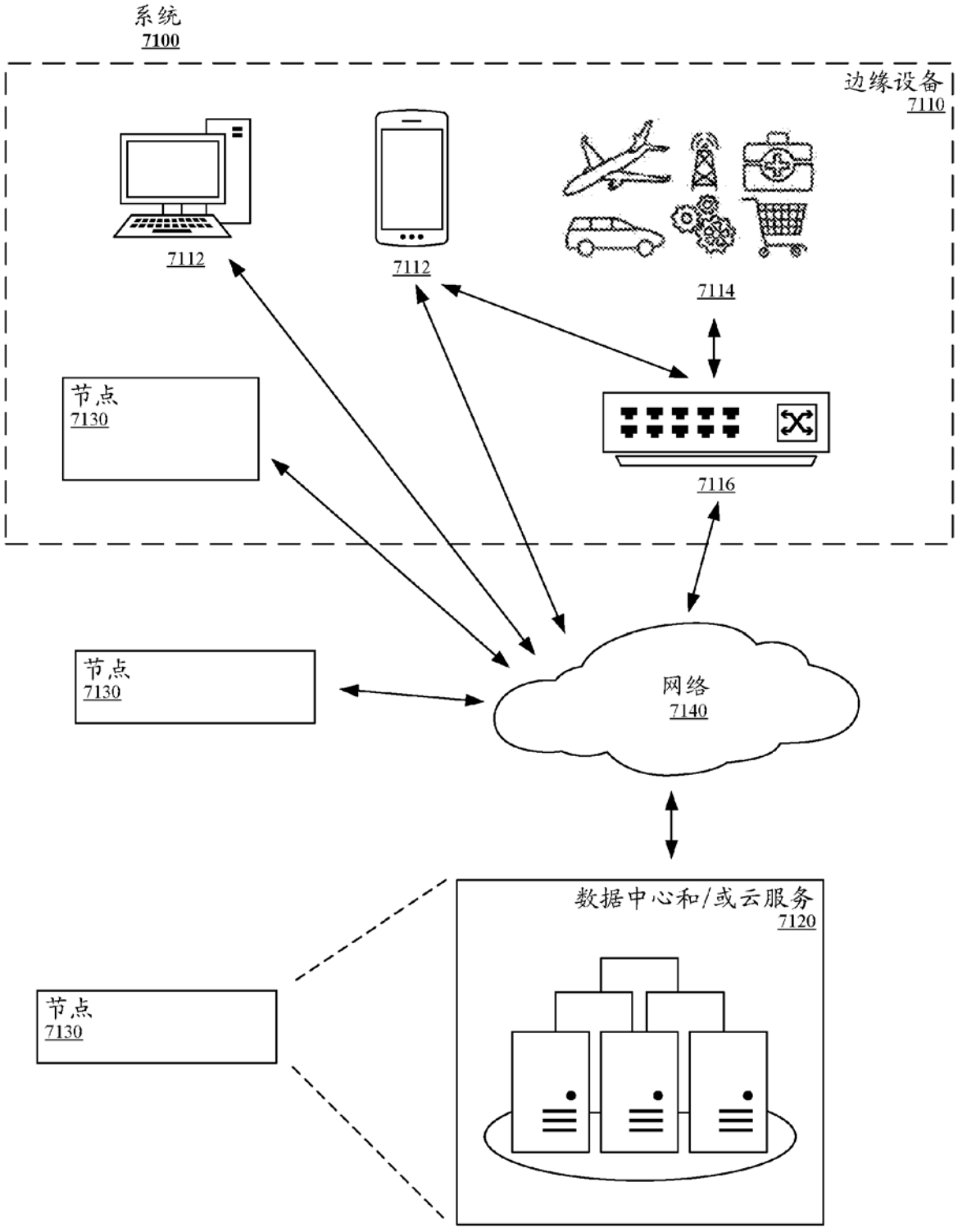

系统
7100
边缘设备
7110
7112
7112
7114
节点
7130
7116
节点
7130
网络
7140
数据中心和/或云服务
7120
节点
7130

图 71

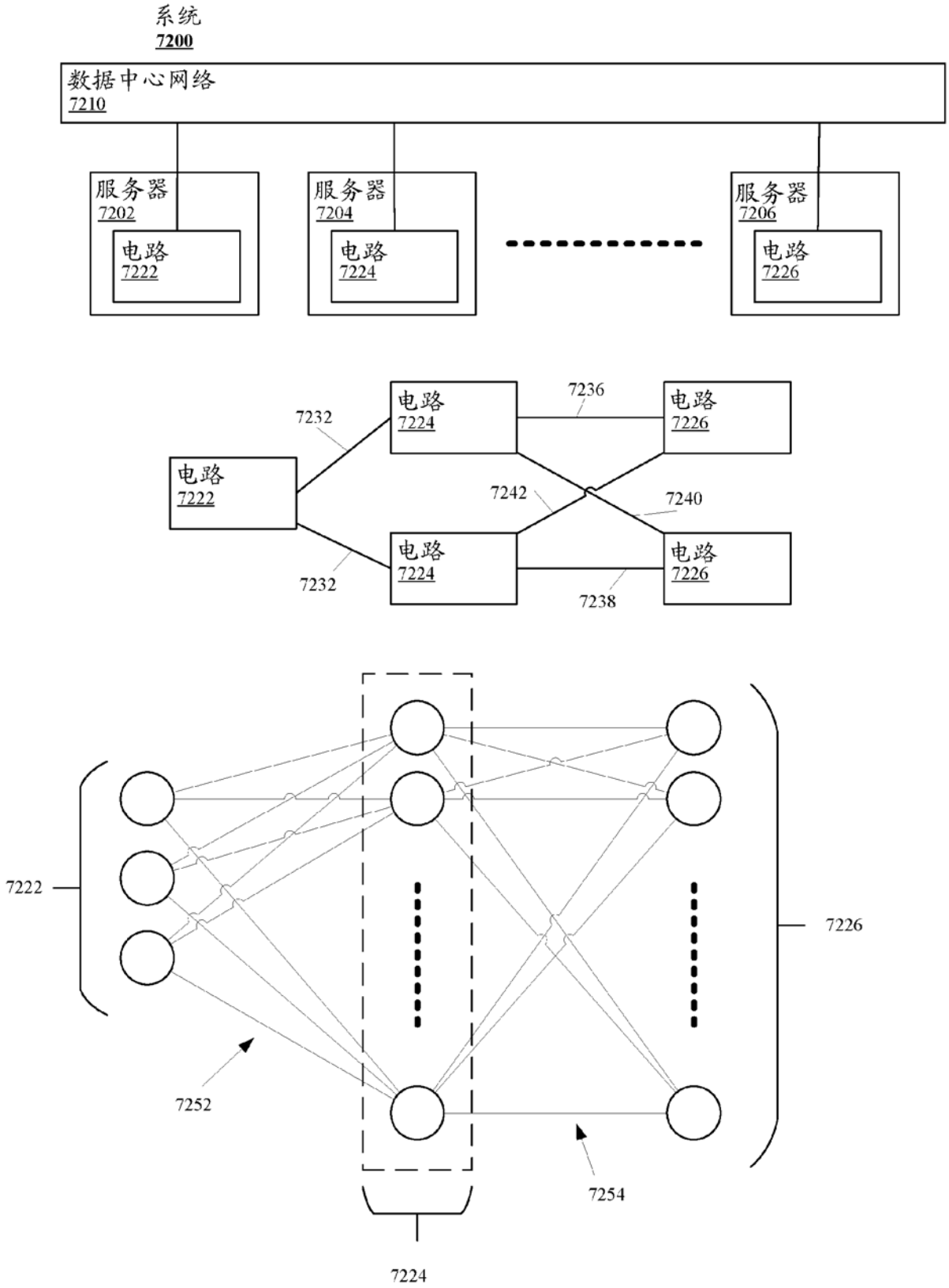

系统
7200
数据中心网络
7210
服务器
7202
电路
7222
服务器
7204
电路
7224
服务器
7206
电路
7226
电路
7222
7232
电路
7224
7236
电路
7226
7242
7240
7232
电路
7224
电路
7226
7238
7222
7226
7252
7254
7224

图 72

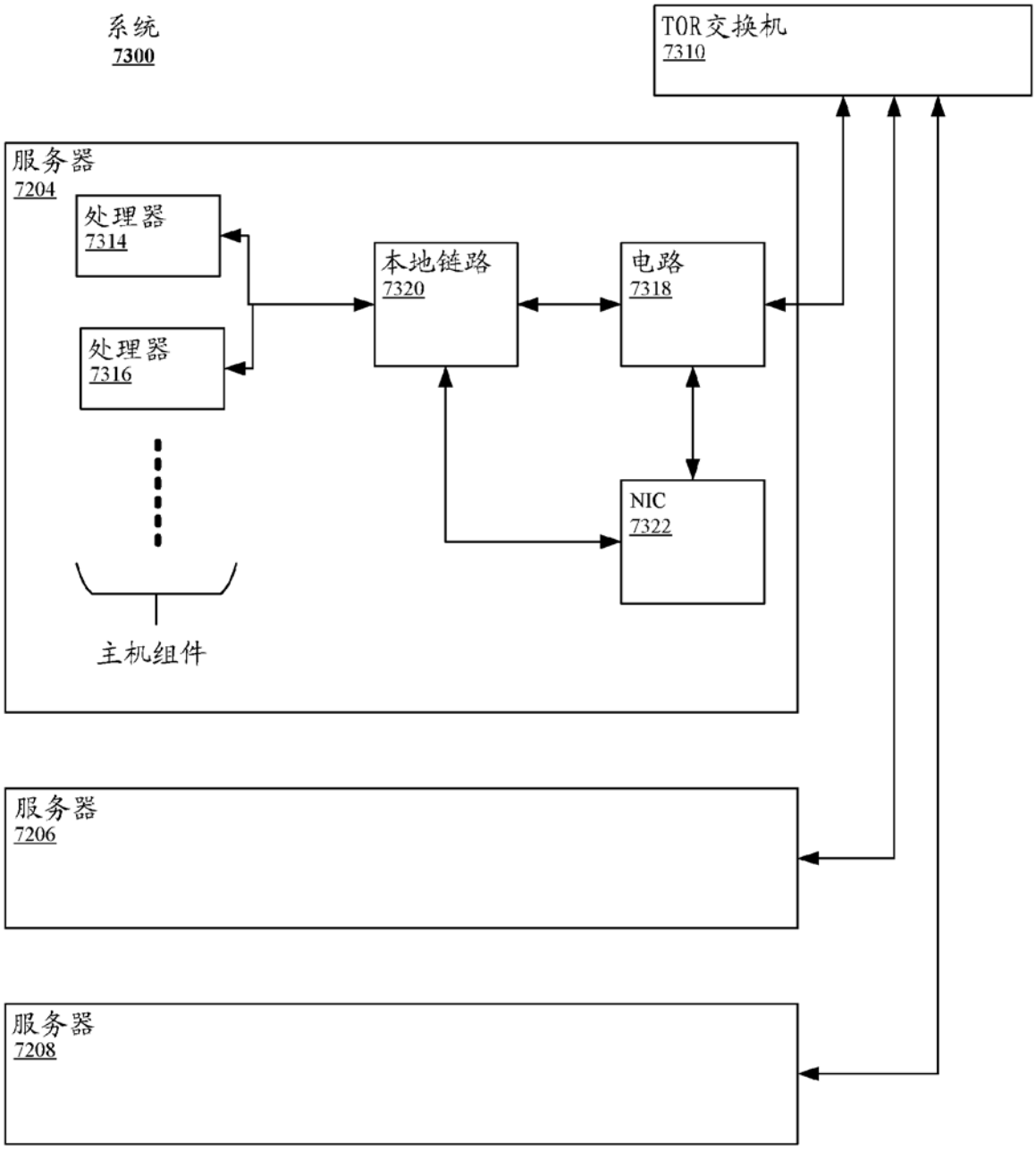

系统
7300
TOR交换机
7310
服务器
7204
处理器
7314
处理器
7316
本地链路
7320
电路
7318
NIC
7322
主机组件
服务器
7206
服务器
7208

图 73