(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 116258183 A
(43)申请公布日 2023.06.13

(21)申请号 202210663806.3
(22)申请日 2022.06.13
(30)优先权数据
10-2021-0171979 2021.12.03 KR
(71)申请人 三星电子株式会社
地址 韩国京畿道
(72)发明人 李元熙
(74)专利代理机构 中科专利商标代理有限责任公司 11021
专利代理师 周祺 李敬文
(51)Int.Cl.
G06N 3/063(2023.01)
G06T 19/00(2011.01)
G06N 3/08(2023.01)
G06N 3/09(2023.01)
G06N 3/088(2023.01)
G06N 3/0464(2023.01)

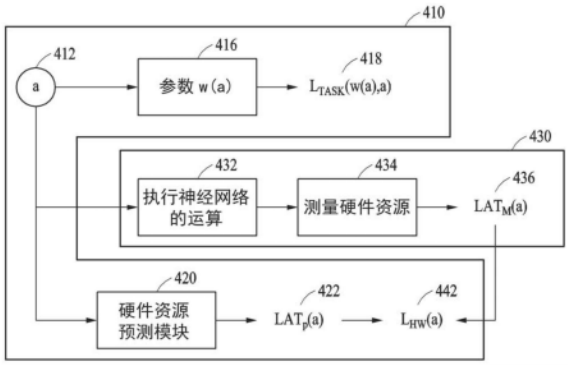
权利要求书2页 说明书13页 附图5页

(54)发明名称

用于神经网络架构搜索的方法和装置

(57)摘要

公开了一种用于搜索神经网络的最佳架构的方法和装置。该装置可以包括处理器，该处理器被配置为：基于神经网络的候选架构的参数产生神经网络损失；测量在具有候选架构的神经网络的运算中使用的第一硬件资源；使用硬件资源预测模型产生第二硬件资源的预测，该第二硬件资源将用于操作具有候选架构的神经网络；基于第一硬件资源和第二硬件资源确定硬件资源损失；以及基于神经网络损失和硬件资源损失确定神经网络的目标架构。



CN 116258183 A

1.一种计算装置,所述计算装置包括:

处理器,被配置为:

基于神经网络的候选架构的参数产生神经网络损失;

测量在具有所述候选架构的所述神经网络的运算中使用的第一硬件资源;

使用硬件资源预测模型产生第二硬件资源的预测,所述第二硬件资源将用于操作具有所述候选架构的所述神经网络;

基于所述第一硬件资源和所述第二硬件资源确定硬件资源损失;以及

基于所述神经网络损失和所述硬件资源损失确定所述神经网络的目标架构。

2.根据权利要求1所述的计算装置,其中,所述硬件资源预测模型包括神经网络,所述神经网络被配置为:接受所述候选架构的参数作为输入,基于所述参数预测所述候选架构的神经网络的所述第二硬件资源,并且输出硬件资源预测值。

3.根据权利要求1所述的计算装置,其中,所述处理器进一步被配置为:

基于所述第一硬件资源和所述第二硬件资源之间的差异确定所述硬件资源损失;以及

更新所述候选架构的参数以最小化所述硬件资源损失。

4.根据权利要求1所述的计算装置,其中,所述处理器进一步被配置为:将所述神经网络损失和所述硬件资源损失的加权和确定为优化损失,并确定所述目标架构以最小化所述优化损失。

5.根据权利要求1所述的计算装置,其中,所述处理器进一步被配置为:确定所述目标架构和目标参数以减少所述神经网络损失和所述硬件资源损失。

6.根据权利要求1所述的计算装置,其中,对于所述神经网络的每个层,通过从相应层的候选运算中选择相应的候选运算来确定相应的候选架构。

7.根据权利要求6所述的计算装置,其中与所选择的候选运算相关联的信息被输入到所述硬件资源预测模型。

8.根据权利要求1所述的计算装置,其中,所述第一硬件资源包括测量的功耗、存储器需求、运算的数量、以及操作具有所述候选架构的神经网络的处理时间中的任何一项或任何组合。

9.根据权利要求1所述的计算装置,其中,所述处理器进一步被配置为:

确定包括所述神经网络损失和所述硬件资源损失的优化损失;以及

通过从所述候选架构的神经网络中包括的每个层的候选运算中选择使所述优化损失最小化的目标运算来确定所述目标架构。

10.根据权利要求1所述的计算装置,其中,所述处理器进一步被配置为:基于由所述候选架构的神经网络处理训练数据而输出的结果数据与验证数据之间的差异,确定所述神经网络损失。

11.一种用于搜索神经网络的最佳架构的处理器实现的方法,所述方法包括:

基于神经网络的候选架构的参数产生神经网络损失;

测量操作所述候选架构的神经网络所需的第一硬件资源;

使用硬件资源预测模块预测操作所述候选架构的神经网络所需的第二硬件资源;

基于所述第一硬件资源和所述第二硬件资源确定硬件资源损失;以及

基于所述神经网络损失和所述硬件资源损失确定所述神经网络的目标架构。

12.根据权利要求11所述的方法,其中,所述硬件资源预测模块包括神经网络,所述神经网络被配置为:接受所述候选架构的参数作为输入,基于所述参数预测所述候选架构的神经网络的所述第二硬件资源,并且输出硬件资源预测值。

13.根据权利要求11所述的方法,其中,确定所述目标架构包括:确定所述目标架构以最小化所述神经网络损失和所述硬件资源损失的加权和。

14.根据权利要求11所述的方法,其中,确定所述目标架构包括:通过从所述候选架构的神经网络中包括的每个层的候选运算中选择目标运算来确定所述目标架构。

15.根据权利要求11所述的方法,其中,所述第一硬件资源和所述第二硬件资源包括功耗、存储器需求、运算的数量、以及操作所述候选架构的神经网络的处理时间中的任何一项或任何组合。

16.一种存储指令的非暂时性计算机可读存储介质,所述指令在由处理器执行时使所述处理器执行根据权利要求11所述的方法。

17.一种用于识别神经网络的架构的处理器实现的方法,所述方法包括:

基于神经网络的候选架构的参数产生神经网络损失;

测量在操作具有所述候选架构的所述神经网络时使用的第一硬件资源;

使用硬件资源预测模块预测具有所述候选架构的神经网络的第二硬件资源;

基于所述第一硬件资源和所述第二硬件资源之间的差异产生硬件资源损失;以及

基于所述神经网络损失和所述硬件资源损失选择作为所述神经网络的目标架构的候选架构,

其中,所述硬件资源预测模块包括硬件资源预测神经网络,所述硬件资源预测神经网络被训练为接受所述神经网络的候选架构的参数作为输入并输出所述第二硬件资源。

18.根据权利要求17所述的方法,其中,选择作为所述神经网络的目标架构的候选架构包括:通过所述神经网络损失与应用了权重的所述硬件资源损失的结果的最小和来确定所述神经网络的目标架构。

19.根据权利要求17所述的方法,其中,产生所述硬件资源损失包括:基于应用考虑所述第一硬件资源与所述第二硬件资源之间的差异的损失函数,产生所述硬件资源损失。

20.根据权利要求17所述的方法,其中,所述候选架构包括神经网络结构,所述神经网络结构包括所述神经网络的每个层的候选运算集合。

用于神经网络架构搜索的方法和装置

技术领域

[0001] 以下描述涉及搜索神经网络的最佳架构。

背景技术

[0002] 典型的神经架构搜索可以是针对给定目的自动搜索神经网络的可能最佳架构的一种方法。这种搜索神经网络的可能合适的结构和架构形状的神经网络取决于通过深度学习解决给定问题的搜索能力。可以通过选择和组合包括预定义算子和函数的原始运算(也被称为搜索空间)来产生可能的最佳神经网络。算子的示例可以包括卷积、池化、级联、跳过连接等。

发明内容

[0003] 提供本发明内容以用简化形式介绍对下面在具体实施方式中进一步描述的构思的选择。本发明内容不意在标识所请求保护的主题的关键特征或基本特征,也不意在帮助确定所请求保护的主题的范围。

[0004] 在一个总体方面,提供了一种计算装置,该装置包括处理器,该处理器被配置为:基于神经网络的候选架构的参数产生神经网络损失;测量在具有候选架构的神经网络的运算中使用的第一硬件资源;使用硬件资源预测模型产生第二硬件资源的预测,该第二硬件资源将用于操作具有候选架构的神经网络;基于第一硬件资源和第二硬件资源确定硬件资源损失;以及基于神经网络损失和硬件资源损失确定神经网络的目标架构。

[0005] 硬件资源预测模型可以包括神经网络,神经网络被配置为:接受候选架构的参数作为输入,基于参数预测候选架构的神经网络的第二硬件资源,并且输出硬件资源预测值。

[0006] 处理器可以被配置为基于第一硬件资源与第二硬件资源之间的差异来确定硬件资源损失,并且更新候选架构的参数以最小化硬件资源损失。

[0007] 处理器可以被配置为:将神经网络损失和硬件资源损失的加权和确定为优化损失,并确定目标架构以最小化优化损失。

[0008] 处理器可以被配置为确定目标架构和目标参数以减少神经网络损失和硬件资源损失。

[0009] 对于神经网络的每个层,可以通过从相应层的候选运算中选择相应的候选运算来确定相应的候选架构。

[0010] 与所选择的候选运算相关联的信息可以被输入到硬件资源预测模型。

[0011] 第一硬件资源可以包括测量的功耗、存储器需求、运算的数量、以及操作具有候选架构的神经网络的处理时间中的任何一项或任何组合。

[0012] 处理器可以被配置为确定包括神经网络损失和硬件资源损失的优化损失,并且通过从候选架构的神经网络中包括的每个层的候选运算中选择使优化损失最小化的目标运算来确定目标架构。

[0013] 处理器可以被配置为基于由候选架构的神经网络处理训练数据而输出的结果数

据与验证数据之间的差异，确定神经网络损失。

[0014] 在另一个总体方面，提供了一种用于提供增强现实(AR)的装置，该装置包括处理器，该处理器被配置为：基于神经网络的候选架构的参数产生神经网络损失；测量在具有候选架构的神经网络的运算中使用的第一硬件资源；使用硬件资源预测模型产生第二硬件资源的预测，该第二硬件资源将用于操作具有候选架构的神经网络；基于第一硬件资源和第二硬件资源确定硬件资源损失；基于神经网络损失和硬件资源损失确定神经网络的目标架构；以及实现具有目标架构的神经网络以处理具有输入数据的运算。

[0015] 该装置可以包括被配置为捕捉图像的相机，其中处理器可以被配置为通过使用具有目标架构的神经网络处理图像或从图像导出的数据来产生AR内容。

[0016] 硬件资源预测模型可以由硬件资源预测模块来实现，该硬件资源预测模块可以包括将候选架构的参数作为输入的神经网络，并且该硬件资源预测模块被配置为基于该参数来预测候选架构的神经网络的第二硬件资源，并输出指示第二硬件资源的硬件资源预测值。

[0017] 处理器可以被配置为确定目标架构以最小化神经网络损失和硬件资源损失的加权和。

[0018] 在另一个总体方面，提供了一种用于搜索神经网络的最佳架构的处理器实现的方法，该方法包括：基于神经网络的候选架构的参数来确定神经网络损失；测量操作候选架构的神经网络所需的第一硬件资源；使用硬件资源预测模块来预测操作候选架构的神经网络所需的第二硬件资源；基于第一硬件资源和第二硬件资源来确定硬件资源损失；以及基于神经网络损失和硬件资源损失来确定神经网络的目标架构。

[0019] 硬件资源预测模块可以包括神经网络，该神经网络被配置为：接受候选架构的参数作为输入，基于参数预测候选架构的神经网络的第二硬件资源，并且输出硬件资源预测值。

[0020] 确定目标架构可以包括：确定目标架构以最小化神经网络损失和硬件资源损失的加权和。

[0021] 确定目标架构可以包括：通过从候选架构的神经网络中包括的每个层的候选运算中选择目标运算来确定目标架构。

[0022] 第一硬件资源和第二硬件资源可以包括功耗、存储器需求、运算的数量、以及操作候选架构的神经网络的处理时间中的任何一项或任何组合。

[0023] 在另一个总体方面，提供了一种用于识别神经网络的架构的处理器实现的方法，该方法包括：基于神经网络的候选架构的参数来确定神经网络损失；测量在操作具有候选架构的神经网络时使用的第一硬件资源；使用硬件资源预测模块来预测具有候选架构的神经网络的第二硬件资源；基于第一硬件资源与第二硬件资源之间的差异来产生硬件资源损失；以及基于神经网络损失和硬件资源损失来选择候选架构作为神经网络的目标架构，其中硬件资源预测模块可以包括硬件资源预测神经网络，该硬件资源预测神经网络被训练为接受神经网络的候选架构的参数作为输入并输出第二硬件资源。

[0024] 选择作为神经网络的目标架构的候选架构可以包括：通过神经网络损失与应用了权重的硬件资源损失的结果的最小和来确定神经网络的目标架构。

[0025] 产生硬件资源损失可以包括：基于应用考虑第一硬件资源与第二硬件资源之间的

差异的损失函数，产生硬件资源损失。

[0026] 候选架构可以包括神经网络结构，该神经网络结构包括神经网络的每个层的候选运算集合。

[0027] 其他特征和方面将通过具体实施方式、附图和权利要求书变得清楚明白。

附图说明

[0028] 图1示出了用于搜索神经网络的最佳架构的示例系统。

[0029] 图2示出了用于搜索神经网络的最佳架构的示例计算装置。

[0030] 图3示出了在每个层的候选运算中选择最佳目标运算的过程的示例。

[0031] 图4示出了用于确定神经网络的目标架构的搜索过程的示例。

[0032] 图5示出了用于搜索神经网络的最佳架构的方法的操作的示例。

[0033] 图6示出了电子设备的示例。

[0034] 在整个附图和详细描述中，除非另有描述或提供，否则相同的附图标记应被理解为指代相同的元件、特征以及结构。附图可以不按比例绘制，并且为了清楚、说明和方便，可以扩大附图中元件的相对尺寸、比例和描绘。

具体实施方式

[0035] 提供以下详细描述以帮助读者获得对本文描述的方法、装置和/或系统的全面理解。然而，在理解了本申请的公开之后，本文中描述的方法、装置和/或系统的各种改变、修改和等同物将是显而易见的。例如，本文中描述的操作顺序仅仅是示例，并且不限于在本文中阐述的那些操作顺序，而是可以在理解本申请的公开之后明显改变，除了必须以一定顺序进行的操作之外。

[0036] 本文描述的特征可以以不同形式来实施，并且不应被解释为限于本文描述的示例。相反，提供本文中描述的示例仅仅是为了说明实现本文中描述的方法、装置和/或系统的许多可行方式中的一些，在理解本申请的公开之后这些方式将显而易见。

[0037] 尽管本文中可以使用诸如“第一”、“第二”和“第三”、“A”、“B”、“C”、(a)、(b)、(c)之类的术语来描述各种构件、组件、区域、层或部分，但是这些构件、组件、区域、层或部分不受这些术语的限制。相反，这些术语仅用于将一个构件、组件、区域、层或部分与另一构件、组件、区域、层或部分加以区分。因此，在不脱离示例的教导的情况下，本文中描述的示例中提及的第一构件、组件、区域、层或部分也可以被称为第二构件、组件、区域、层或部分。

[0038] 贯穿说明书，当组件被描述为“连接到”或“耦接到”另一组件时，它可以直接“连接到”或“耦接到”该另一组件，或者可以存在介于其间的一个或多个其他组件。相反，当元件被描述为“直接连接到”或“直接耦接到”另一元件时，可以不存在介于其间的其他元件。

[0039] 如本文中使用的，单数形式“一”、“一个”和“所述”意图还包括复数形式，除非上下文明确地给出相反的指示。如本文中所使用的，术语“和/或”包括关联列出的项目中的任何一个和任何两个或更多个的任何组合。如本文所用，术语“包括”、“包含”和“具有”表示存在所阐述的特征、数目、操作、构件、元件和/或其组合，但并不排除存在或添加一个或多个其他特征、数目、操作、构件、元件和/或其组合。除非另有定义，本文中使用的所有术语(包括技术或科学术语)具有与根据本公开内容通常所理解的和在理解本公开内容之后通常所理

解的含义相同的含义。除非本文中另外定义，否则诸如通常使用的字典中定义的术语之类的术语应被解释为具有与相关领域和本公开内容中的上下文含义相匹配的含义，并且不应被解释为理想的或过度形式化的含义。

[0040] 在本文中，关于示例或实施例(例如，关于示例或实施例可以包括或实现什么)的术语“可以”的使用意味着存在至少一个示例或实施例，其中这样的特征是被包括或实现的，而所有示例不限于此。在下文中，将参考附图来详细描述示例。当参考附图描述示例时，相同的附图标记表示相同的组件，并且将省略与其相关的重复描述。

[0041] 图1示出了用于搜索神经网络的最佳架构的系统的示例。

[0042] 神经网络或人工神经网络(ANN)可以产生输入信息与输出信息之间的映射，并且可以具有针对尚未用于训练的输入信息推断出相对正确的输出的泛化能力。作为非限制性示例，神经网络可以指具有解决问题或执行任务的能力的通用模型，其中，节点通过连接和通过训练的其他参数调整来形成网络。

[0043] 神经网络可以实现为具有多个层的架构，该多个层包括输入层、隐藏信息以及输出层。在神经网络层中，输入图像或映射可以与被称为核的滤波器进行卷积，并且作为结果，可以输出多个特征映射。输出的特征映射可以在后续的卷积层中再次作为输入特征映射与另一个核进行卷积，并且可以输出多个新的特征映射。作为非限制性示例，在重复执行卷积运算以及可能地执行其他层操作之后，最终可以输出通过神经网络针对输入图像的特征的识别或分类结果。

[0044] 神经网络可以是机器学习模型结构。在另一示例中，神经网络层可以从输入数据中提取特征数据并基于该特征数据提供推断。特征数据也可以是与通过对输入数据进行抽象而获得的特征相关联的数据。神经网络可以基于深度学习以非线性关系映射输入数据和输出数据，以产生这种推断。深度学习，例如，通过神经网络的多个隐藏层的反向传播可以针对各种目的或任务(例如，根据大数据集进行的语音识别或语音音译)产生经训练的神经网络，可以通过有监督学习和/或无监督学习将输入数据与输出数据进行互相映射，但这仅仅是示例。

[0045] 在示例中，训练人工神经网络可以指示确定并调整层之间的权重和偏差或属于彼此相邻的不同层的多个节点之间的权重和偏差，但这仅仅是此类参数的非限制性示例。

[0046] 参考图1，系统100可以是基于硬件的框架，其被配置为通过机器学习搜索基本神经网络120的最佳架构(或神经网络结构)。基本神经网络120可以是尚未训练的神经网络(或未经训练的神经网络)，其中每个层的操作和参数(例如，连接权重)未被确定。基本神经网络120可以包括多个神经网络层(或被简称为“层”)。在示例中，基本神经网络120可以是例如深度神经网络(DNN)、卷积神经网络(CNN)、循环神经网络(RNN)、受限玻尔兹曼机(RBM)、深度信念网络(DBN)、双向循环DNN(BRDNN)、深度Q网络、或其中两种或更多种的组合，但其示例不限于前述示例。基本神经网络120可以包括可通过处理器执行指令来实现的硬件结构。

[0047] 系统100可以基于数据库(DB)110中存储的训练数据对基本神经网络120执行机器学习。在图1的示例中，可以通过监督学习或部分监督学习方法来执行机器学习。

[0048] 在示例中，系统100可以通过监督学习来训练基本神经网络120。系统100可以基于调整算法(例如，随机梯度下降方案)和损失函数来执行训练。用于训练的训练数据可以包

括要输入到神经网络的输入数据和对应于该输入数据的验证数据(或真实数据)。基本神经网络120可以处理训练数据中包括的输入数据,以输出结果数据。系统100可以基于验证数据与从基本神经网络120输出的结果数据之间的比较结果来确定神经网络损失,并且可以搜索使神经网络损失最小化的最佳架构。

[0049] 系统100可以通过执行多目标神经架构搜索(NAS)方法来搜索目标神经网络130的最佳架构。系统100可以不对基本神经网络120的架构进行采样。系统100可以针对基础神经网络120的每个层设置多个候选运算,在这些候选运算中选择最合适的候选运算,并且搜索最佳架构。通过这种搜索方法,系统100可以对神经网络执行有效的优化,并节省时间、能源和计算资源。

[0050] 系统100可以通过训练过程基于给定目的(例如,对象分类、对象识别、语音识别等)输出具有最佳架构的目标神经网络130。搜索最佳架构可以包括确定由神经网络的每个层执行的运算以及确定神经网络的参数的最佳值。系统100可以由本文描述的用于搜索神经网络的最佳架构的装置(例如,图2中的计算装置200)来执行。

[0051] 在示例中,当搜索目标神经网络130的最佳架构时,系统100可以考虑硬件资源约束。系统100可以通过考虑在执行神经网络时使用的硬件资源和由神经网络执行的任务的验证损失来执行优化。系统100可以通过考虑操作神经网络所需的硬件资源来搜索目标神经网络130。硬件资源可以是例如功耗、存储器需求、运算的数量(例如,乘法累加(MAC)运算的数量)、处理时间、以及图形处理单元(GPU)占用率。系统100可以考虑一个或多个硬件资源,并且除了上述硬件资源之外,可以无限制地考虑在数字上可以观察到的任何硬件资源。

[0052] 当确定基本神经网络120的候选架构时,系统100可以确定候选架构的神经网络损失和硬件资源的硬件资源损失,并且可以搜索使神经网络损失和硬件资源损失最小化的目标架构。神经网络损失和硬件资源损失可以包括用于确定目标架构的优化损失。

[0053] 当要确定硬件资源损失时,系统100可以基于使用由候选架构的神经网络所需的硬件资源的实际测量值所输出的硬件资源的预测值和硬件资源预测模块(例如,图4中的硬件资源预测模块420)来确定硬件资源损失。将在下文中详细描述,硬件资源预测模块是针对各种训练目标提供预测值的硬件模块,该预测值用于预测候选架构的神经网络的硬件资源。硬件资源预测模块可以由经训练的神经网络实现,以基于候选架构的输入参数,例如由表示硬件资源预测模块的处理器或本文中的任何其他处理器来输出由候选架构的神经网络所需的硬件资源的预测值。硬件资源预测模块可以具有可微分特性,并且可以通过硬件资源预测模块在搜索过程中保持可微分性。在示例中,当保持可微分性时,可以执行示例端到端的学习。系统100可以通过硬件资源预测模块在优化损失中反映用于神经网络的架构的硬件资源。

[0054] 如上所述,系统100可以考虑硬件资源约束来搜索神经网络的最佳架构,并且可以在短时间段内执行优化。此外,系统100可以通过考虑实际的硬件资源测量值来搜索最佳架构。

[0055] 图2示出了用于搜索神经网络的最佳架构的计算装置的结构的示例。

[0056] 参考图2,计算装置200可以是用于搜索神经网络的最佳架构并且可以执行参考图1所描述的系统100的设备。计算装置200可以执行本文描述或示出的与数据处理方法相关的一个或多个操作。计算装置200可以包括处理器210和存储器220。存储设备230可以存储

用于架构搜索的数据(例如,训练数据)和用于训练的神经网络。

[0057] 存储器220可以存储由计算装置200的组件(例如,处理器210)使用的各种数据。各种数据可以包括例如计算机可读指令以及用于与其相关的操作的输入数据或输出数据。存储器220可以包括易失性存储器和非易失性存储器中的任意一种或任意组合。

[0058] 易失性存储器件可以实现为动态随机存取存储器(DRAM)、静态随机存取存储器(SRAM)、晶闸管RAM(T-RAM)、零电容RAM(Z-RAM)或双晶体管RAM(TTRAM)。

[0059] 非易失性存储器件可以实现为电可擦可编程只读存储器(EEPROM)、闪存、磁RAM(MRAM)、自旋力矩(STT)-MRAM、导电桥接RAM(CBRAM))、铁电RAM(FeRAM)、相变RAM(PRAM)、电阻RAM(RRAM)、纳米管RRAM、聚合物RAM(PoRAM)、纳米浮栅存储器(NFGM)、全息存储器、分子电子存储设备、或绝缘体电阻变化存储器。在下文中提供关于存储器220的进一步细节。

[0060] 处理器210可以控制计算装置200的整体操作,并且可以执行用于执行计算装置200的操作的相应处理器可读指令。处理器210可以执行例如软件来控制与处理器210连接的计算装置200的一个或多个硬件组件(例如,下面在图6中描述的其他组件),并且可以执行各种数据处理或操作、以及对这些组件的控制。

[0061] 在示例中,作为数据处理或操作的至少一部分,处理器210可以将指令或数据存储在存储器220中,执行存储器220中存储的指令和/或处理数据,并且将从其获得的结果数据存储在存储器220中。处理器200可以是由硬件实现的数据处理设备,该硬件包括具有执行期望操作的物理结构的电路。例如,期望操作可以包括程序中包括的代码或指令。

[0062] 硬件实现的数据处理设备可以包括例如主处理器(例如,中央处理单元(CPU)、现场可编程门阵列(FPGA)、或应用处理器(AP))、或者可独立于主处理器或与主处理器联合操作的辅处理器(例如,GPU、神经处理单元(NPU)、图像信号处理器(ISP)、传感器集线器处理器或通信处理器(CP))。在下文中提供关于处理器210的进一步细节。

[0063] 处理器210可以从存储器220读取神经网络数据/向存储器220写入神经网络数据,神经网络数据例如是文本数据、语音数据、图像数据、特征映射数据、内核数据、偏差、权重(例如,连接权重数据)、超参数和其他参数等,并且处理器210可以使用读取/写入的数据来实现神经网络。当实现神经网络时,处理器210可以重复执行输入和参数之间的运算,以便产生关于输出的数据。这里,在示例卷积层中,可以取决于诸如输入或输入特征映射的通道的数量、内核的通道的数量、输入特征映射的大小、内核的大小、内核的数量、以及值的精度之类的各种因素来确定卷积运算的数量。这种神经网络可以实现为复杂的架构,其中处理器210以高达数亿至数百亿的运算计数来执行卷积运算,并且处理器210为了卷积运算而访问存储器920的频率迅速增大。

[0064] 处理器210可以使用训练数据来学习神经网络(例如,图1中的基本神经网络120)的候选架构并且可以确定神经网络损失。尚未训练的神经网络可以包括多个层,每个层包括一个或多个节点,并且可以预定义可以由每个层执行的候选运算。候选运算可以包括例如基于3×3内核的卷积运算、基于5×5内核的卷积运算、以及池化操作,但不限于此。对于神经网络的每个层,可以通过选择每个层的候选运算中的任何一个来确定候选架构。处理器210可以基于神经网络的候选架构的参数来确定神经网络损失。处理器210可以基于通过处理候选架构的神经网络的训练数据所输出的结果数据与验证数据之间的差异来确定神经网络损失。神经网络损失可以由损失函数来确定。在示例中,可以预定义损失函数。

[0065] 处理器210可以测量操作候选架构的神经网络所需(或所用)的硬件资源。测量的硬件资源可以包括例如功耗、存储器需求、运算的数量、以及当候选架构的神经网络操作时的处理时间中的一项或多项,但不限于此。处理器210可以测量硬件资源以确定硬件资源测量值。

[0066] 处理器210可以使用硬件资源预测模块(例如,图4中的硬件资源预测模块420)来预测当候选架构的神经网络操作时所需要的硬件资源。硬件资源预测模块可以包括模型(例如,机器学习模型),例如接收候选架构的参数(例如,与从每个层的候选运算中选择的候选运算相关联的信息)作为输入的神经网络,并且硬件资源预测模块可以通过基于输入参数来预测候选架构的神经网络的硬件资源来输出资源预测值。

[0067] 处理器210可以基于测量的硬件资源和预测的硬件资源来确定硬件资源损失。处理器210可以基于测量的硬件资源与预测的硬件资源之间的差异以及损失函数来确定硬件资源损失。例如,当候选架构的神经网络操作时,处理器210可以通过将实际处理时间测量值与从硬件资源预测模型输出的预期处理时间值之间的差应用于损失函数,来确定硬件资源损失。

[0068] 处理器210可以基于神经网络损失和硬件资源损失来确定神经网络的目标架构。处理器210可以确定用于减少神经网络损失和硬件资源损失的目标架构和目标参数。处理器210可以更新候选架构的参数,使得硬件资源损失最小化,例如(作为非限制性示例),在朝向更小或最小损失的方向上最小化,或者基于最小阈值。处理器210可以确定包括神经网络损失和硬件资源损失的优化损失,并且可以从候选架构的神经网络中包括的每个层的候选运算中选择使优化损失最小化的目标运算。处理器210可以选择每个层的使整个神经网络的神经网络损失和硬件资源损失最小化的目标运算,并且可以更新神经网络的参数。处理器210可以确定基于候选架构参数的神经网络损失和基于候选架构参数的硬件资源损失的加权和作为优化损失,并且可以确定使优化损失最小化的目标架构。

[0069] 由上述计算装置200执行的操作可以不同地应用于可以在嵌入式系统以及移动设备(例如,可穿戴设备、智能电话等)中运行的基于神经网络的算法。

[0070] 图3示出了在每个层的候选运算中选择最佳目标运算的过程的示例。

[0071] 参考图3,在操作310中,尚未训练的神经网络(例如,图1中的基本神经网络120)可以包括多个层312、314、316和318、以及可以针对层312、314、316和318中的每一个定义的多个候选运算。在示例中,神经网络可以具有层312、314、316和318中的每一层的三个候选运算。在层之间执行的候选运算可以具有不同的运算方法。例如,在层312与层314之间执行的候选运算可以是不同类型的运算。

[0072] 计算装置(例如,图2中的计算装置200)可以在训练过程中选择作为候选运算中的最佳候选运算的目标运算。在操作320中,计算装置可以在层312、314、316和318中的每一层的候选运算322、324、326、328和329中选择任何一个候选运算,并且可以确定由所选候选运算322、324、326、328和329的组合形成的候选架构的优化损失。在搜索空间中,计算装置可以多次组合层312、314、316和318中的每一层的候选运算,并且可以计算优化损失的每个组合以最小化优化损失(或目标运算)。在操作330中,当各种组合的训练过程完成时,可以基于针对层312、314、316和318中的每一层选择的目标运算来确定目标架构。目标架构可以包括每个层的目标运算。

[0073] 图4示出了用于确定神经网络的目标架构的搜索过程的示例。

[0074] 参考图4，当在神经网络的目标架构的搜索过程中给出神经网络的候选架构“a”412时，可以确定候选架构“a”412的参数w(a)416。候选架构“a”412可以表示包括针对每个层选择的候选运算的集合的神经网络结构。

[0075] 候选架构“a”412的参数可以包括神经网络的每个层的候选运算中的所选候选运算的参数、以及指示每个所选候选运算的运算特性的参数。例如，假设神经网络中包括的层的候选运算包括基于3×3内核的卷积运算和基于5×5内核的卷积运算，则候选架构“a”412的参数可以包括对哪个卷积运算选自这些卷积运算加以指示的参数以及所选卷积运算的内核参数。这里，卷积运算可以由卷积层来实现。

[0076] 计算装置(例如，图2中的计算装置200)可以基于参数w(a)416针对候选架构“a”412的神经网络的任务确定神经网络损失$L_{TASK}(w(a),a)$418。神经网络损失$L_{TASK}(w(a),a)$418可以是使由神经网络执行的任务的损失最小化的损失。神经网络损失$L_{TASK}(w(a),a)$418可以是有监督或无监督的基于验证的损失。

[0077] 在操作432中，计算装置可以使用具有候选架构“a”412的神经网络来执行操作，并且在操作434中，计算装置可以实际测量整个神经网络的在执行相应操作的过程中所需的硬件资源。要测量的硬件资源可以包括例如功耗、存储器需求、运算的数量、处理时间等。当在操作434中测量硬件资源时，可以确定候选架构“a”412的硬件资源测量值$LAT_M(a)$436。在示例中，可以通过对通过最大运算所确定的架构执行神经网络运算来测量硬件资源。由于包括这种处理的过程430不具有可微分性，因此可能无法确定过程430的前向运算和后向运算定义。为了解决这种问题，可以使用硬件资源预测模块420。

[0078] 计算装置可以预测整个神经网络的当具有候选架构“a”412的神经网络使用硬件资源预测模块420来执行运算时所需的硬件资源。可以通过硬件资源预测模块420来确定候选架构“a”412的硬件资源预测值$LAT_P(a)$422。硬件资源预测模块420可以接收候选架构“a”412的参数作为输入，基于输入参数来预测候选架构“a”412的神经网络的硬件资源，并且可以输出硬件资源预测值$LAT_P(a)$422。硬件资源预测模块420中的运算可以被执行为可微分运算。

[0079] 硬件资源预测模块420可以是通过训练过程训练的模型或神经网络，使得基于神经网络架构的参数来输出预测神经网络的架构所需或所用的硬件资源的预测值。然而，除了模型或神经网络之外，硬件资源预测模块420还可以通过可以基于候选架构“a”412来预测神经网络的硬件资源的其他方法来实现。

[0080] 计算装置可以基于硬件资源测量值$LAT_M(a)$436和硬件资源预测值$LAT_P(a)$422来确定候选架构“a”412的硬件资源损失LHW(a)442。硬件资源损失$L_{HW}(a)$442可以被定义为使得硬件资源测量值$LAT_M(a)$436与硬件资源预测值$LAT_P(a)$422之间的差被最小化。

[0081] 在示例中，硬件资源损失$L_{HW}(a)$442可以基于对硬件资源测量值$LAT_M(a)$436与硬件资源预测值$LAT_P(a)$422之间的差的损失加以指示的$L_{HW1}(a)$、以及作为用于优化硬件资源的因素(例如，用于最小化延迟的因素)的$L_{HW2}(a)$来确定。$L_{HW1}(a)$和$L_{HW2}(a)$可以分别由下面的等式1和2来确定。

[0082] [等式1]

[0083] $$L_{HW1}(a) = (LAT_M(a) - LAT_P(a))^2$$

[0084] [等式2]

[0085] $L_{HW2}(a) = (LAT_M(a))^2$

[0086] 硬件资源损失$L_{HW}(a)$442可以被确定为$L_{HW1}(a)$与$L_{HW2}(a)$之间的加权和，例如，如下面的等式3所表示。

[0087] [等式3]

[0088] $L_{HW}(a) = L_{HW1}(a) + w \times L_{HW2}(a)$

[0089] 在等式3中，w表示应用于$L_{HW2}(a)$的权重，并且可以是例如预设常数。在示例中，权重可以仅应用于$L_{HW1}(a)$，或者不同的权重可以分别应用于$L_{HW1}(a)$和$L_{HW2}(a)$。

[0090] 计算装置可以通过将硬件资源测量值$LAT_M(a)$436与硬件资源预测值$LAT_P(a)$422之间的差应用于损失函数来确定硬件资源损失$L_{HW}(a)$442。在示例中，可以预定义在等式1至等式3中描述的损失函数L_{HW}。

[0091] 计算装置可以确定包括神经网络损失$L_{TASK}(w(a),a)$418和硬件资源损失$L_{HW}(a)$442在内的优化损失，并且可以通过从候选架构“a”412的神经网络中包括的每个层的候选运算中选择使优化损失最小化的目标运算，来确定目标架构。例如，可以通过搜索候选架构“a”412和使优化损失最小化的候选架构“a”的参数$w(a)$416来确定目标架构，如下面的等式4所示。在等式4的示例中，优化损失可以通过神经网络损失$L_{TASK}(w(a),a)$418与应用了权重λ的硬件资源损失$L_{HW}(a)$442的结果的最小和来确定。

[0092] [等式4]

[0093] $$\min_{a}\min_{w} L_{TASK}(w(a),a) + \lambda \cdot L_{\mathrm{HW}}(a)$$

[0094] 计算装置可以在短时间段内有效地搜索神经网络的目标架构，并且可以在选择了目标架构时将神经网络所需的硬件资源视为优化约束。在图4中，过程410可以具有可微分特性，而过程430可以具有不可微分特性。计算设备可以通过使用硬件资源预测模块420基于神经网络的每个层的运算来预测整个神经网络的硬件资源，从而保持可微分性。在示例中，硬件资源预测模块420可以实现为神经网络。当保持可微分性时，可以执行端到端学习。此外，上述搜索过程可以通过反映整个神经网络的硬件资源约束来将神经网络的架构优化为可微分，并且可以仅使用短优化时间，因为目标架构通过单个学习过程被找到。此外，在目标架构的搜索过程中可以保持前向操作和后向操作的一致性。

[0095] 图5示出了搜索神经网络的最佳架构的方法的操作的示例。图5中的操作可以按照所示的顺序和方式来执行，然而在不脱离所描述的说明性示例的精神和范围的情况下，可以改变一些操作的顺序，或者省略一些操作。图5所示的许多操作可以并行或同时执行。图5的一个或多个块和这些块的组合可以通过执行指定功能的基于专用硬件的计算机(例如，处理器)或者专用硬件和计算机指令的组合来实现。例如，该方法的操作可以由计算装置(例如，图2中的计算装置200)来执行。除了下面对图5的描述之外，图1至图4的描述也适用于图5，并且通过引用并入本文。因此，这里可以不再重复以上描述。

[0096] 参考图5，在操作510，计算装置可以选择神经网络(例如，图1中的基本神经网络120)的候选架构。计算装置可以通过在神经网络的每个层的候选运算中选择任何一个所定义的候选运算来选择候选架构。

[0097] 在操作520，计算装置可以基于神经网络的候选架构的参数来确定神经网络损失。计算装置可以使用训练数据来学习神经网络的候选架构并且可以确定神经网络损失。计算

装置可以基于由神经网络处理的验证数据与结果数据之间的差异来确定神经网络损失。当从候选架构的神经网络输出的结果数据与目标验证数据之间的差异增加时，神经网络损失会增加。

[0098] 在操作530，计算装置可以测量用于操作候选架构的神经网络的物理硬件资源。测量的硬件资源可以包括例如功耗、存储器需求、运算的数量、以及当候选架构的神经网络操作时的处理时间中的一项或多项，但不限于此。

[0099] 在操作540，计算装置可以使用硬件资源预测模块(例如，图4中的硬件资源预测模块420)来预测当候选架构的神经网络操作时所需要或所使用的硬件资源。与包括可以由神经网络的每个层执行的候选运算中的候选运算在内的所选候选运算相关联的信息可以被输入到硬件资源预测模型，并且硬件资源预测模型可以基于输入信息来提供相应神经网络所需要或所使用的硬件资源的预测值。

[0100] 在操作550，计算装置可以基于测量的硬件资源和预测的硬件资源来确定硬件资源损失。计算装置可以基于测量的硬件资源与预测的硬件资源之间的差异以及损失函数来确定硬件资源损失。在示例中，可以预定义损失函数。

[0101] 在操作560，计算装置可以基于神经网络损失和硬件资源损失来确定神经网络的目标架构和目标参数。计算装置可以确定使包括神经网络损失和硬件资源损失在内的优化损失最小化的目标架构和目标参数。计算设备可以通过在候选架构的神经网络中包括的每个层的候选运算中选择使优化损失最小化的目标运算，来确定目标架构。计算装置可以确定基于候选架构参数的神经网络损失和基于候选架构参数的硬件资源损失的加权和作为优化损失，并且确定使优化损失最小化的目标架构。

[0102] 图6示出了电子设备的示例。

[0103] 参考图6，电子设备600可以实现为各种类型的计算设备(例如，个人计算机(PC)、数据服务器或便携式设备)或在其中实现。在示例中，便携式设备可以实现为膝上型计算机、移动电话、智能电话、平板PC、移动互联网设备(MID)、个人数字助理(PDA)、企业数字助理(EDA)、数字静止相机、数字视频相机、便携式多媒体播放器(PMP)、个人导航设备或便携式导航设备(PND)、手持游戏控制台、电子书、智能汽车、自动驾驶汽车、或智能设备。在示例中，电子设备600可以是可穿戴设备，例如，用于提供增强现实(AR)的设备(以下被简称为“AR提供设备”)，例如，AR眼镜、头戴式显示器(HMD)、智能手表和产品检测设备。

[0104] 电子设备600可以包括处理器610、存储器620、相机630、传感器640、输出设备650和通信设备660。电子设备600的组件中的至少一些可以相互耦接并经由外设间通信接口670(例如，总线、通用输入和输出(GPIO)接口、串行外围接口(SPI)、移动行业处理器接口(MIPI))在它们之间交换信号(例如，命令或数据)。

[0105] 处理器610可以是由硬件实现的处理设备，该硬件包括具有执行操作的物理结构的电路。例如，可以通过执行配置处理设备执行所述操作的任何一个或任何组合的计算机可读指令来实现操作。

[0106] 例如，硬件实现的数据处理设备可以包括微处理器、中央处理单元(CPU)、处理器核、多核处理器、多处理器，专用集成电路(ASIC)和现场可编程门阵列(FPGA)中的任一种。下面提供关于处理器610的进一步细节。

[0107] 处理器610可以控制电子设备600的整体操作并且执行要由电子设备600执行的功

能和指令。处理器610可以执行上面参考图1至图5描述的计算装置(例如,图2中的计算装置200)的操作。

[0108] 存储器620可以存储可由处理器610执行的指令、输入/输出数据、以及各种神经网络参数。存储器620可以包括易失性存储器和/或非易失性存储器。易失性存储器件可以实现为动态随机存取存储器(DRAM)、静态随机存取存储器(SRAM)、晶闸管RAM(T-RAM)、零电容RAM(Z-RAM)或双晶体管RAM(TTRAM)。

[0109] 非易失性存储器件可以实现为电可擦可编程只读存储器(EEPROM)、闪存、磁RAM(MRAM)、自旋力矩(STT)-MRAM、导电桥接RAM(CBRAM))、铁电RAM(FeRAM)、相变RAM(PRAM)、电阻RAM(RRAM)、纳米管RRAM、聚合物RAM(PoRAM)、纳米浮栅存储器(NFGM)、全息存储器、分子电子存储设备、或绝缘体电阻变化存储器。在下文中提供关于存储器620的进一步细节。

[0110] 相机630可以捕捉图像。相机630可以获得例如彩色图像、黑白图像、灰色图像、红外(IR)图像或深度图像。例如,由相机630捕捉的图像可以用作到CNN的卷积层的输入。

[0111] 传感器640可以检测电子设备600的操作状态(例如,电力或温度)或电子设备600外部的环境状态(例如,用户的状态),并且可以产生与检测到的状态相对应的电信号或数据值。传感器640可以包括例如手势传感器、陀螺仪传感器、大气压力传感器、磁性传感器、加速度传感器、握持传感器、接近传感器、颜色传感器、IR传感器、生物特征传感器、温度传感器、湿度传感器或照度传感器。传感器640可以包括用于测量电子设备600的各种资源的传感器。

[0112] 输出设备650可以通过视觉、听觉或触觉通道向用户提供电子设备600的输出。输出设备650可以包括例如显示设备,例如液晶显示器或发光二极管(LED)/有机LED显示器、微型LED、触摸屏、扬声器、振动发生器设备、或可以向用户提供输出的任何其他设备。在示例中,输出设备650还可以被配置为接收来自用户的输入,例如语音输入、手势输入或触摸输入。

[0113] 通信设备660可以被解释为在电子设备600与外部电子设备之间建立直接(或有线)通信信道或无线通信信道,并且可以支持通过所建立的通信信道进行通信。在示例中,通信设备660可以包括无线通信模块(例如,蜂窝通信模块、短距离无线通信模块或全球导航卫星系统(GNSS)通信模块)或有线通信模块(例如,局域网(LAN)通信模块或电力线通信(PLC)模块)。无线通信设备可以通过短距离通信网络(例如,蓝牙™、无线保真(Wi-Fi)直接或红外数据协会(IrDA))或长距离通信网络(例如,传统蜂窝网络、5G网络、下一代通信网络、互联网)、或计算机网络(例如,LAN或广域网(WAN))与外部设备进行通信。

[0114] 在示例中,电子设备600可以是使用基于神经网络的算法的AR提供装置(或设备)(例如,AR眼镜)。AR提供装置可以佩戴在用户的脸上以向用户提供与AR服务和/或虚拟现实(VR)服务相关的内容。处理器610可以使用具有目标架构的经训练的神经网络来执行处理操作。相机630可以捕捉用于产生AR内容的图像,并且处理器610可以通过使用具有目标架构的神经网络处理图像来产生AR内容。例如,处理器610可以通过识别通过相机630获得的图像中的对象并将虚拟内容叠加在所识别的对象区域或对象周围的区域上来产生AR内容。

[0115] 处理器610可以通过参考图2和图5所描述的相同过程来确定神经网络的目标架构。例如,处理器610可以基于用于神经网络(例如,图1中的基本神经网络120)的候选架构的参数来确定神经网络损失,并且当候选架构的神经网络操作时,可以测量所需的硬件资

源。处理器610可以使用硬件资源预测模块(例如,图4中的硬件资源预测模块420)来预测当候选架构的神经网络操作时所需的硬件资源,并且基于测量的硬件资源和预测的硬件资源来确定硬件资源损失。处理器610可以基于神经网络损失和硬件资源损失来确定目标架构。处理器610可以根据基于候选架构参数的神经网络损失和基于候选架构参数的硬件资源损失的加权和来确定优化损失,并且可以确定使优化损失最小化的目标架构和候选架构的参数。

[0116] 计算装置200、处理器210、处理器610、以及本文所描述的其他装置、设备、单元、模块和组件由硬件组件来实现。在适当的情况下可用于执行本申请中所描述的操作的硬件组件的示例包括控制器、传感器、生成器、驱动器、存储器、比较器、算术逻辑单元、加法器、减法器、乘法器、除法器、积分器、以及被配置为执行本申请所描述的操作的任何其他电子组件。在其他示例中,用于执行本申请中所描述的操作的一个或多个硬件组件由计算硬件(例如,由一个或多个处理器或计算机)实现。处理器或计算机可以由一个或多个处理元件(例如,逻辑门阵列、控制器和算术逻辑单元、数字信号处理器、微计算机、可编程逻辑控制器、现场可编程门阵列、可编程逻辑阵列、微处理器、或被配置为以定义的方式响应并执行指令以实现期望的结果的任何其他设备或设备的组合)来实现。在一个示例中,处理器或计算机包括(或者,连接到)存储由处理器或计算机执行的指令或软件的一个或多个存储器。由处理器或计算机实现的硬件组件可以执行指令或软件,例如,操作系统(OS)和在OS上运行的一个或多个软件应用,以执行本申请中描述的操作。硬件组件还可以响应于执行指令或软件来访问、操纵、处理、创建和存储数据。为了简洁起见,在本申请中描述的示例的描述中可以使用单数术语“处理器”或“计算机”,但是在其他示例中可以使用多个处理器或计算机,或者处理器或计算机可以包括多个处理元件、或多种类型的处理元件、或两者兼有。例如,单个硬件组件或者两个或更多个硬件组件可以由单个处理器、或者两个或更多个处理器、或者处理器和控制器来实现。一个或多个硬件组件可以由一个或多个处理器、或者处理器和控制器来实现,并且一个或多个其他硬件组件可以由一个或多个其他处理器、或者另一处理器和另一控制器来实现。一个或多个处理器、或者处理器和控制器可以实现单个硬件组件、或者两个或更多个硬件组件。硬件组件可以具有任何一种或多种不同的处理配置,其示例包括单处理器、独立处理器、并行处理器、单指令单数据(SISD)多处理、单指令多数据(SIMD)多处理、多指令单数据(MISD)多处理、多指令多数据(MIMD)多处理、控制器和算术逻辑单元(ALU)、DSP、微型计算机、专用集成电路(ASIC))、现场可编程门阵列(FPGA)、可编程逻辑单元(PLU)、中央处理单元(CPU)、图形处理单元(GPU)、神经处理单元(NPU)、或任何其他能够以定义的方式响应并执行指令的设备。

[0117] 执行本申请中所描述的操作的方法由计算硬件执行,例如,由执行指令或软件的如上所述地实现的一个或多个处理器或计算机执行,以执行本申请中所描述的通过这些方法执行的操作。例如,单个操作或者两个或更多个操作可以由单个处理器、或者两个或更多个处理器、或者处理器和控制器来执行。一个或多个操作可以由一个或多个处理器、或者处理器和控制器来执行,并且一个或多个其他操作可以由一个或多个其它处理器、或者另一处理器和另一控制器来执行。一个或多个处理器、或者处理器和控制器可以执行单个操作、或者两个或更多个操作。

[0118] 用于控制处理器或计算机如上所述地实现硬件组件并执行所述方法的指令或软

件被写为计算机程序、代码段、指令或其任何组合，用于单独地或共同地指示或配置处理器或计算机作为机器或专用计算机来操作，以执行由硬件组件执行的操作和上述方法。在一个示例中，指令或软件包括由处理器或计算机直接执行的机器代码，例如由编译器产生的机器代码。在示例中，指令或软件包括小程序、动态链接库(DLL)、中间件、固件、设备驱动程序、存储用于搜索神经网络的最佳架构的方法的应用程序中的至少一种。在另一示例中，指令或软件包括由处理器或计算机使用解释器执行的高级代码。本领域的普通程序员能够基于附图中所示的框图和流程图以及说明书中的对应描述来容易地编写指令或软件，其中公开了用于执行由硬件组件和如上所述的方法执行的操作的算法。

[0119] 控制处理器或计算机实现硬件组件并执行如上所述的方法的指令或软件、以及任何相关联数据、数据文件和数据结构被记录、存储或固定在一个或多个非暂时性计算机可读存储介质之中或之上。非暂时性计算机可读存储介质的示例包括只读存储器(ROM)、随机存取可编程只读存储器(PROM)、电可擦可编程只读存储器(EEPROM)、随机存取存储器(RAM)、磁RAM(MRAM)、自旋转移矩(STT)-MRAM、静态随机存取存储器(SRAM)、晶闸管RAM(T-RAM)、零电容RAM(Z-RAM)、双晶体管RAM(TTRAM)、导电桥接RAM(CBRAM)、铁电RAM(FeRAM)、相变RAM(PRAM)、电阻RAM(RRAM)、纳米管RRAM、聚合物RAM(PoRAM)、纳米浮栅存储器(NFGM)、全息存储器、分子电子存储器件)、绝缘体电阻变化存储器、动态随机存取存储器(DRAM)、静态随机存取存储器(SRAM)、闪存、非易失性存储器、CD-ROM、CD-R、CD+R、CD-RW、CD+RW、DVD-ROM、DVD-R、DVD+R、DVD-RW、DVD+RW、DVD-RAM、BD-ROM、BD-R、BD-RLTH、BD-RE、蓝光或光盘存储设备、硬盘驱动器(HDD)、固态驱动器(SSD)、闪存、卡类型的存储器(比如，多媒体卡或微型卡(例如，安全数字(SD)或极限数字(XD)))、磁带、软盘、磁光数据存储设备、光学数据存储设备、硬盘、固态盘、以及被如下配置的任何其它设备：以非暂时性方式存储指令或软件以及任何相关联的数据、数据文件和数据结构，并且向处理器或计算机提供指令或软件以及任何相关联的数据、数据文件和数据结构，使得处理器或计算机可以执行该指令。在示例中，指令或软件以及任何相关联的数据、数据文件和数据结构分布在联网的计算机系统上，使得一个或多个处理器或计算机以分布方式存储、访问和执行该指令和软件以及任何相关联的数据、数据文件和数据结构。

[0120] 尽管本公开包括特定示例，但是在理解了本申请的公开内容之后将显而易见的是，在不脱离权利要求及其等同物的精神和范围的情况下，可以对这些示例进行形式和细节上的各种改变。本文中描述的示例应仅被认为是描述性的，而不是为了限制的目的。每个示例中的特征或方面的描述被认为适用于其他示例中的类似特征或方面。如果所描述的技术以不同的顺序执行和/或如果所描述的系统、架构、设备或电路中的组件以不同的方式组合和/或被其他组件或其等同物替换或补充，则可以实现合适的结果。因此，本公开的范围不是由具体实施方式来限定，而是由权利要求及其等同物来限定，并且在权利要求及其等同物的范围内的所有变化都被解释为包括在本公开中。

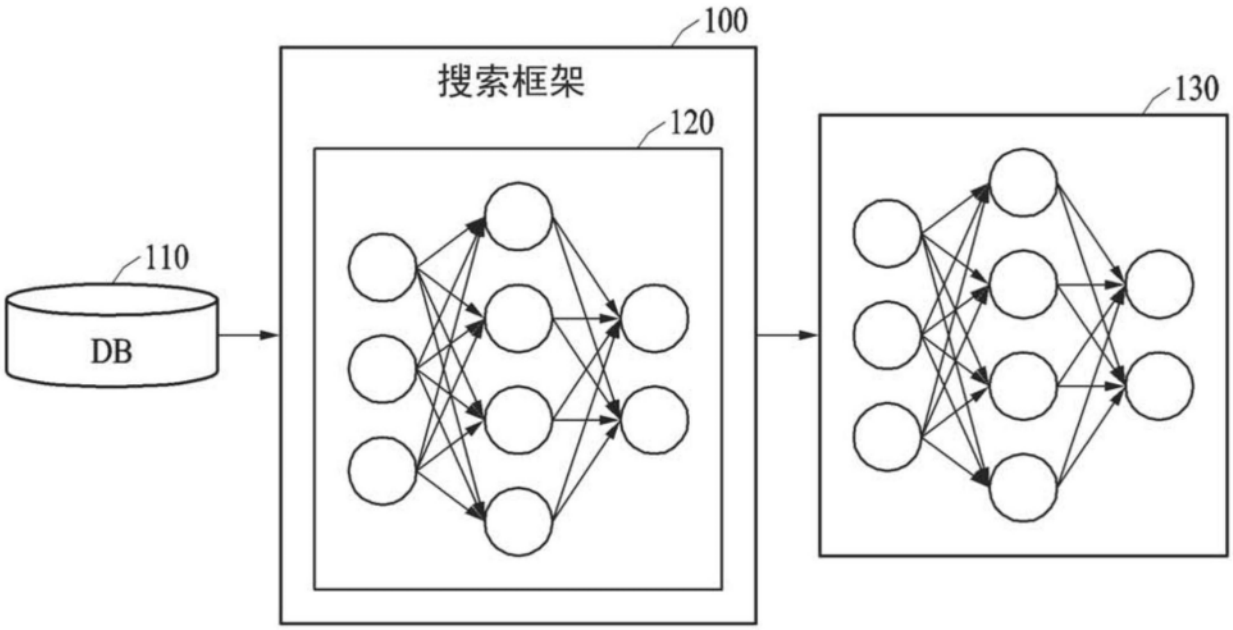


图1

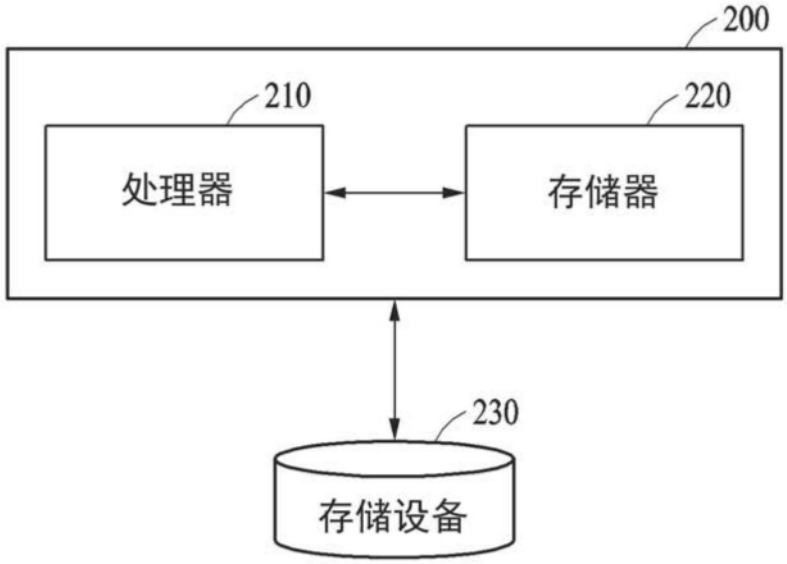


图2

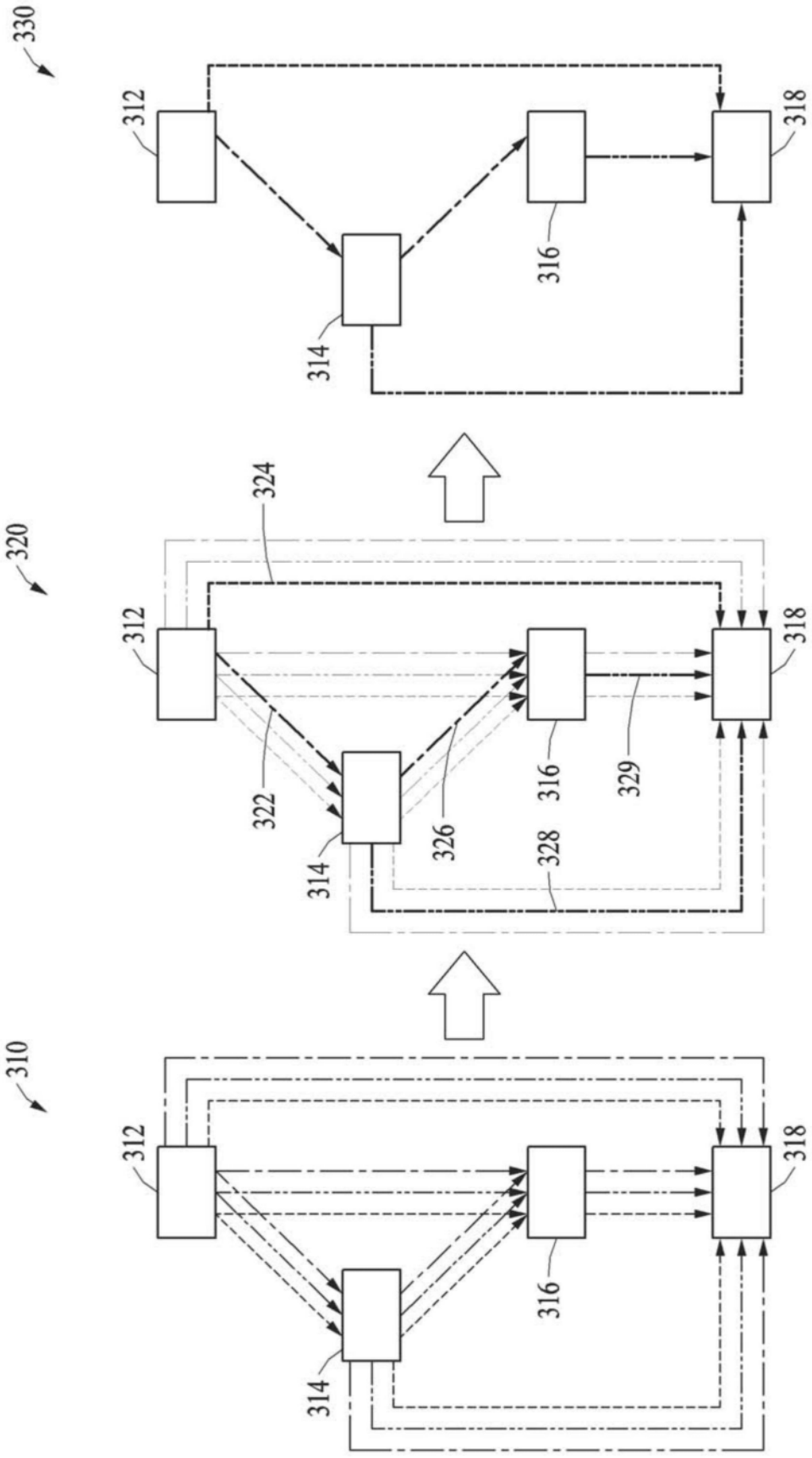


图3

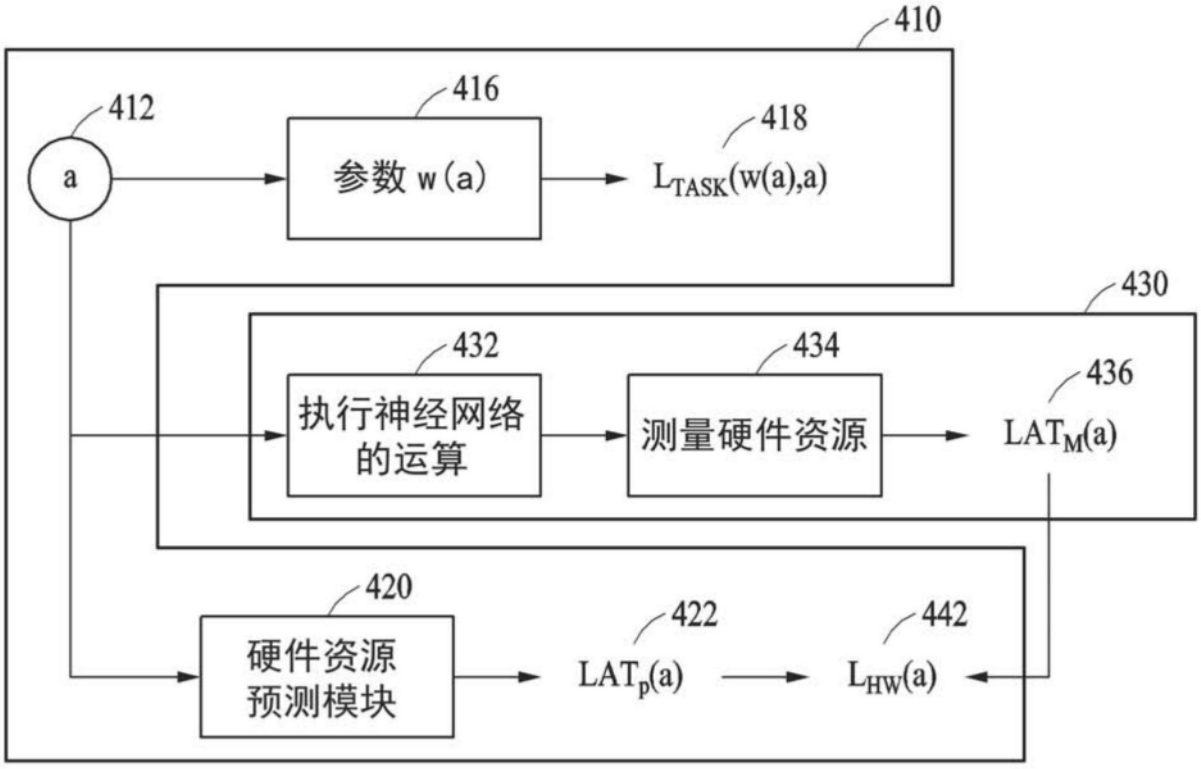

410
412
a
416
参数 w(a)
418
$L_{TASK}(w(a),a)$
430
432
执行神经网络的运算
434
测量硬件资源
436
$LAT_M(a)$
420
硬件资源预测模块
422
$LAT_p(a)$
442
$L_{HW}(a)$

图4

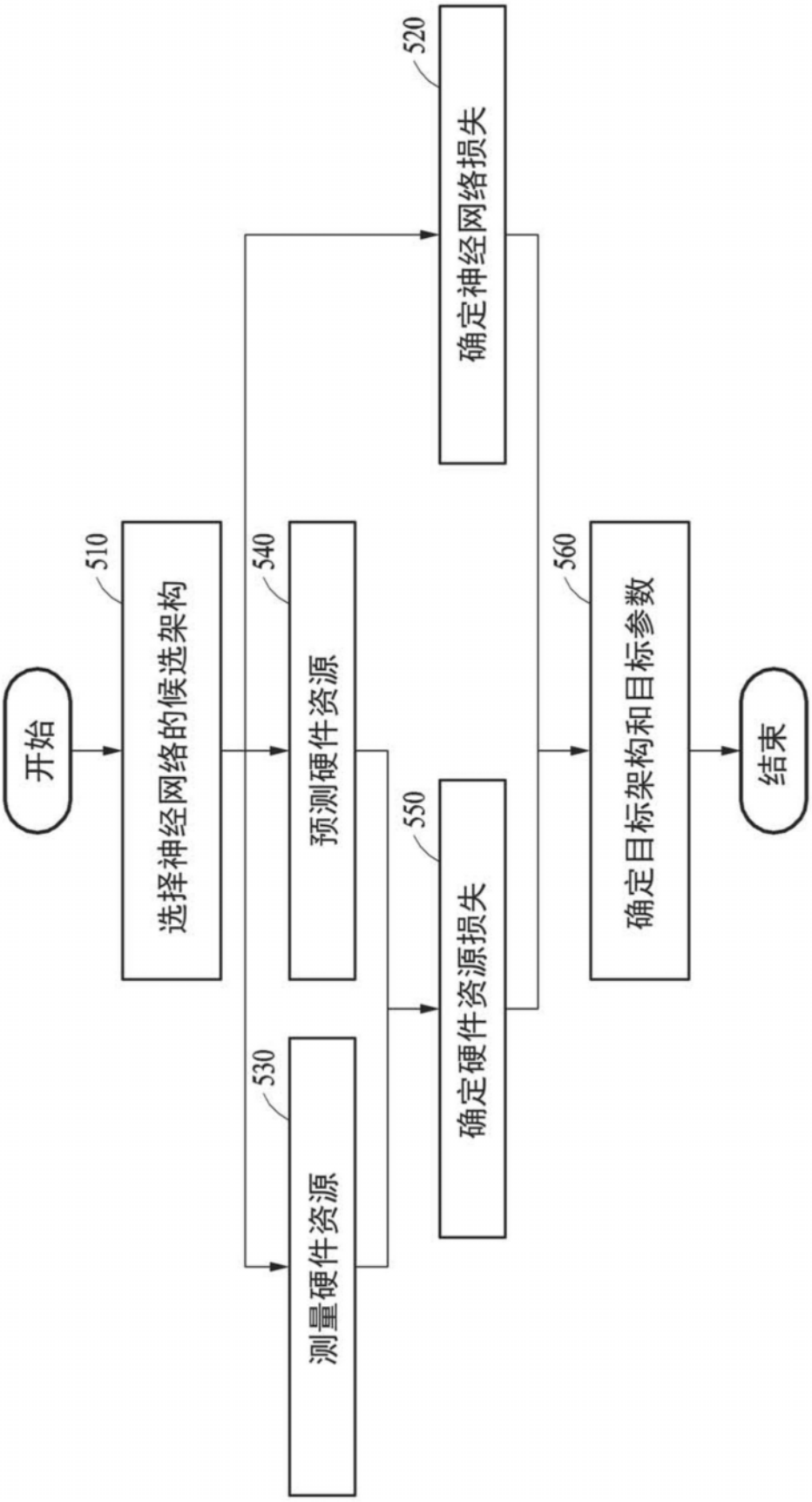

开始
510
选择神经网络的候选架构
520
确定神经网络损失
540
预测硬件资源
530
测量硬件资源
550
确定硬件资源损失
560
确定目标架构和目标参数
结束

图5

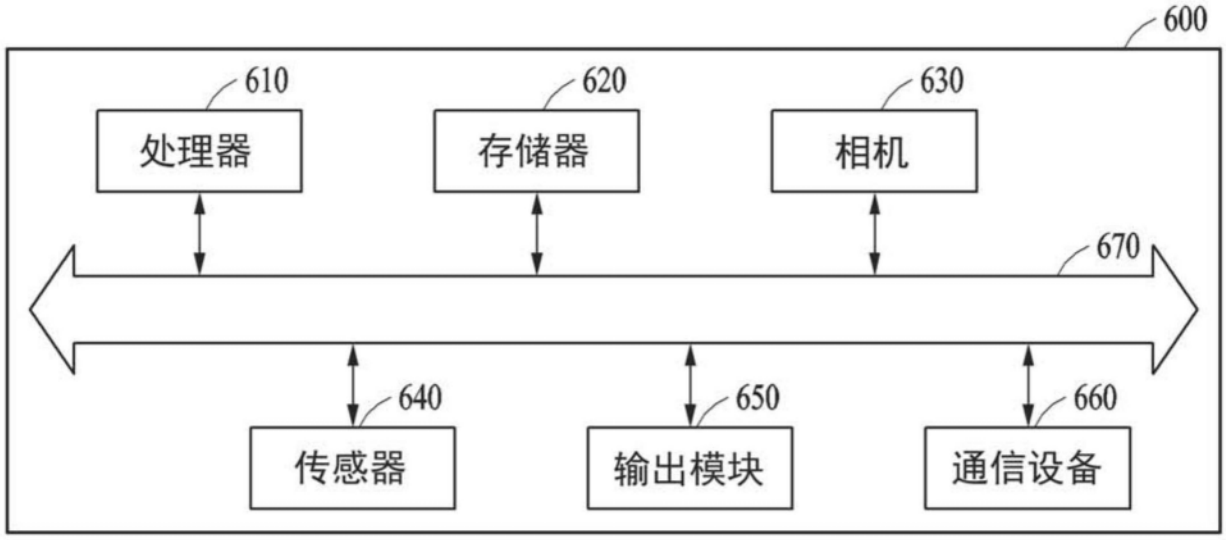


图6