(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115943390 A
(43)申请公布日 2023.04.07

(21)申请号 202180040202.7
(22)申请日 2021.05.05
(30)优先权数据
20305485.3 2020.05.12 EP
(85)PCT国际申请进入国家阶段日
2022.12.01
(86)PCT国际申请的申请数据
PCT/EP2021/061798 2021.05.05
(87)PCT国际申请的公布数据
WO2021/228641 EN 2021.11.18
(71)申请人 交互数字CE专利控股公司
地址 法国巴黎
(72)发明人 Q·K·N·董 T·菲洛奇 F·勒博尔泽 F·施尼茨勒 P·方丹
(74)专利代理机构 北京润平知识产权代理有限公司 11283
专利代理师 肖冰滨
(51)Int.Cl.
G06N 3/08(2006.01)
G06N 3/04(2006.01)
G06N 3/063(2006.01)
G06N 5/00(2006.01)
G06N 3/10(2006.01)

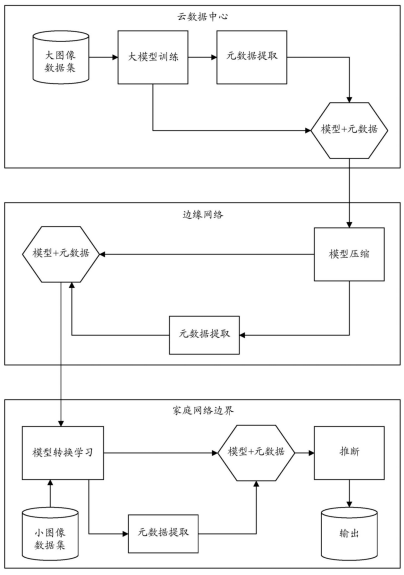
权利要求书2页 说明书16页 附图10页

(54)发明名称

用于训练和/或部署深度神经网络的系统和方法

(57)摘要

本公开涉及一种方法，该方法包括在训练第一深度神经网络时获得元数据，以及将获得的元数据嵌入在信号中。本公开涉及一种方法，该方法包括获得与第一深度神经网络的先前训练相关的元数据，以及使用获得的元数据来适配第二深度神经网络的模型。本公开还涉及对应的设备、计算机存储介质和信号。



CN 115943390 A

1.一种设备,所述设备包括至少一个处理器,所述至少一个处理器被配置用于:

-获得从第一深度神经网络的先前训练确定的至少一个元数据;

-使用所述获得的元数据来适配第二深度神经网络的模型。

2.一种方法,所述方法包括:

-从第一深度神经网络的先前训练获得至少一个元数据;

-使用所述获得的元数据来适配第二深度神经网络的模型。

3.根据权利要求1所述的设备或根据权利要求2所述的方法,其中所述元数据属于包括以下项的组:

-至少一个批大小,

-至少一个n-优化器,

-至少一个drop-out,

-至少一个学习率,

-至少一个损失函数的指定和/或参数,

-与训练的精度相关的至少一个性能指标;

-与所述DNN的至少一层的至少一个权重的重要性相关的至少一个指标;

-关于在所述训练期间使用的训练集的至少一个元素上执行的一个或多个预处理的至少一个信息

-与所述第一DNN的一个或多个操作模式相关的至少一个信息

-表示在所述第一DNN内能够进行预测的至少一个位置的至少一个信息;

-上述元数据中的至少两者的组合。

4.根据权利要求1或3所述的设备或根据权利要求2或3所述的方法,其中适配所述第二深度神经网络的所述模型包括使用所述第一元数据来确定所述模型的权重的子集。

5.根据权利要求1或3所述的设备或根据权利要求2或3所述的方法,其中适配所述第二深度神经网络的所述模型包括使用所述第一元数据来减少所述模型的权重的数量。

6.根据权利要求1或3所述的设备或根据权利要求2或3所述的方法,其中所述适配包括使用所述第一元数据丢弃所述模型的至少一部分。

7.根据权利要求1或3-6中任一项所述的设备或根据权利要求2-6中任一项所述的方法,其中所述适配包括使用所述第一元数据来编码所述模型的至少一部分。

8.根据权利要求1或3-6中任一项所述的设备或根据权利要求2-6中任一项所述的方法,其中所述适配包括使用所述第一元数据来微调所述模型。

9.根据权利要求8所述的设备或方法,其中所述适配包括预处理在所述微调期间使用的训练数据集的至少一部分。

10.根据权利要求1或3-9中任一项所述的设备或根据权利要求2-9中任一项所述的方法,其中所述获得包括解码所述元数据。

11.根据权利要求10所述的设备或方法,其中所述元数据从信号解码,所述元数据作为通过所述第一深度神经网络的所述训练输出的所述第一深度神经网络模型的参数的边信息。

12.根据权利要求10所述的设备或方法,其中所述元数据从信号解码,所述信号是除了嵌入由所述第一深度神经网络的所述训练输出的所述第一深度神经网络模型的参数的信

号之外的信号。

13.根据权利要求1或3-12中任一项所述的设备或根据权利要求2-12中任一项所述的方法,其中所述第二深度神经网络是所述已训练的第一深度神经网络,并且其中所述第一元数据和所述模型分开获得。

14.根据权利要求1或3至13中任一项所述的设备,所述至少一个处理器被适配用于,或者根据权利要求2至13中任一项所述的方法包括:在用户界面上呈现表示所述元数据中的一个或多个元数据的至少一个信息。

15.一种非暂态计算机可读存储介质,所述非暂态计算机可读存储介质承载软件程序,所述软件程序包括程序代码指令,所述程序代码指令用于当所述软件程序由计算机执行时,执行根据权利要求2至14中任一项所述的方法。

用于训练和/或部署深度神经网络的系统和方法

[0001] 引言

[0002] 本公开的一个或多个实施方案的技术领域涉及深度学习技术的使用，如深度神经网络(DNN)的使用。由于深度学习技术可以在许多技术领域中使用，因此本公开的实施方案可以在许多技术领域中实现，例如在多媒体处理技术领域中，例如在图像处理、视频处理和/或音频处理中。

[0003] 描述

[0004] 根据第一方面，本原理使得能够通过提出一种用于训练深度神经网络的方法来解决至少一些缺点。

[0005] 至少一些实施方案涉及存储/编码/发送/解码与深度神经网络的训练相关的元数据。元数据可以作为边信息与DNN相关联，或者可以分别被存储/编码/发送/解码到它们所涉及的模型。

[0006] 至少一些实施方案涉及编码/解码与深度神经网络的训练和/或通过训练输出的DNN模型的参数相关的元数据。

[0007] 本公开的至少一些实施方案提出一种用于编码深度神经网络的方法。

[0008] 本公开的至少一些实施方案提出一种用于解码深度神经网络的方法。

[0009] 本公开的至少一些实施方案涉及一种设备，该设备包括至少一个处理器，该至少一个处理器被配置用于获得从第一深度神经网络(DNN)的先前训练确定的至少一个元数据；以及使用所述获得的元数据适配第二深度神经网络的模型。

[0010] 本公开的至少一些实施方案涉及一种方法，该方法包括：从第一深度神经网络的先前训练获得至少一个元数据，以及使用所述获得的元数据来适配第二深度神经网络的模型。

[0011] 根据本公开的至少一些实施方案，元数据属于包括以下项的组：

[0012] -至少一个批大小，

[0013] -至少一个优化器，

[0014] -至少一个drop-out设置，

[0015] -至少一个学习率设置，

[0016] -至少一个损失函数的指定和/或参数，

[0017] -与训练的精度相关的至少一个性能指标；

[0018] -与DNN的至少一层的至少一个权重或至少一个过滤器的重要性相关的至少一个指标；

[0019] -关于在所述训练期间使用的训练集的至少一个元素上执行的一个或多个预处理的至少一个信息

[0020] -与DNN模型的一个或多个操作模式相关的至少一个信息

[0021] -表示在DNN模型内能够进行预测的至少一个位置的至少一个信息；

[0022] -上述元数据中的至少两者的组合。

[0023] 根据本公开的至少一些实施方案，所述适配包括使用所述第一元数据来微调所述

模型。

[0024] 根据本公开的至少一些实施方案,所述适配包括预处理在所述微调期间使用的训练数据集的至少一部分。

[0025] 根据本公开的至少一些实施方案,所述适配包括使用所述第一元数据丢弃所述模型的至少一部分。

[0026] 根据本公开的至少一些实施方案,所述适配可以包括通过进一步训练和/或模型配置来选择用于推断的所需/最佳操作点的模型适配。

[0027] 根据本公开的至少一些实施方案,所述第二DNN是所述已训练的第一DNN,并且所述第一元数据和所述模型分开获得。

[0028] 根据本公开的至少一些实施方案,所述第二DNN是所述已训练的第一DNN,并且其中所述第一元数据作为所述模型的边信息被获得。

[0029] 根据本公开的至少一些实施方案,至少一个处理器被适配用于或者方法包括在用户界面上呈现表示所述元数据中的一个或多个元数据的至少一个信息。

[0030] 虽然未明确描述,但上述设备可被适配为在其实施方案中的任一个实施方案中执行本公开的上述方法。

[0031] 根据本公开的至少一些实施方案,所述适配包括使用所述第一元数据加载所述模型的子集。例如,所述子集模型可以与计算能力、芯片组架构、省电约束或延迟约束相关。

[0032] 例如,本公开的方法的至少一个实施方案涉及一种用于解码与第一DNN的先前训练和/或和训练第二DNN深度神经网络(例如微调第二深度神经网络的参数)相关的元数据的方法。

[0033] 通常,在编码过程中,对数据进行熵编码以获得压缩数据。为了重构数据,通过对应于熵编码和量化的逆过程来解码压缩的量化数据。

[0034] 根据另一方面,提供了一种装置。该装置包括处理器。该处理器可被配置为训练、编码和/或解码与深度神经网络相关的元数据和/或通过执行前述方法中的任一方法来训练、编码和/或解码深度神经网络。

[0035] 根据至少一个实施方案的另一一般方面,提供了一种设备,该设备包括:根据解码实施方案中的任一实施方案的装置;以及以下项中的至少一者:(i)天线,该天线被配置为接收信号,(ii)频带限制器,该频带限制器被配置为将接收到的信号限制到包括该信号的部分的频带,或(iii)显示器,该显示器被配置为显示表示该信号的一部分的输出。

[0036] 虽然未明确描述,但与方法或对应设备相关的本实施方案可以任何组合或子组合来使用。

[0037] 根据另一方面,本公开涉及一种非暂态计算机可读程序产品,该非暂态计算机可读程序产品包括程序代码指令,该程序代码指令用于当非暂态计算机可读程序产品由计算机执行时在其实施方案中的任一实施方案中执行本公开的方法中的至少一个方法。

[0038] 根据另一方面,本公开涉及一种非暂态计算机可读存储介质,该非暂态计算机可读存储介质承载软件程序,该软件程序包括程序代码指令,该程序代码指令用于当软件程序由计算机执行时,在其实施方案中的任一实施方案中执行本公开的方法中的至少一个方法。根据至少一个实施方案的另一一般方面,提供了一种非暂态计算机可读介质,该非暂态计算机可读介质包括根据所描述的编码实施方案或变型中的任一者生成的数据内容。

[0039] 根据至少一个实施方案的另一一般方面，提供了一种信号，该信号包括根据所描述的编码实施方案或变型中的任一者生成的数据。

[0040] 根据至少一个实施方案的另一一般方面，比特流被格式化以包括根据所描述的编码实施方案或变型中的任一者生成的数据内容。

[0041] 根据至少一个实施方案的另一一般方面，提供了一种计算机程序产品，该计算机程序产品包括指令，当程序由计算机执行时该指令使得计算机执行所描述的解码实施方案或变型中的任一者。

附图说明

[0042] ●图1示出了通用的标准编码方案。

[0043] ●图2示出了通用的标准解码方案。

[0044] ●图3示出了可实现所描述的实施方案的典型处理器布置；

[0045] ●图4示出了根据本公开的至少一些实施方案的DNN编码方案；

[0046] ●图5示出了根据本公开的至少一些实施方案的DNN解码方案；

[0047] ●图6A至图6C示出了本公开所提出的架构的一些示例性工作流程；

[0048] ●图7A和图7B示出了用于获得和/或使用元数据的本公开的方法的一些示例性实施方案；

[0049] ●图8示出了在DNN的开发和部署的不同层级处实现的元数据的示例性使用和提取。

[0050] 应当注意，附图例示了示例性实施方案，并且本公开的实施方案不限于所例示的实施方案。

具体实施方式

[0051] 深度神经网络(DNN)已经在各种领域(诸如多媒体处理、计算机视觉、语音识别、自然语言处理等)中表现出先进的性能。然而，这种性能是以巨大的计算成本为代价的，因为DNN往往有大量的参数，通常达到数百万，有时甚至数十亿。

[0052] 为了获得良好的机器学习模型(更具体地，深度神经网络(DNN)模型)，研究人员和工程师花费大量时间来实验不同的模型架构、参数(如“超参数”)设置、预处理方法等。“超参数”在本文中应被理解为除了描述DNN本身(如DNN的层的权重(或神经元))的参数之外的、但与DNN被训练、微调或验证(推断)的方式相关的参数。

[0053] 深度神经网络模型或者使用一个或多个深度神经网络模型的设备或应用的用户通常不知道模型是如何被训练的。因此，训练DNN模型和/或已训练的DNN的部署通常不能依赖于在此类模型的训练阶段可获得的并且在稍后阶段(如再训练或微调)有用的信息和/或知识。

[0054] 本公开的至少一些实施方案提出保留与模型的训练相关的知识的至少一部分，如训练参数和应用度量。例如，在训练期间，可以观察到关于模型的各种有用知识，诸如损失函数的演变、精度、梯度变化等。此类知识可以保存在表示此类知识的模拟和/或数字信息中或从该模拟和/或数字信息中检索，该模拟和/或数字信息可以记录在存储介质上和/或嵌入在信号中。表示信息(本文中也称为元数据)可保存在计算机可读介质上和/或被编码

和/或发送以促进至少一个基于DNN的模型的部署。例如，一些表示信息可有助于针对新背景和/或新应用对DNN模型进行微调、变换、适配和/或个性化(例如，以使DNN适应于本地环境(如，设备的操作系统和/或处理能力)，或借助于个人、私人数据(如，家庭成员的图像)通过微调来对DNN进行个性化)。

[0055] 应当指出，根据实施方案，可以与DNN的参数相关联地(例如作为边信息)或者分开地(例如为了使用表示第一DNN的架构或训练阶段的一些信息来设计和/或训练第二DNN)存储/编码/发送/解码表示信息。

[0056] 因此，本公开的至少一些实施方案可有助于之后(如当模型被部署在执行训练的设备之外的设备中时)利用那些表示信息。

[0057] 根据实施方案，表示信息可在不同阶段被不同地使用，如下文详述的示例性实施方案所示。

[0058] 在图6A至图6C中示出根据本公开的至少一个实施方案的示例性机器学习系统的高级工作流程(训练和部署)。此类工作流程可允许在稍后阶段考虑在训练阶段(如图6A至图6C所示)期间使用和/或收集的信息。

[0059] 更准确地说，根据图6A和图6C的系统的示例性实施方案，元数据(框6500)可从在训练DNN模型时(在模型的训练之前和/或期间和/或从训练的输出)生成和/或收集的信息获得(例如提取)和/或可选地被编码(如框6600)，以供以后使用。

[0060] 如图6B所示，元数据可用于部署阶段，可选地在被解码(如框6700)之后。例如，它们可被用于模型部署阶段期间的模型适配(如框6300)。

[0061] 下面结合图6A至图6C以及图7A和图7B给出关于工作流程中每个框的更多细节。

[0062] 图6A特别示出了示例性用例中第一DNN的训练和编码。图7A结合图6A的第一DNN示出了可在电子设备(如云设备或编码设备)中实现的本公开的一些实施方案的至少一种方法。如图所示，图7A的方法可包括获得710第一DNN模型(即，从中执行训练(如图6A的框6100所示)的DNN架构)。获得第一模型可包括设计第一DNN模型或重新使用已设计(或部分设计)的第一DNN模型。架构例如可以从“现成”模型或随机地初始化(例如初始化权重值)。

[0063] 图7A的方法还可以包括训练720获得的第一模型(如图6A的框6100所示)。训练可以例如涉及使用存储在至少一个第一文件(如图6A的框610的元素DB1所示的数据库)中的一个或多个元素。这些元素可以是音频、图像、视频、文本、时间序列等形式。

[0064] 根据实施方案，可以不同地执行训练。例如，其可取决于在其操作期间(即，在推断期间)待指派给DNN的一个任务或多个任务。

[0065] 第一模型的训练产生预训练的第一模型(图6A的框6200)和与第一DNN模型的训练相关的多个元数据。元数据可以在第一模型的训练期间已经获得(例如从第一DNN模型生成和/或提取)并且被存储，或者可以在较早阶段期间(如在第一模型的设计期间)已经获得，或者可以已经通过另一(“第二”)DNN模型的先前训练(例如使用本公开的图7A所示的方法)输出。可选地，如图6A的框6600所示，可以对获得的元数据进行编码。元数据可被进一步处理(730,740)，例如它们可被发送和/或保存以供将来使用。如图7A所示，元数据可作为边信息(如图6A和图6B所示)与预训练的第一模型相关联以供进一步处理730，或可与预训练的第一模型分开保存以供进一步处理740(例如用于训练另一DNN模型)。

[0066] 因此，在一些实施方案中，该方法可以包括获得750另一模型的元数据，如训练元

数据。要指出的是，由于DNN的训练可以执行若干次，所以“其他”模型可以是相同DNN的先前版本，换句话说，是通过更新(由于先前训练)相同DNN的模型而获得的模型。在此类实施方案中，由“第一”模型的训练输出的第一元数据可以包括与先前训练相关的元数据的至少一部分。

[0067] 根据实施方案，可以获得不同种类的元数据。元数据的示例包括：

[0068] -与在至少一个DNN模型的训练期间使用的一个或多个参数(在本文中也称为“超参数”)的设置相关的指示，如一个或多个批大小(换句话说，用于计算迭代中的梯度的样本数目)、与优化器设置相关的一个或多个参数、drop-out设置、学习率设置、一个或多个损失函数、一个或多个评估度量等。

[0069] 例如，对于灵活的递归神经网络(RNN)训练，可以使用不同的损失函数条件来训练网络。作为示例，在第一步骤中，可以使用更宽松的损失函数条件来收敛到模型，然后可以利用更有约束力的条件来训练该模型。在这种情况下，为了保持与两步训练相关的信息，可以存储不同的损失函数条件。在另一示例中，仅可存储最后损失函数条件。

[0070] -与DNN中的至少一个权重(或神经元)的重要性相关的一个或多个重要性指标。可定义此类重要性指标的方式可以根据实施方案而变化。例如，可以至少部分地基于DNN训练期间至少一个权重(或神经元)的值的变化，和/或至少部分地基于至少一个权重(或神经元)在DNN训练期间使用的损失函数中的贡献，或至少部分地基于至少一个权重(或神经元)在至少一个预测的精度中的贡献等，来计算关于至少一个权重(或神经元)的重要性指标。

[0071] -关于在训练DNN之前使用的一个或多个预处理方法的至少一个信息，例如应用于训练集的至少一个元素的方法，诸如数据缩放、归一化、数据变换(例如，图像裁剪、旋转等)、数据增强方法等。

[0072] -与模型的一个或多个操作模式(例如具有不同精度、计算成本和/或存储占用权衡的不同操作模式)相关的至少一个信息。该信息可以是诸如对若干模式(如部署模式)之一的指定的信息或者与模式相关联的补充信息，关于与部署模式相关的一些其他信息，如可用计算资源或关于推断时间、存储要求、精度等的其他约束，

[0073] -至少一个信息，其表示网络中能够进行预测的至少一个位置(从而允许设备在该位置之后丢弃DNN模型的任何进一步计算，从而使用“轻”模型)、表示参数的不同量化、近似或压缩级别、表示附加层、表示能够被丢弃以降低精度、成本和占用的网络部分、表示具有不同权衡的相同系列的一个或多个不同模型等。

[0074] 如上文所指出，元数据可与DNN网络的参数分开地进一步处理740(例如，编码、存储于记录介质上和/或发送)(如图6A的框650所示)。元数据还可以与模型参数(架构、参数)一起被进一步处理730(例如，编码、存储在记录介质上和/或发送)(如图6A的框6200所示)。根据实施方案，用于发送元数据的信号可以使用不同的格式来编码。在一些实施方案中，元数据可以与模型参数分开编码(例如，在专用文件中)。在一些实施方案中，元数据可以与模型参数一起被编码，例如使用用于保存和交换神经网络模型的API或格式，如开放神经网络交换(ONNX)格式和/或神经网络交换格式(NNEF)。

[0075] 例如，在信号与ONNX格式兼容的一些实施方案中，元数据可以嵌入在ONNX格式的扩展中。根据另一示例，在信号与NNEF格式兼容的一些实施方案中，元数据可以作为附加信息被嵌入，如通过添加文件而被嵌入到初始化容器的扩展中。

[0076] 作为基于标准(如NNEF)的示例,其中容器的内容可以包括若干文件(例如,容器在子文件夹和文件中存储特定信息,诸如自定义格式的优化网络数据),并且其引入了描述网络结构的文本文件、结构描述中的每个可变张量的二进制数据文件(根据张量文件格式构造)、包括输出张量的量化算法细节的可选量化文件(根据量化文件格式构造),元数据可以被存储在包括特定元数据信息的附加文件中。元数据的进一步处理的示例可包括将元数据发送到另一电子设备,如边缘设备或其中将部署DNN网络的设备(如个人计算机、膝上型计算机、智能电话、平板计算机、连接的设备等)。

[0077] 发送元数据可以在训练之后立即发生,或者在训练完成之后在存储介质中存储元数据一段时间之后发生。

[0078] 图7B结合图6B和图6C示出本公开的一些实施方案的至少一种方法,其可在获得(例如,接收)预训练的第一DNN模型的元数据时和/或之后在电子设备(这里称为"目标电子设备")中实现,如另一个云设备、边缘设备或将部署DNN网络的设备(如个人计算机、移动设备、膝上型计算机、智能电话、平板计算机、连接的设备、CE设备(包括智能电视、智能助理、AI加速器))。图6B更具体地示出了"中间设备"中的示例性用例,其中获得的元数据用于DNN的训练(或者另一DNN的训练或者与接收的元数据相关的DNN相同的DNN的进一步训练)。图6C更具体地示出了示例性用例,其中元数据用于在"目标设备"中的推断之前对DNN进行微调(换句话说,用于获得微调的DNN(模型2框640)以处理推断)。

[0079] 图7B的方法可以包括获得770与第一DNN模型相关的第一元数据,以及获得760第二DNN模型(例如,已训练的第一DNN模型,其训练已经输出了接收的元数据,或者已训练的第一DNN模型之外的模型)。

[0080] 第一元数据和/或第二DNN模型可从存储介质或从由目标设备接收的信号获得。

[0081] 根据实施方案,可以分开执行第一元数据的获得或第二DNN模型的获得,或者第一元数据可以与第二DNN模型一起获得(例如当第一DNN和第二DNN相同时作为边信息)。在一些实施方案中,一些元数据可与第二DNN模型一起获得,其他元数据与第二DNN模型分开(或独立地)获得。

[0082] 可选地,如图6B和图6C的框6700所示,该方法可以包括一起或分开解码获得的元数据和/或获得的模型。图7B的方法还可以包括使用获得的第一元数据中的一些第一元数据来处理780获得的第二模型。例如,可以执行包括可选地压缩和/或解压缩模型的中间适配。处理780可涉及使用存储在至少一个文件(类似于如图6B的框620的元素DB2所示的数据库)中的一个或多个元素来训练第二DNN模型。例如,这些元素可以是音频、图像、视频、文本、时间序列的形式。根据实施方案,可以不同地执行训练780。例如,其可取决于待指派给电子设备上的DNN的一个任务或多个任务。

[0083] 在一些示例性实施方案中(例如在中间设备(如边缘设备)中实现的),可以加载第一元数据和/或第二DNN模型以用于模型适配(框6300)(例如,模型的微调)。

[0084] 在其中第二DNN模型为已训练的第一模型(模型1)的图6B的示例性用例中,使用第一元数据微调已训练的第一模型。在微调之后,获得第二DNN模型和第二元数据。

[0085] 实际上,基于深度学习技术的应用可能需要使模型适应不同任务、不同条件或优化至少一个目标设备的性能(计算成本)。微调780可使用比用于第一DNN模型的预训练720的元素(例如,DB1的元素)少的元素(例如,DB2的元素)。因此,在微调期间获得的元数据

6500(包括第一元数据)可以可选地被压缩(如由框6600所示的那样被编码)和/或被发送到一个或多个目标设备(如由图6B的框640、650所示),类似于上文已经结合图6A和图7A所解释的。

[0086] 在图6C和图7B的示例性用例(例如在目标设备中实现)中,该方法可以涉及DNN模型的推断(6400,790)。即使图6C未示出,但应理解,在一些实施方案中,也可以在目标设备中执行模型适配(例如,使用目标设备的用户的私人信息(如家庭成员图像和/或音频样本)来进行的DNN的个性化)。

[0087] 在一些示例性用例中,模型适配(在中间设备或目标设备中)可涉及(如图6B的框6300所示)用新数据集(如图6B中的DB2)微调DNN模型。在一些示例性用例中,模型适配可涉及模型压缩或修剪。模型适配还可以包括将模型分割成若干部分,这些部分稍后可以部署在不同设备上。通过在不同设备之间分割计算,此类实施方案对于加速推断时间和节省存储器可能是有益的。根据本公开的一些实施方案,除了DNN的层结构以及层的权重和偏差之外,模型适配还可以考虑与训练相关的一些元数据。

[0088] 此类元数据的示例包括与先前训练相关的一个或多个超参数。在模型适配期间元数据的使用可以是自动的(例如当元数据被用作软件应用的输入参数时),或者在一些实施方案中,表示元数据的一些信息可以例如在用户界面上提供。实际上,提供与模型的先前训练的一些设置相关的超参数可以帮助用户在用新的/个性化的数据微调模型时选择或调整一些参数设置。在一些实施方案中,例如当预处理已经被应用于在模型的先前训练期间使用的数据时,与预处理方法相关的元数据可以是有用的,以便允许类似地、完全自动地或在用户的控制下预处理在模型的适配和/或微调期间使用的数据(或至少获得与用于适配和/或微调的预处理数据类似类型的数据)。

[0089] 值得注意的是,作为元数据保存的参数的至少一部分可以在模型的适配和/或微调期间重新使用。

[0090] 此外,与一些DNN权重的重要性相关的元数据可以有助于与DNN权重相关的处理。在一些实施方案中,例如在训练集(如DB2)的个性化数据不像用于训练模型的初始数据(如DB1)那么大的实施方案中,可以仅对DNN的最重要神经元的子集执行微调,而非对整个模型权重。

[0091] 在这种情况下,与神经元的重要性相关的元数据可以有助于确定(或选择)将用于微调的子集。

[0092] 在包括减小模型大小(以便与设备(如云设备、边缘设备和/或目标设备)的一些计算/存储资源限制一致)的一些实施方案中,重要性元数据的使用可以帮助修剪不太重要的神经元以使模型更轻。此外,训练元数据可以帮助提高编码效率(例如,神经元的重要性可以被实现为排序列表)。

[0093] 作为另一示例,与DNN模型的不同操作模式相关的元数据可帮助使DNN模型适应于在推断期间DNN将执行的至少一个任务的一个或多个要求,且适应于部署DNN的设备。此类要求(或约束)可包括精度要求、能量要求、计算要求和/或存储要求。元数据可以用于例如从模型族中选择模型,挑选模型的子部分,和/或选择一组参数设置。

[0094] 元数据可因此有助于使DNN模型适应于DNN将推断的目标设备,以用于例如性能优化或节能。

[0095] 在训练模型时获得的元数据还可在编码DNN参数的至少一部分时(如在DNN的权重的量化和/或近似期间)使用,在将模型划分成待在不同设备(如用户装备、边缘设备和/或云设备)上执行的不同部件(因此在设备之间划分计算)时使用。

[0096] 预训练的NN模型的元数据因此可帮助改进DNN的部署期间的适配和/或灵活性,同时不显著影响模型的总体大小(当作为边信息与模型相关联时)。

[0097] 模型适配可以输出已适配的模型(如图6B和图6C的框640所示)。已适配的模型可以用于目标设备(例如任何CE设备,如智能电视、移动电话、机顶盒等)、边缘设备或云设备(或此类设备的组合)中的推断(即,执行特定任务)。

[0098] 结合图6B的框6300示出模型适配。然而,在一些实施方案中,可以在目标设备中(如上文所解释)或在云设备中执行模型适配。例如,其可以在云数据中心离线执行,并且新模型被保存以供将来使用。该模型随后可用于云数据中心中的推断或被发送以供在部署期间在消费电子(CE)设备中使用。

[0099] 在一些实施方案中,模型适配步骤(框6300)可以在边缘设备或云设备中在线执行,新模型被发送以供立即使用,或直接在CE设备中执行。

[0100] 在一些实施方案中,模型适配可以使用获得的元数据并且可以进一步获得附加元数据(通过实现元数据提取和/或元数据编码),该附加元数据可以与先前获得的元数据一起与已适配的模型一起或分开地被添加、保存和/或发送。

[0101] 元数据还可以例如帮助满足一些人工智能(AI)/机器学习(ML)模型分布的一些潜在服务要求,例如使模型适应目标设备的有限容量计算和能量资源的服务要求,或者更新AI/ML模型以适应变化的任务和环境的服务要求。

[0102] 图8示出了根据本公开的一些实施方案的通信网络系统的部署架构的示例,其具有DNN的开发和部署的不同层级。所示系统包括云数据中心(例如,可以实现图6A的云数据中心)、边缘网络(包括一个或多个网络设备)(例如,可以实现图6B的边缘网络/设备)和家庭网络(包括至少一个网络设备)(例如,可以实现图6C的家庭网络/设备)。

[0103] 在云数据中心中,执行DNN模型的第一训练,从训练提取元数据,并且已训练的模型和训练元数据被输入到边缘网络,在该边缘网络中,它们可以被压缩(编码和/或解码)并且被发送到家庭网络。家庭网络可以接收元数据和已训练的模型,其可以用于模型转换学习,该模型转换学习使用比在云数据中心中使用的数据集更小的数据集。如图所示,可以通过模型转换学习来获得(提取和/或生成)附加元数据。由模型转换学习输出的已适配的模型以及,可选地,获得的元数据和附加元数据可以用于DNN的推断。

[0104] 当然,图8仅是示例性实施方案。在本公开的一些其他实施方案中,示出在云中的一些框可以在边缘网络和/或家庭网络中实现,和/或示出在边缘网络中的一些框可以在云和/或家庭网络中实现,并且/或者示出在家庭网络中的一些框可以在边缘网络和/或云网络中实现(例如,压缩可以在云中执行或者压缩可以在家庭网络中执行以适应特定设备)。

[0105] 还应当指出,根据实施方案,训练的顺序方面可以被保留或省略。例如,元数据可以在架构中的不同步骤处被添加/移除/替换(因此,涉及最新模型训练的元数据可以与涉及初始模型的元数据共存或不共存)。

[0106] 本公开的至少一些实施方案可以包括与至少一个DNN模型的训练相关的至少一些元数据的压缩。如上文所解释,根据实施方案,那些元数据可作为DNN模型的边信息(例如,

描述DNN的架构和/或DNN的层的参数(如权重和偏差)的数据的边信息)而存储、编码、发送和/或解码。至少一个DNN的元数据(如训练元数据)和/或其相关联的数据的压缩可以促进元数据和/或其相关联的数据的发送和/或存储。图4和图5示出了一般工作流程(其可以在本公开的一些示例性实施方案中实现),该一般工作流程涵盖与至少一个深度神经网络的至少一层相关联的至少一个张量的参数的压缩。在一些实施方案中,可以在相同DNN的两个或更多个层上迭代地执行压缩(如图4和图5所示),并且特别地,在一些实施方案中,在相同DNN的每一层上迭代地执行压缩。

[0107] 根据本公开的实施方案,所有的至少一个层可以是卷积层、或完全连接器层,或者至少一个层可以包括至少一个卷积层和/或至少一个完全连接器层。

[0108] 在图4的示例性实施方案中,方法400可以包括获得410(或者换句话说,获取)与待压缩的层相关联的张量的参数。例如可以通过从存储单元检索至少一个层的参数或者通过经由通信接口从数据源接收参数来执行该获得。

[0109] 在本公开的至少一个实施方案中,执行神经网络的层的压缩可以包括:

[0110] -神经网络的层的参数(如权重和偏差)的量化430,以用较少数量的比特来表示它们;

[0111] -已量化信息的无损熵编码440。

[0112] 在一些实施方案中,在量化430之前,压缩400还可以包括通过利用神经网络中的固有冗余来减少420神经网络的参数(如权重和偏差)的数量的步骤。因此,与关联于层的张量的维度相比,减少420提供了降维的张量。该减少420是可选的,因此在一些实施方案中可以被省略。

[0113] 图5描绘了解码方法500,该解码方法可以用于解码通过已经结合图4描述的方法400获得的比特流。如图5所示,解码方法500可以包括解析和解码510对应于DNN的一个或多个层的比特流。更精确地,解析和解码510可以包括解码512比特流的头部部分。通过解码头部获得的已解码的头部信息可以包括例如先前用于量化对应原始张量的值的参数。方法500还可以包括解码514比特流的主体。

[0114] 当可以通过解码方法500解码若干层时,可以在每层迭代地执行方法500,直到最后一层的参数被编码(550)。

[0115] 另外的实施方案和信息

[0116] 本申请描述了各个方面,包括工具、特征、实施方案、模型、方法等。具体描述了这些方面中的许多方面,并且至少示出个体特性,通常以可能听起来有限的方式描述。然而,这是为了描述清楚,并不限制这些方面的应用或范围。实际上,所有不同的方面可组合和互换以提供进一步的方面。此外,这些方面也可与先前提交中描述的方面组合和互换。

[0117] 本专利申请中描述和设想的方面可以许多不同的形式实现。下面的图1、图2和图3提供了一些实施方案,但是设想了其他实施方案,并且图1、图2和图3的讨论不限制具体实施的广度。这些方面中的至少一个方面通常涉及编码和解码框架,其可应用于对与DNN相关的数据进行编码或解码,并且至少一个其他方面通常涉及发送生成或编码的比特流。这些和其他方面可实现为方法、装置、其上存储有用于根据该方法中任一种对数据(如与DNN相关的数据)进行编码或解码的指令的计算机可读存储介质,和/或其上存储有根据该方法中任一种生成的比特流的计算机可读存储介质。

[0118] 在本申请中，术语“重构的”和“解码的”可以互换使用，通常，但不是必须，术语“重构的”在编码器侧使用，而“解码的”在解码器侧使用。

[0119] 本文描述了各种方法，并且每种方法包括用于实现方法的一个或多个步骤或动作。除非正确操作方法需要特定顺序的步骤或动作，否则可修改或组合特定步骤和/或动作的顺序和/或用途。

[0120] 本申请中描述的各种方法和其他方面可用于修改编码器100和解码器200的模块(例如，熵编码和/或解码模块(360,150,330))，如图1和图2所示出的。此外，本发明方面不限于给定的标准，并且可以应用于例如其他标准和推荐(无论是预先存在的还是未来开发的)，以及任何此类标准和推荐的扩展。除非另外指明或技术上排除在外，否则本申请中所述的方面可单独或组合使用。

[0121] 本申请中使用了各种数值(例如关于重要性度量)。具体值是为了示例目的，并且所述方面不限于这些具体值。

[0122] 图1示出了编码器100。设想了这一编码器100的变型，但是为了清楚起见，下文描述了编码器100而不描述所有预期的变型。

[0123] 在被编码之前，数据序列可以经过预编码处理(110)，例如以便获取对压缩更具弹性的信号分布。元数据可与预处理相关联并且附接到比特流。

[0124] 在编码器100中，数据由编码器元件进行编码，如下文所描述。待编码的数据可以被分区(120)并以例如CU为单位被处理。每个单位被编码。数据可以被变换(130)和量化(140)。量化(和可选的变换)系数以及其他语法元素被熵编码(150)以输出比特流。该编码器可跳过变换，并对未变换的数据直接进行量化。该编码器可绕过变换和量化两者，即，在不应用变换或量化过程的情况下直接对数据进行编码。

[0125] 图2示出了解码器200的框图。在解码器200中，如下所述，比特流由解码器元件进行解码。解码器200通常执行与如图1所描述的编码过程相反的解码过程。编码器100通常还执行解码作为对数据编码的一部分。

[0126] 具体地，解码器的输入包括比特流，该比特流可由编码器100生成。比特流首先被熵解码(210)以获得变换系数和其他编码信息(例如关于DNN的编码层的数量和/或DNN的编码层的标识的编码信息)。分区信息指示数据是如何被分区的。因此，解码器可以根据已解码的分区信息划分(220)数据。对变换系数进行解量化(230)和逆变换(240)。

[0127] 已解码的数据可以进一步经过解码后处理(250)，例如，用于执行预编码处理(110)中执行的过程的逆过程。解码后处理可使用在预编码处理中导出并且在比特流中有信号通知的元数据。

[0128] 图3示出了实现各个方面和实施方案的系统的示例的框图。系统1000可体现为包括下文所述的各个部件的设备，并且被配置为执行本文档中所述的一个或多个方面。此类设备的示例包括但不限于各种电子设备，诸如个人计算机、膝上型计算机、智能电话、平板计算机、数字多媒体机顶盒、数字电视机接收器、个人视频录制系统、连接的家用电器和服务器。系统1000的元件可以单独地或组合地体现在单个集成电路(IC)、多个IC和/或分立部件中。例如，在至少一个实施方案中，系统1000的处理和编码器/解码器元件分布在多个IC和/或分立元件上。在各种实施方案中，系统1000经由例如通信总线或通过专用输入和/或输出端口通信地耦接到一个或多个其他系统或其他电子设备。在各种实施方案中，系统

1000被配置为实现本文档中描述的方面中的一个或多个方面。

[0129] 系统1000包括至少一个处理器1010，该至少一个处理器被配置为执行加载到其中的指令，以用于实现例如本文档中描述的各个方面。处理器1010可包括嵌入式存储器、输入输出接口和本领域已知的各种其他电路。系统1000包括至少一个存储器1020(例如，易失性存储器设备和/或非易失性存储器设备)。系统1000包括存储设备1040，该存储设备可包括非易失性存储器和/或易失性存储器，包括但不限于电可擦除可编程只读存储器(EEPROM)、只读存储器(ROM)、可编程只读存储器(PROM)、随机存取存储器(RAM)、动态随机存取存储器(DRAM)、静态随机存取存储器(SRAM)、闪存、磁盘驱动器和/或光盘驱动器。作为非限制性示例，存储设备1040可包括内部存储设备、附接存储设备(包括可拆和不可拆的存储设备)和/或网络可访问的存储设备。

[0130] 系统1000包括编码器/解码器模块1030，该编码器/解码器模块被配置为例如处理数据以提供已编码的DNN层或已解码的DNN层，并且编码器/解码器模块1030可包括其自身的处理器和存储器。编码器/解码器模块1030表示可被包括在设备中以执行编码和/或解码功能的模块。众所周知，设备可包括编码模块和解码模块中的一者或两者。此外，编码器/解码器模块1030可实现为系统1000的独立元件，或者可结合在处理器1010内作为本领域技术人员已知的硬件和软件的组合。

[0131] 要加载到处理器1010或编码器/解码器1030上以执行本文档中所述的各个方面的程序代码可存储在存储设备1040中，并且随后被加载到存储器1020上以供处理器1010执行。根据各种实施方案，处理器1010、存储器1020、存储设备1040和编码器/解码器模块1030中的一者或多者可在本文档中所描述的过程的执行期间存储各个项目中的一个或多个项目。此类存储项目可以包括但不限于输入张量、已解码的张量或已解码的张量的部分、比特流、矩阵、变量以及处理等式、公式、运算和运算逻辑的中间或最终结果。

[0132] 在一些实施方案中，处理器1010和/或编码器/解码器模块1030内部的存储器用于存储指令和提供工作存储器以用于在编码或解码期间需要的处理。然而，在其他实施方案中，处理设备外部的存储器(例如，处理设备可以是处理器1010或编码器/解码器模块1030)用于这些功能中的一个或多个功能。外部存储器可以是存储器1020和/或存储设备1040，例如动态易失性存储器和/或非易失性闪存存储器。在若干实施方案中，外部非易失性闪存存储器用于存储例如电视机的操作系统。在至少一个实施方案中，快速外部动态易失性存储器(诸如，RAM)用作用于编码及解码操作的工作存储器，该编码及解码操作诸如与元数据、DNN和/或视频相关的编码及解码操作，该编码及解码操作诸如用于MPEG-2(MPEG是指运动图片专家组，MPEG-2也被称为ISO/IEC 13818，且13818-1也被称为H.222，且13818-2也被称为H.262)、HEVC(HEVC是指高效视频编码，也被称为H.265及MPEG-H Part 2)或VVC(通用视频编码，由联合视频专家组JVET开发的新标准)、开放神经网络交换(ONNX)格式、神经网络交换格式(NNEF)、用于多媒体内容描述及分析的神经网络压缩(MPEG-NNR)格式、未来网络(包括5G)-机器学习焦点组(ITU FG-ML5G)格式、或第三代合作伙伴计划(3GPP)格式(如3GPP规范组TSG-SA(TSG服务和系统方面)。

[0133] 对系统1000的元件的输入可通过如块1130中所示的各种输入设备提供。此类输入设备包括但不限于：(i)射频(RF)部分，其接收例如由广播器通过空中发射的RF信号；(ii)分量(COMP)输入端子(或一组COMP输入端子)；(iii)通用串行总线(USB)输入端子；和/或

(iv)高清晰度多媒体接口(HDMI)输入端子。图3中未示出的其他示例包括复合视频。

[0134] 在各种实施方案中,块1130的输入设备具有本领域已知的相关联的相应输入处理元件。例如,RF部分可与适于以下项的元件相关联:(i)选择期望的频率(也称为选择信号,或将信号频带限制到一个频带),(ii)下变频选择的信号,(iii)再次频带限制到更窄频带以选择(例如)在某些实施方案中可称为信道的信号频带,(iv)解调经下变频和频带限制的信号,(v)执行纠错,以及(vi)解复用以选择期望的数据包流。各种实施方案的RF部分包括用于执行这些功能的一个或多个元件,例如频率选择器、信号选择器、频带限制器、信道选择器、滤波器、下变频器、解调器、纠错器和解复用器。RF部分可包括执行这些功能中的各种功能的调谐器,这些功能包含例如下变频接收信号至更低频率(例如,中频或近基带频率)或至基带。在一个机顶盒实施方案中,RF部分及其相关联的输入处理元件接收通过有线(例如,电缆)介质发射的RF信号,并且通过滤波、下变频和再次滤波至期望的频带来执行频率选择。各种实施方案重新布置上述(和其他)元件的顺序,移除这些元件中的一些元件,和/或添加执行类似或不同功能的其他元件。添加元件可包括在现有元件之间插入元件,例如,插入放大器和模数变换器。在各种实施方案中,RF部分包括天线。

[0135] 此外,USB和/或HDMI端子可包括用于跨USB和/或HDMI连接将系统1000连接到其他电子设备的相应接口处理器。应当理解,输入处理(例如Reed-Solomon纠错)的各个方面可根据需要例如在单独的输入处理IC内或在处理器1010内实现。类似地,USB或HDMI接口处理的方面可根据需要在单独的接口IC内或在处理器1010内实现。将经解调、纠错和解复用的流提供给各种处理元件,包括例如处理器1010以及编码器/解码器1030,该处理元件与存储器和存储元件结合操作以根据需要处理数据流以呈现在输出设备上。

[0136] 系统1000的各种元件可设置在集成外壳内。在集成外壳内,各种元件可使用合适的连接布置1140(例如,如本领域已知的内部总线,包括IC间(I2C)总线、布线和印刷电路板)互连并且在其间传输数据。

[0137] 系统1000包括能够经由通信信道1060与其他设备通信的通信接口1050。通信接口1050可包括但不限于被配置为通过通信信道1060传输和接收数据的收发器。通信接口1050可包括但不限于调制解调器或网卡,并且通信信道1060可例如在有线和/或无线介质内实现。

[0138] 在各种实施方案中,使用无线网络诸如Wi-Fi网络例如IEEE 802.11(IEEE是指电气和电子工程师协会)将数据流式发射或以其他方式提供给系统1000。这些实施方案中的Wi-Fi信号通过适用于Wi-Fi通信的通信信道1060和通信接口1050进行接收。这些实施方案的通信信道1060通常连接到接入点或路由器,该接入点或路由器提供对包括互联网的外部网络的访问,以用于允许流式应用和其他云上通信。其他实施方案使用通过输入块1130的HDMI连接递送数据的机顶盒向系统1000提供流式数据。另外的其他实施方案使用输入块1130的RF连接向系统1000提供流式数据。如上所述,各种实施方案以非流式的方式提供数据。另外,各种实施方案使用除了Wi-Fi以外的无线网络,例如蜂窝网络或蓝牙网络。

[0139] 系统1000可将输出信号提供到各种输出设备,包括显示器1100、扬声器1110和其他外围设备1120。各种实施方案的显示器1100包括例如触摸屏显示器、有机发光二极管(OLED)显示器、曲面显示器和/或可折叠显示器中的一者或多者。显示器1100可用于电视机、平板计算机、笔记本计算机、蜂窝电话(移动电话)或另外的设备。显示器1100还可以与

其他部件集成在一起(例如,如在智能电话中),或者是单独的(例如,笔记本计算机的外部监视器)。在实施方案的各种示例中,其他外围设备1120包括独立数字视频光盘(或数字通用光盘,两个术语均是DVR)、光盘播放器、立体声系统和/或照明系统中的一者或多者。各种实施方案使用提供基于系统1000的输出的功能的一个或多个外围设备1120。例如,盘播放器执行播放系统1000的输出的功能。

[0140] 在各种实施方案中,使用诸如AV等信令在系统1000与显示器1100、扬声器1110或其他外围设备1120之间传送控制信号。链路(Link)、消费电子控制(CEC)或其他能够在有或没有用户干涉的情况下实现设备到设备控制的通信协议。输出设备可通过相应接口1070、1080和1090经由专用连接通信地耦接到系统1000。另选地,输出设备可使用通信信道1060经由通信接口1050连接到系统1000。显示器1100和扬声器1110可以与电子设备(诸如电视机)中的系统1000的其他部件集成在单个单元中。在各种实施方案中,显示器接口1070包括显示驱动器,诸如例如定时控制器(T Con)芯片。

[0141] 另选地,如果输入1130的RF部分是单独机顶盒的一部分,则显示器1100和扬声器1110可选地与其他部件中的一个或多个部件分开。在显示器1100和扬声器1110为外部部件的各种实施方案中,输出信号可经由专用输出连接(包括例如,HDMI端口、USB端口或COMP输出)提供。

[0142] 这些实施方案可由处理器1010或由硬件或由硬件和软件的组合实现的计算机软件执行。作为非限制性示例,这些实施方案可由一个或多个集成电路实现。作为非限制性示例,存储器1020可以是适合于技术环境的任意类型,并且可使用任何适当的数据存储技术来实现,诸如光学存储器设备、磁存储器设备、基于半导体的存储器设备、固定存储器和可移动存储器。作为非限制性示例,处理器1010可以是适合于技术环境的任意类型,并且可涵盖微处理器、通用计算机、专用计算机和基于多核架构的处理器中的一者或多者。

[0143] 各种具体实施参与解码。如本申请中所用,“解码”可涵盖例如对所接收的编码序列执行的过程的全部或部分,以便产生适于显示的最终输出。在各种实施方案中,此类过程包括通常由解码器执行的一个或多个过程,例如熵解码、逆量化、逆变换和差分解码。在各种实施方案中,此类过程还包括或另选地包括由本应用中所述的各种具体实施的解码器执行的过程。

[0144] 作为进一步的示例,在实施方案中,“解码”仅是指熵解码,在另一个实施方案中,“解码”仅是指差分解码,并且在又一个实施方案中,“解码”是指熵解码和差分解码的组合。短语“解码过程”是具体地指代操作的子集还是广义地指代更广泛的解码过程基于具体描述的上下文将是清楚的,并且据信将被本领域的技术人员很好地理解。

[0145] 各种具体实施参与编码。以与上面关于“解码”的讨论类似的方式,如在本申请中使用的“编码”可涵盖例如对输入数据序列执行以便产生编码比特流的全部或部分过程。在各种实施方案中,此类过程包括通常由编码器执行的一个或多个过程,例如,分区、差分编码、变换、量化和熵编码。在各种实施方案中,此类过程还包括或另选地包括由本应用中所述的各种具体实施的编码器执行的过程。

[0146] 作为进一步的示例,在实施方案中,“编码”仅是指熵编码,在另一个实施方案中,“编码”仅是指差分编码,并且在又一个实施方案中,“编码”是指差分编码和熵编码的组合。短语“编码过程”是具体地指代操作的子集还是广义地指代更广泛的编码过程基于具体描

述的上下文将是清楚的，并且据信将被本领域的技术人员很好地理解。

[0147] 注意，本文所使用的语法元素是描述性术语。因此，它们不排除使用其他语法元素名称。

[0148] 当附图呈现为流程图时，应当理解，其还提供了对应装置的框图。类似地，当附图呈现为框图时，应当理解，其还提供了对应的方法/过程的流程图。

[0149] 各种实施方案是指参数模型或速率失真优化。具体地，在编码过程期间，通常考虑速率和失真之间的平衡或权衡，这常常考虑到计算复杂性的约束。可以通过速率失真优化(RDO)度量或通过最小均方(LMS)、绝对误差平均值(MAE)或其他此类测量值来测量。速率失真优化通常表述为最小化速率失真函数，该速率失真函数是速率和失真的加权和。存在不同的方法解决速率失真优化问题。例如，这些方法可基于对所有编码选项(包括所有考虑的模式或编码参数值)的广泛测试，并且完整评估其编码成本以及重构信号在编码和解码之后的相关失真。更快的方法还可用于降低编码复杂性，特别是对基于预测或预测残差信号而不是重构的残差信号的近似失真的计算。也可使用这两种方法的混合，诸如通过针对可能的编码选项中的仅一些编码选项使用近似失真，而针对其他编码选项使用完全失真。其他方法仅评估可能的编码选项的子集。更一般地，许多方法采用各种技术中任一种来执行优化，但是优化不一定是对编码成本和相关失真两者的完整评估。

[0150] 本文所述的具体实施和方面可在例如方法或过程、装置、软件程序、数据流或信号中实现。即使仅在单个形式的具体实施的上下文中讨论(例如，仅作为方法讨论)，讨论的特征的具体实施也可以其他形式(例如，装置或程序)实现。装置可在例如适当的硬件、软件和固件中实现。方法可在例如一般是指处理设备的处理器中实现，该处理设备包括例如计算机、微处理器、集成电路或可编程逻辑设备。处理器还包括通信设备，诸如例如计算机、手机、便携式/个人数字助理(“PDA”)以及便于最终用户之间信息通信的其他设备。

[0151] 提及“一个实施方案”或“实施方案”或“一个具体实施”或“具体实施”以及它们的其他变型，意味着结合实施方案描述的特定的特征、结构、特性等包括在至少一个实施方案中。因此，短语“在一个实施方案中”或“在实施方案中”或“在一个具体实施中”或“在具体实施中”的出现以及出现在本申请通篇的各个地方的任何其他变型不一定都是指相同的实施方案。

[0152] 另外，本申请可涉及“确定”各种信息。确定信息可包括例如估计信息、计算信息、预测信息或从存储器检索信息中的一者或多者。

[0153] 此外，本申请可涉及“访问”各种信息。访问信息可包括例如接收信息、检索信息(例如，从存储器)、存储信息、移动信息、复制信息、计算信息、确定信息、预测信息或估计信息中的一者或多者。

[0154] 另外，本申请可涉及“接收”各种信息。与“访问”一样，接收旨在为广义的术语。接收信息可包括例如访问信息或检索信息(例如，从存储器)中的一者或多者。此外，在诸如例如存储信息、处理信息、发射信息、移动信息、复制信息、擦除信息、计算信息、确定信息、预测信息或估计信息的操作期间，“接收”通常以一种方式或另一种方式参与。

[0155] 应当理解，例如，在“A/B”、“A和/或B”以及“A和B中的至少一者”的情况下，使用以下“/”、“和/或”和“至少一种”中的任一种旨在涵盖仅选择第一列出的选项(A)，或仅选择第二列出的选项(B)，或选择两个选项(A和B)。作为进一步的示例，在“A、B和/或C”和“A、B和C

中的至少一者”的情况下，此类短语旨在涵盖仅选择第一列出的选项(A)，或仅选择第二列出的选项(B)，或仅选择第三列出的选项(C)，或仅选择第一列出的选项和第二列出的选项(A和B)，或仅选择第一列出的选项和第三列出的选项(A和C)，或仅选择第二列出的选项和第三列出的选项(B和C)，或选择所有三个选项(A和B和C)。如对于本领域和相关领域的普通技术人员显而易见的是，这可扩展到所列出的尽可能多的项目。

[0156] 而且，如本文所用，词语“发信号通知”是指(除了别的以外)向对应解码器指示某物。例如，在某些实施方案中，编码器向多个变换、编码模式或标记中的至少一者发信号通知。这样，在一个实施方案中，在编码器侧和解码器侧两者均使用相同的参数。因此，例如，编码器可将特定参数发射(显式信令)到解码器，使得解码器可使用相同的特定参数。相反，如果解码器已具有特定参数以及其他，则可在不发射(隐式信令)的情况下使用信令，以简单允许解码器知道和选择特定参数。通过避免发射任何实际功能，在各种实施方案中实现了位节省。应当理解，信令可以各种方式实现。例如，在各种实施方案中，使用一个或多个语法元素、标记等将信息发信号通知至对应解码器。虽然前面涉及词语“signal(发信号通知)”的动词形式，但是词语“signal(信号)”在本文也可用作名词。

[0157] 对于本领域的普通技术人员将显而易见的是，具体实施可产生格式化为携带例如可存储或可发送的信息的各种信号。信息可包括例如用于执行方法的指令或由所述具体实施中的一个具体实施产生的数据。例如，可格式化信号以携带所述实施方案的比特流。可格式化此类信号例如为电磁波(例如，使用频谱的射频部分)或基带信号。格式化可包括例如对数据流编码并且用编码的数据流调制载体。信号携带的信息可以是例如模拟或数字信息。已知的是，信号可通过各种不同的有线或无线链路发射。信号可存储在处理器可读介质上。

[0158] 我们描述了多个实施方案。这些实施方案的特征可在各种权利要求类别和类型中单独地或以任何组合提供。此外，实施方案可包括以下特征、设备或方面中的一个或多个，单独地或以任何组合，跨各种权利要求类别和类型：

[0159] ●一种过程或设备，其用于执行获得和/或存储与神经网络的训练相关的一个或多个元数据。

[0160] ●一种过程或设备，其用于执行获得和/或存储与深度神经网络模型的训练相关的一个或多个元数据作为深度神经网络模型和/或参数的边信息。

[0161] ●一种过程或设备，其用于执行与神经网络的训练相关的一个或多个元数据的编码和解码，以实现元数据压缩。

[0162] ●一种过程或设备，其用于执行与神经网络的训练相关的一个或多

[0163] 个元数据的编码和解码，以实现深度神经网络压缩和元数据压缩作为深度神经网络模型和/或参数的边信息。。

[0164] ●一种过程或设备，其用于利用表示参数的比特流中的插入信息执

[0165] 行编码和解码，以实现对包括一个或多个层的预训练的深度神经网络、以及作为深度神经网络模型和/或参数的边信息的与深度神经网络的先前训练相关的元数据进行深度神经网络压缩。

[0166] ●一种过程或设备，其用于利用表示参数的比特流中的插入信息执

[0167] 行编码和解码，以实现对预训练的深度神经网络、以及作为深度神经网络模型和/

或参数的边信息的与深度神经网络的先前训练相关的元数据进行深度神经网络压缩，直到达到压缩标准。

[0168] ●包括所描述的语法元素中的一个或多个语法元素或其变型的比特

[0169] 流或信号。

[0170] ●包括传递根据所述实施方案中任一项生成的信息的语法的比特流

[0171] 或信号。

[0172] ●根据所描述的实施方案中任一项所述的创建和/或发送和/或接收和/

[0173] 或解码比特流或信号。

[0174] ●根据所述实施方案中任一项所述的方法、过程、装置、存储指令

[0175] 的介质、存储数据的介质或信号。

[0176] ●在信令中插入语法元素，这使得解码器能够以与编码器所使用的

[0177] 方式相对应的方式确定编码模式。

[0178] ●对包括所描述的语法元素中的一个或多个语法元素或其变型的比

[0179] 特流或信号进行创建和/或发送和/或接收和/或解码。

[0180] ●根据所描述的实施方案中的任一实施方案执行变换方法的电视、机顶盒、蜂窝电话、平板计算机或其他电子设备。

[0181] ●根据所描述的实施方案中的任一实施方案执行变换方法确定并显

[0182] 示所得图像(例如，使用监视器、屏幕或其他类型的显示器)的电视、机顶盒、蜂窝电话、平板计算机或其他电子设备。

[0183] ●根据所描述的实施方案中的任一实施方案选择、频带限制或调谐(例如，使用调谐器)信道以接收包括编码图像的信号并执行变换方法的电视、机顶盒、蜂窝电话、平板计算机或其他电子设备。

[0184] ●通过空中接收(例如，使用天线)包括编码图像的信号并且执行

[0185] 变换方法的电视机、机顶盒、蜂窝电话、平板计算机或其他电子设备。

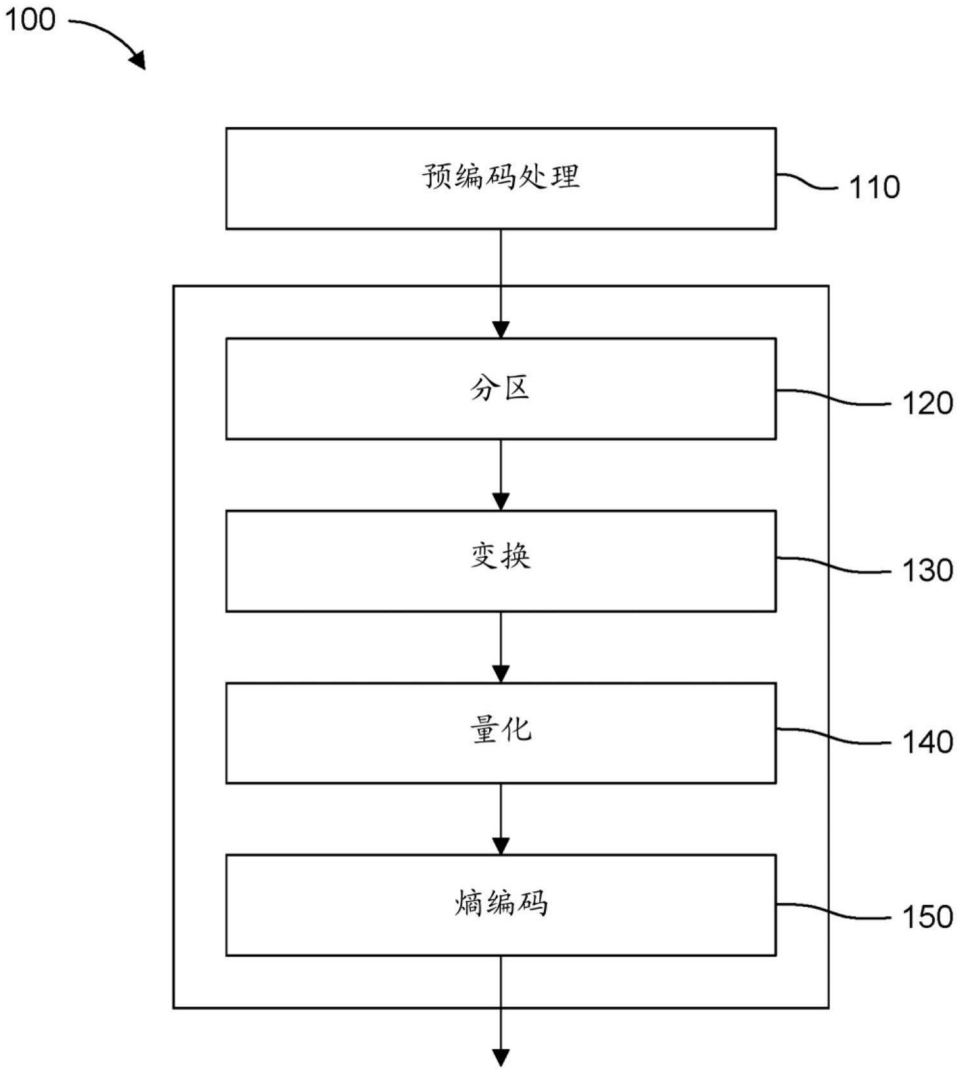


图1

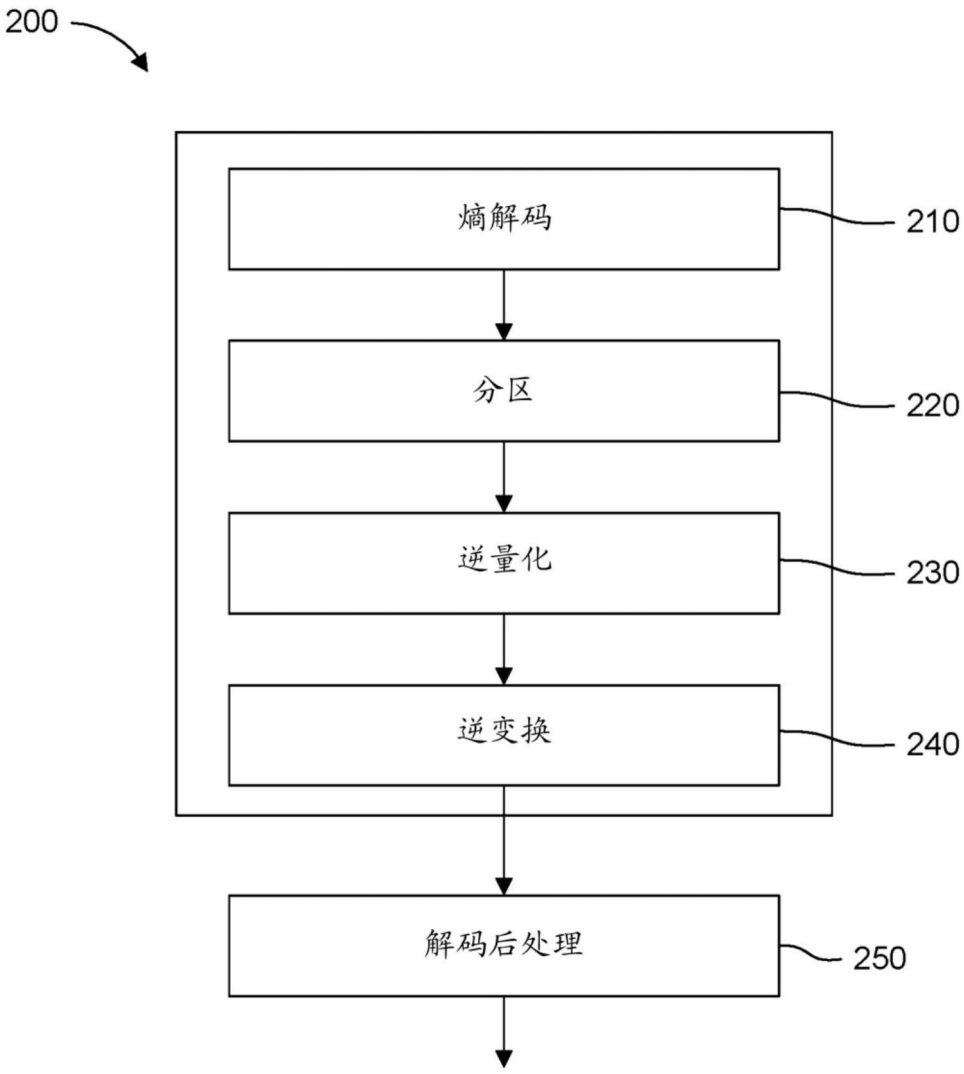


图2

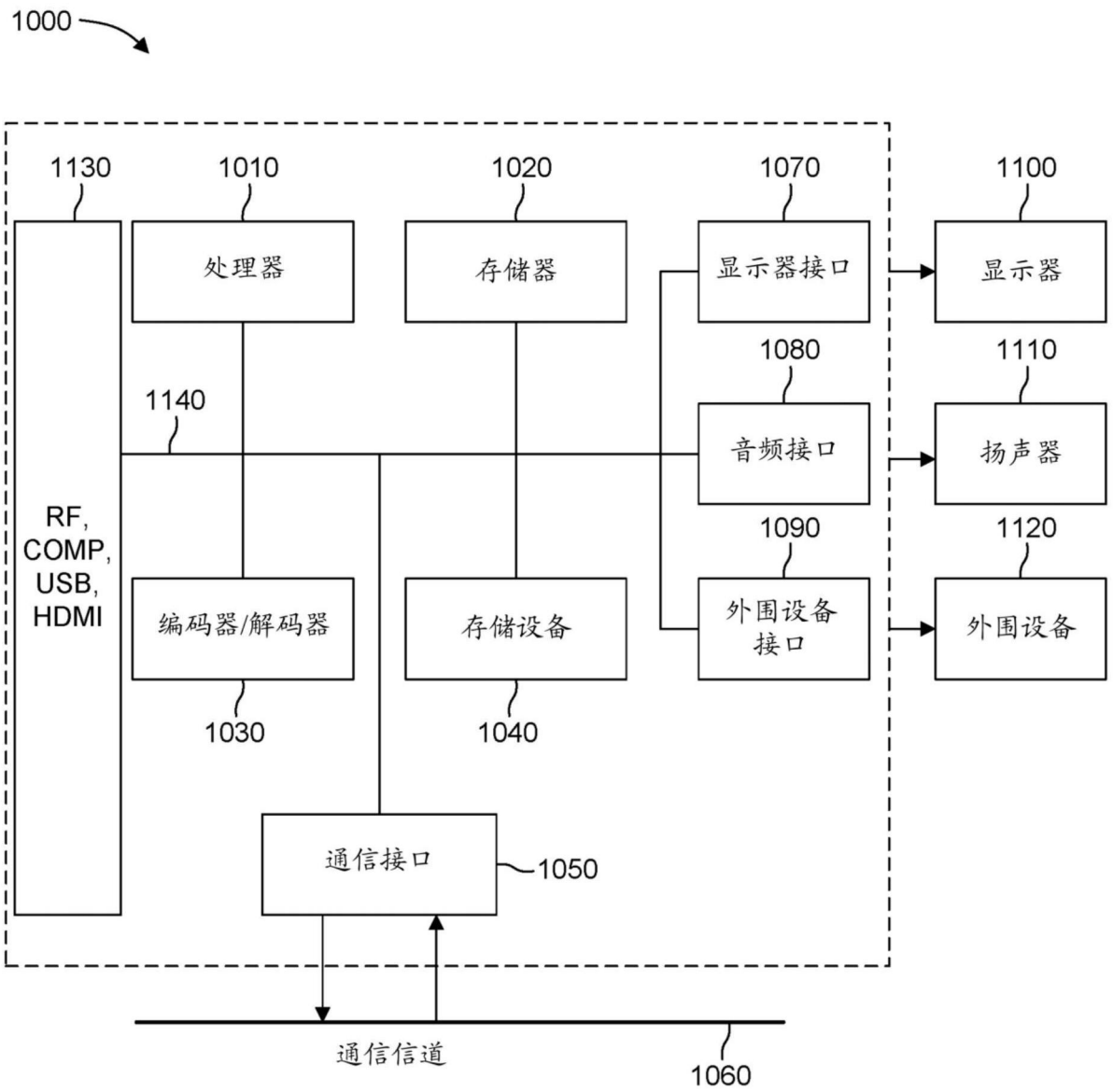

1000
1130
1010
1020
1070
1100
处理器
存储器
显示器接口
显示器
1080
1110
1140
音频接口
扬声器
RF,
COMP,
USB,
HDMI
1090
1120
编码器/解码器
存储设备
外围设备
接口
外围设备
1030
1040
通信接口
1050
通信信道
1060

图3

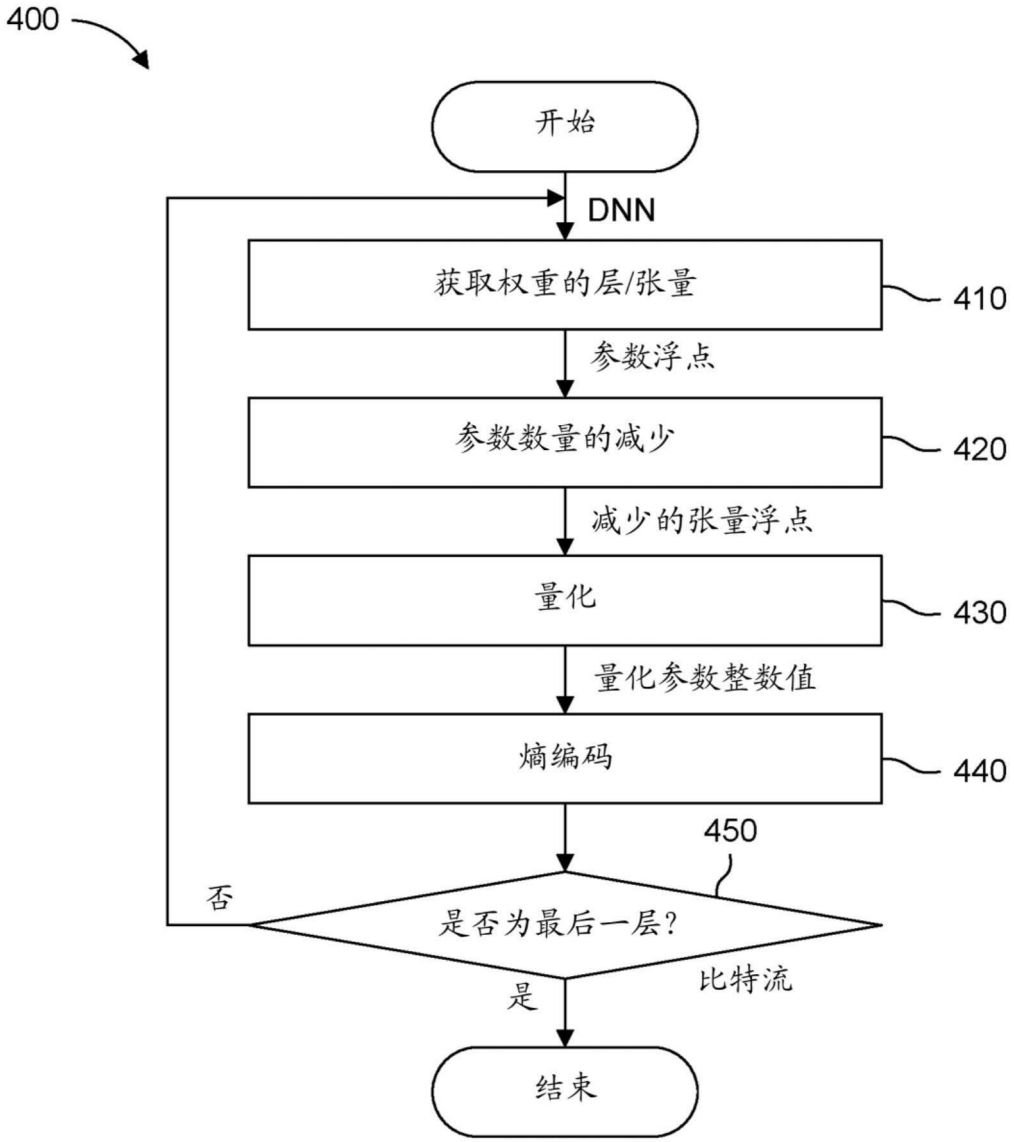

400
开始
DNN
获取权重的层/张量
410
参数浮点
参数数量的减少
420
减少的张量浮点
量化
430
量化参数整数值
熵编码
440
450
否
是否为最后一层?
比特流
是
结束

图4

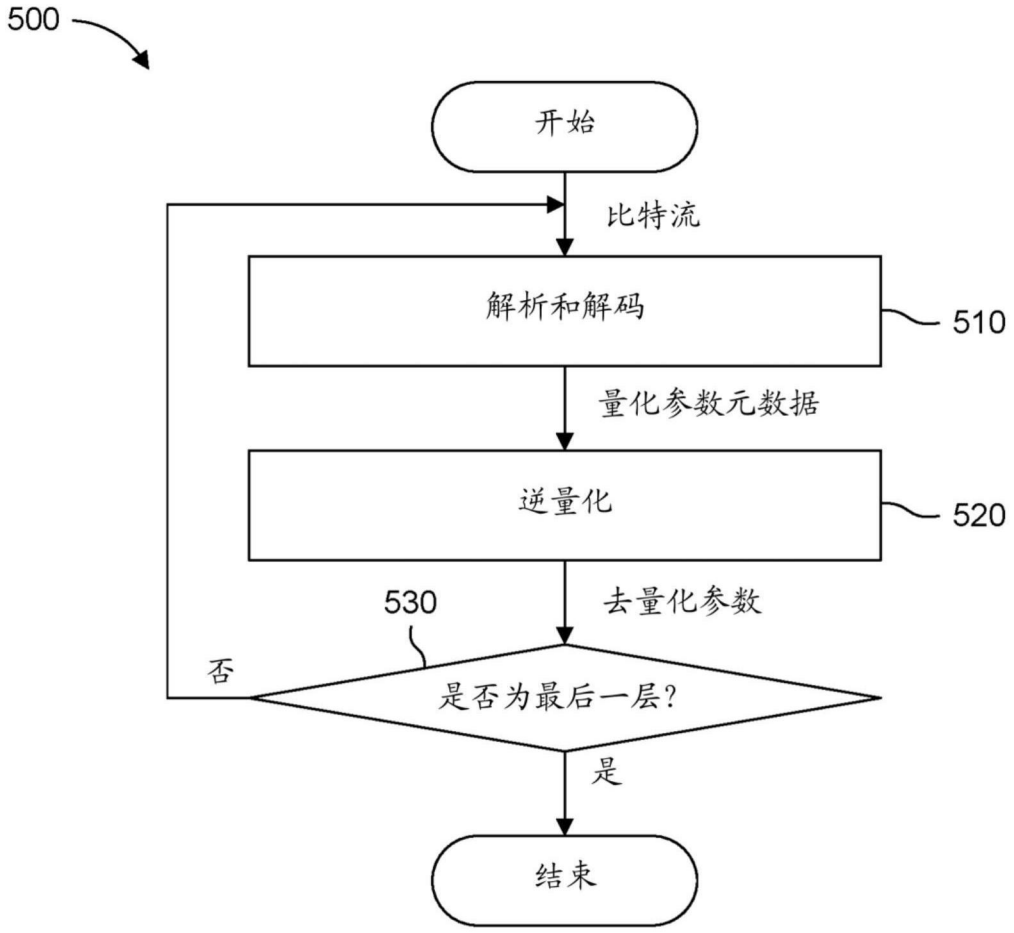


图5

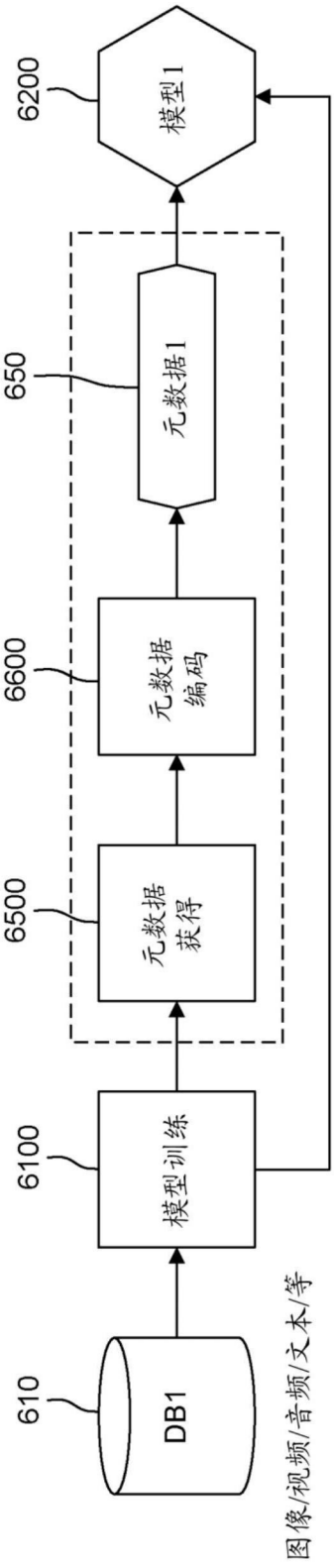

6200
模型1
650
元数据1
6600
元数据编码
6500
元数据获得
6100
模型训练
610
DB1
图像/视频/音频/文本/等

图6A

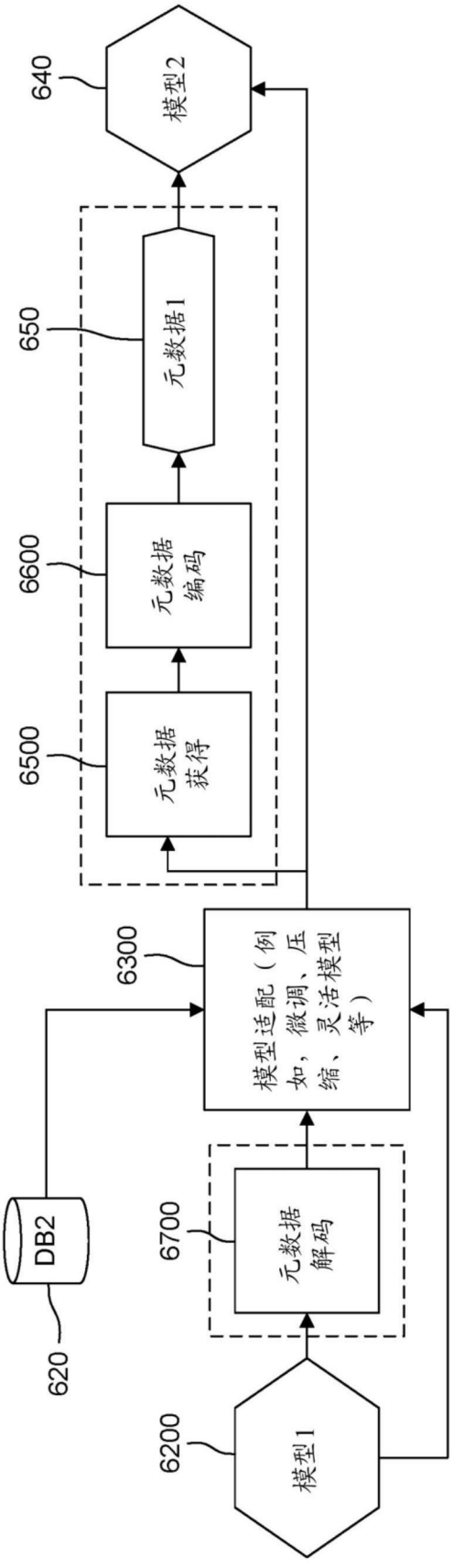

640
模型2
650
元数据1
6600
元数据编码
6500
元数据获得
6300
模型适配（例如，微调、压缩、灵活模型等）
6700
元数据解码
DB2
620
6200
模型1

图6B

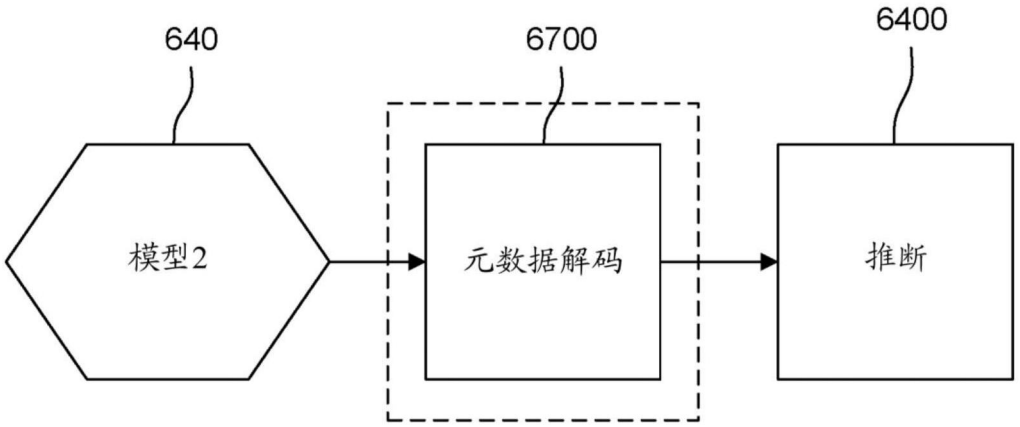

640
6700
6400
模型2
元数据解码
推断

图6C

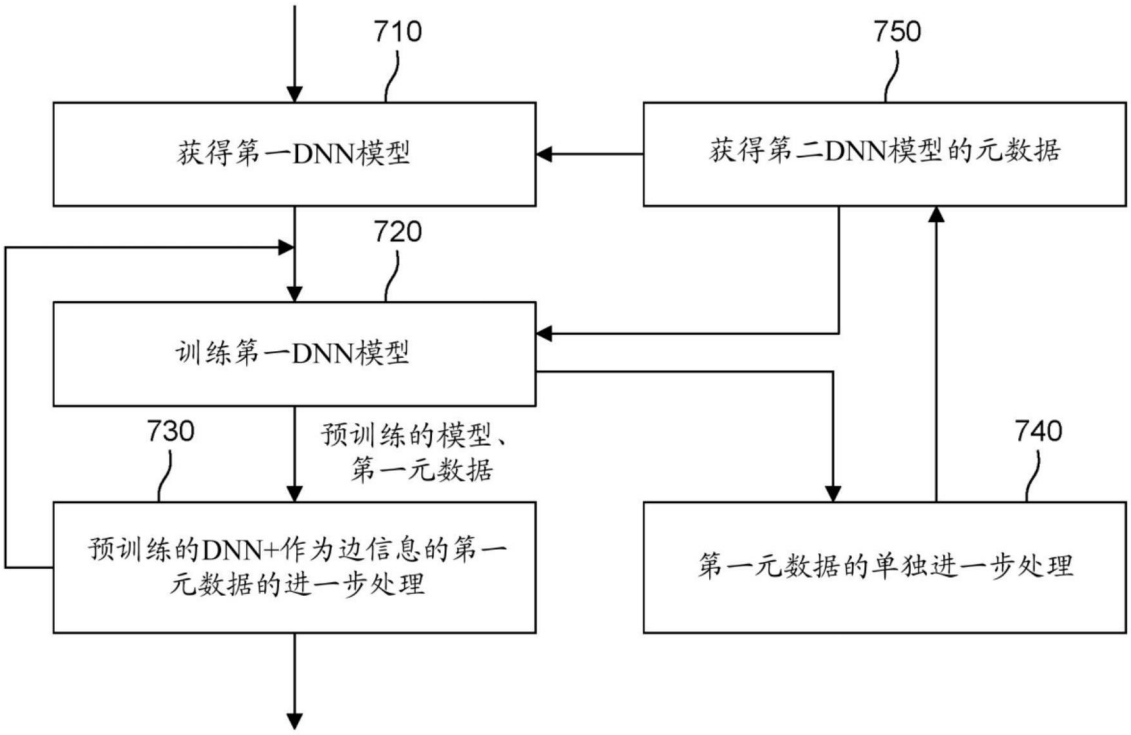

710
750
获得第一DNN模型
获得第二DNN模型的元数据
720
训练第一DNN模型
730
预训练的模型、
第一元数据
740
预训练的DNN+作为边信息的第一
元数据的进一步处理
第一元数据的单独进一步处理

图7A

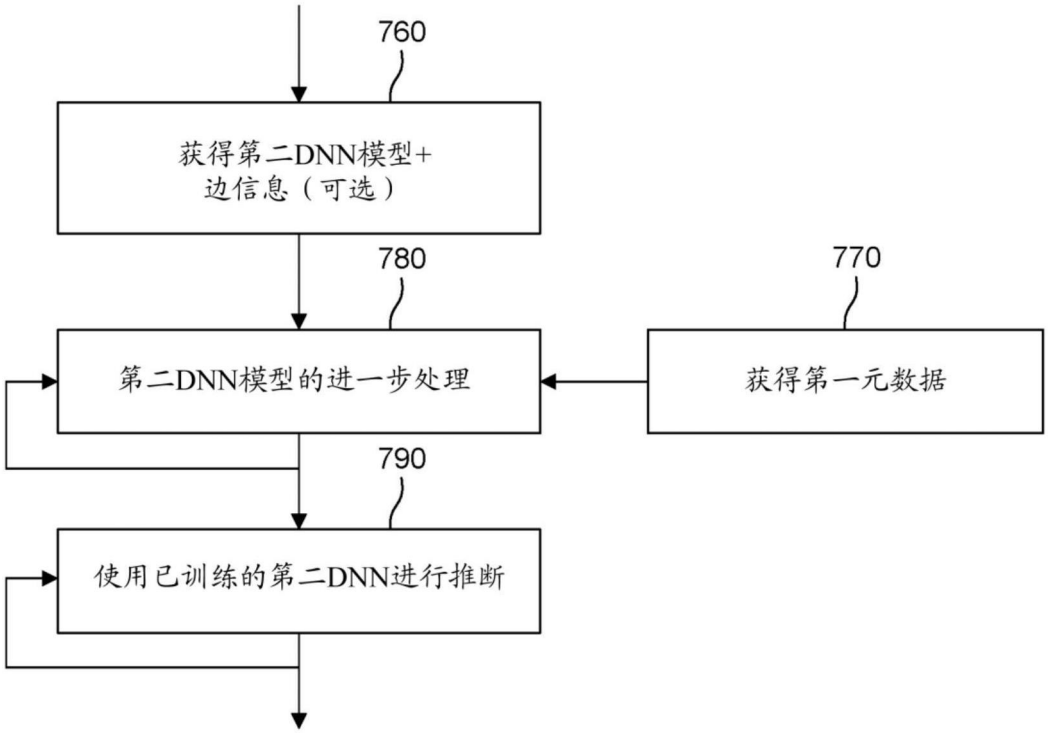

760
获得第二DNN模型+
边信息（可选）
780
770
第二DNN模型的进一步处理
获得第一元数据
790
使用已训练的第二DNN进行推断

图7B

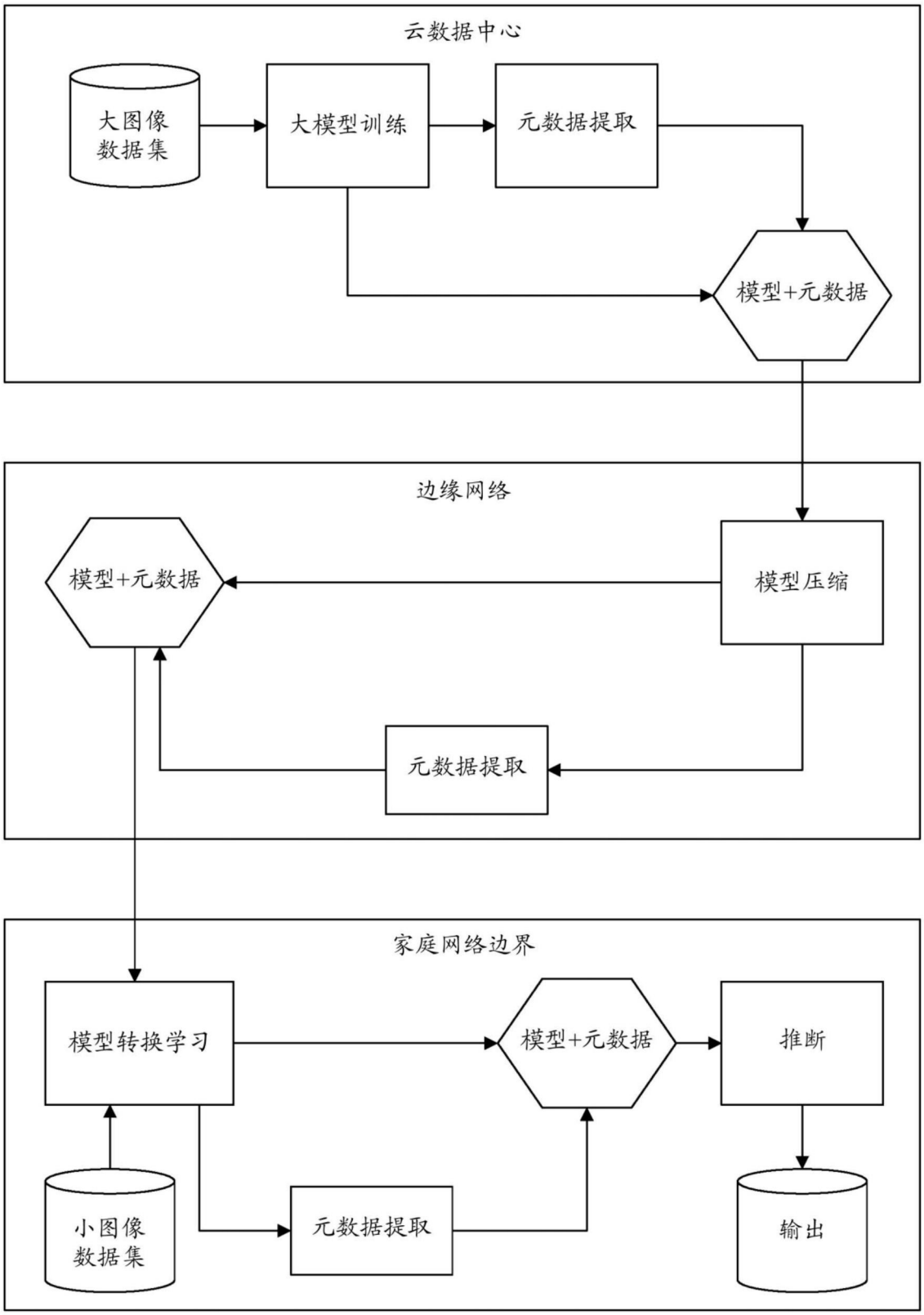

云数据中心
大图像
数据集
大模型训练
元数据提取
模型+元数据
边缘网络
模型+元数据
模型压缩
元数据提取
家庭网络边界
模型转换学习
模型+元数据
推断
小图像
数据集
元数据提取
输出

图8