(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 111406267 A
(43)申请公布日 2020.07.10

(21)申请号 201880076005.9

(22)申请日 2018.11.30

(30)优先权数据
62/593,213 2017.11.30 US

(85)PCT国际申请进入国家阶段日
2020.05.25

(86)PCT国际申请的申请数据
PCT/US2018/063293 2018.11.30

(87)PCT国际申请的公布数据
WO2019/108923 EN 2019.06.06

(71)申请人 谷歌有限责任公司
地址 美国加利福尼亚州

(72)发明人 W.华 B.佐夫 J.什伦斯 刘晨曦 J.黄 李佳 F-F.李 K.P.墨菲

(74)专利代理机构 北京市柳沈律师事务所 11105
代理人 金玉洁

(51)Int.Cl.
G06N 3/08(2006.01)
G06N 3/04(2006.01)

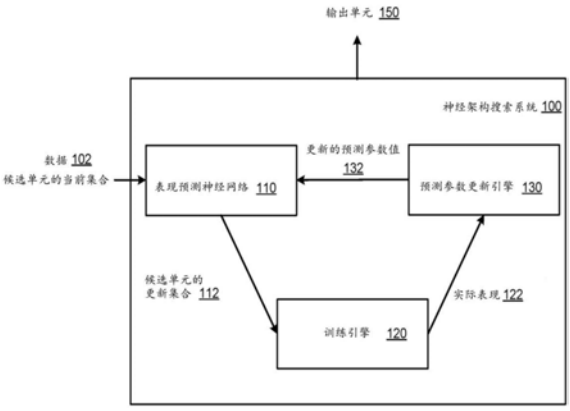
权利要求书2页 说明书10页 附图4页

(54)发明名称

使用性能预测神经网络的神经架构搜索

(57)摘要

描述了一种用于确定配置为执行特定的机器学习任务的任务神经网络的架构的方法。该方法包括:获得数据,该数据指定用于任务神经网络的候选架构的当前集合;对于当前集合中的每个候选架构:使用具有多个性能预测参数的性能预测神经网络来处理指定候选架构的数据,性能预测神经网络被配置为根据性能预测参数的当前值处理指定候选架构的数据以生成性能预测,该性能预测表征具有候选架构的神经网络在关于特定的机器学习任务的训练之后将执行得有多好;以及通过基于当前集合中的候选架构的性能预测来选择当前集合中的一个或多个候选架构来生成候选架构的更新集合。



CN 111406267 A

1.一种由一个或多个计算机执行的方法,该方法包括:

确定被配置为执行特定的机器学习任务的任务神经网络的架构,包括:

获得数据,该数据指定任务神经网络的候选架构的当前集合;

对于当前集合中的每个候选架构:

使用具有多个性能预测参数的性能预测神经网络来处理指定候选架构的数据,其中,性能预测神经网络被配置为根据性能预测参数的当前值处理指定候选架构的数据,以生成性能预测,该性能预测表征具有候选架构的神经网络在关于特定的机器学习任务的训练之后将执行得有多好;以及

通过基于对于当前集合中的候选架构的性能预测选择当前集合中的一个或多个候选架构来生成候选架构的更新集合。

2.根据权利要求1所述的方法,其中,特定的机器学习任务包括图像处理。

3.根据权利要求1或权利要求2所述的方法,其中,特定的机器学习任务包括图像分类或视频分类。

4.根据权利要求1所述的方法,其中,特定的机器学习任务包括语音识别。

5.根据前述权利要求中的任一项所述的方法,还包括:

训练具有所确定的架构的任务神经网络;以及

使用具有所确定的架构的训练后的任务神经网络对接收到的网络输入执行特定的机器学习任务。

6.根据权利要求1至5中的任一项所述的方法,还包括:

对于更新集合中的每个候选架构:

生成具有候选架构的任务神经网络的实例;

训练所述实例以执行特定的机器学习任务;以及

评估训练后的实例关于特定的机器学习任务的性能,以确定训练后的实例的实际性能;以及

使用训练后的实例的实际性能来调整性能预测神经网络的性能预测参数的当前值。

7.根据权利要求6所述的方法,还包括:

对于更新集合中的每个候选架构:

通过针对每个新候选架构将一个或多个运算的相应集合添加到候选架构来从候选架构生成多个新候选架构;

对于每个新候选架构:

使用性能预测神经网络并根据性能预测参数的更新值来处理指定新候选架构的数据,以生成对于新候选架构的性能预测;以及

通过基于新候选架构的性能预测选择一个或多个新候选架构来生成候选架构的新集合。

8.根据权利要求7所述的方法,还包括:

在新集合中选择新候选架构中的一个作为任务神经网络的架构。

9.根据权利要求8所述的方法,其中,所述选择包括:

对于新集合中的每个新候选架构:

生成具有新候选架构的任务神经网络的实例;

训练所述实例以执行特定的机器学习任务;以及

评估训练后的实例关于特定的机器学习任务的性能,以确定训练后的实例的实际性能;以及

选择与具有最佳实际性能的训练后的实例相对应的新候选架构作为用于任务神经网络的架构。

10.根据权利要求7-9中的任一项所述的方法,其中,任务神经网络的架构包括每个共享一个或多个超参数的多个卷积单元,每个卷积单元包括一个或多个运算块,每个运算块接收一个或多个相应的输入隐藏状态并生成相应的输出隐藏状态,以及其中,每个候选架构和每个新候选架构定义用于由每个卷积单元所共享的超参数的值。

11.根据权利要求10所述的方法,其中,每个候选架构定义用于具有第一数量的运算块的卷积单元的架构,以及其中,通过针对每个新候选架构将一个或多个运算的相应集合添加到候选架构来从候选架构生成多个新候选架构包括:

针对每个新候选单元将具有相应超参数的新运算块添加到候选架构。

12.根据权利要求1-11中的任一项所述的方法,其中,指定候选架构的数据是定义候选架构的嵌入序列,以及其中,性能预测神经网络是循环神经网络。

13.根据权利要求12所述的方法,其中,性能预测是循环神经网络的在处理所述序列中的最后一个嵌入之后的输出。

14.一种系统,包括存储指令的一个或多个计算机和一个或多个存储设备,该指令当由所述一个或多个计算机运行时使所述一个或多个计算机执行根据权利要求1-13中的任一项所述的相应方法的操作。

15.一种或多种存储指令的计算机存储介质,该指令当由一个或多个计算机运行时使所述一个或多个计算机执行根据权利要求1-13中的任一项所述的相应方法的操作。

使用性能预测神经网络的神经架构搜索

[0001] 相关申请的交叉引用

[0002] 本申请是2017年11月30日提交的美国临时专利申请第62/593,213号的非临时申请并要求享有其优先权,该临时专利申请的全部内容通过引用合并于此。

背景技术

[0003] 本说明书涉及确定神经网络的架构。

[0004] 神经网络是机器学习模型,其采用一层或多层非线性单元来预测接收到的输入的输出。一些神经网络除了包括输出层之外还包括一个或多个隐藏层。每个隐藏层的输出用作网络中下一个层(即下一个隐藏层或输出层)的输入。网络的每一层根据参数的相应集合的当前值从接收到的输入生成输出。

[0005] 一些神经网络是循环神经网络。循环神经网络是接收输入序列并从输入序列生成输出序列的神经网络。特别地,循环神经网络可以在计算当前时间步长处的输出时使用网络的来自先前时间步长的内部状态的一些或全部。循环神经网络的示例是包括一个或多个长短期记忆(LSTM)存储器块的LSTM神经网络。每个LSTM存储器块可以包括一个或多个单元,每个单元包括输入门、遗忘门和输出门,输入门、遗忘门和输出门允许该单元存储该单元的先前状态,以例如用于生成当前激活或被提供给LSTM神经网络的其他组件。

发明内容

[0006] 本说明书描述了一种系统,该系统被实现为在一个或多个位置中的一个或多个计算机上的计算机程序,该系统确定配置为执行特定的机器学习任务的任务神经网络的网络架构。

[0007] 本说明书中描述的主题可以在特定实施例中实现,从而实现以下优点中的一个或多个。通过使用本规范中描述的技术来确定任务神经网络的架构,系统可以确定在各种机器学习任务(例如图像分类或另一图像处理任务)中的任何一个上达到或甚至超过最先进性能的网络架构。另外,该系统可以以比现有技术在计算上效率高得多(即比现有技术消耗更少的计算资源)的特定方式确定任务神经网络的架构(例如,确定在任务神经网络的整个架构中重复的输出单元)。特别地,许多现有技术依赖于通过训练具有候选架构的网络来评估大量候选架构的性能。这种训练既耗时又是计算密集的。所描述的技术通过改为采用性能预测神经网络而不需要实际训练具有候选架构的网络极大地减少了任务神经网络的需要训练的实例数量,该性能预测神经网络有效地预测具有候选架构的训练后的网络的性能。在一些描述的实施方式中,此方法与其他资源节约型方法(即有效地限制最终输出架构的可能架构的搜索空间而不会不利地影响并且在一些情况下甚至改善包括输出架构的多个实例的所得任务神经网络的性能的技术)相结合,以实现更高的计算效率。例如,其他资源节约型方法可以包括学习卷积单元或包含多个运算块的其他类型的单元的架构、然后根据预定的模板重复所学习的单元以生成任务神经网络的架构。

[0008] 本说明书中描述的主题的一个或多个实施例的细节在附图和以下描述中阐述。所

述主题的其他特征、方面和优点将由说明书、附图和权利要求变得明显。

附图说明

[0009] 图1示出了示例神经架构搜索(NAS)系统的架构。

[0010] 图2示出了任务神经网络的示例单元的架构。

[0011] 图3示出了示例任务神经网络的架构。

[0012] 图4是用于确定输出单元的架构的示例过程的流程图。

[0013] 各个附图中同样的附图标记和标号指示同样的元件。

具体实施方式

[0014] 本说明书描述了被实现为一个或多个位置中的一个或多个计算机上的计算机程序的神经架构搜索系统，该神经架构搜索系统确定用于任务神经网络的网络架构。任务神经网络被配置为执行特定的机器学习任务。

[0015] 一般，任务神经网络被配置为接收网络输入并处理该网络输入以生成对于该输入的网络输出。

[0016] 在一些情况下，任务神经网络是卷积神经网络，其被配置为接收输入图像并处理该输入图像以生成对于该输入图像的网络输出，即执行某种图像处理任务。

[0017] 例如，任务可以是图像分类，并且由神经网络针对给定图像生成的输出可以是对于对象类别的集合中的每个的得分，其中每个得分代表该图像包含属于该类别的对象的图像的所估计的可能性。

[0018] 作为另一示例，任务可以是图像嵌入生成，并且由神经网络生成的输出可以是输入图像的数字嵌入。

[0019] 作为又一示例，任务可以是对象检测，并且由神经网络生成的输出可以识别输入图像中的在此处绘出特定类型的对象的位置。

[0020] 在另一些情况下，任务可以是视频分类，并且任务神经网络被配置为接收视频或视频的一部分作为输入，并生成确定输入的视频或视频部分涉及什么样的一个主题或多个主题的输出。

[0021] 在另一些情况下，任务可以是语音识别，并且任务神经网络被配置为接收音频数据作为输入，并生成针对给定的所讲出的话语确定该话语代表的一个或多个术语的输出。

[0022] 在另一些情况下，任务可以是文本分类，并且任务神经网络被配置为接收输入文本片段，并生成确定输入文本片段涉及什么样的一个主题或多个主题的输出。

[0023] 图1示出了示例神经架构搜索(NAS)系统100。神经架构搜索系统100是被实现为在一个或多个位置中的一个或多个计算机上的计算机程序的系统的示例，下面描述的系统、组件和技术可以在计算机程序中实现。

[0024] 在一些实施方式中，NAS系统100被配置为通过确定在整个网络架构中重复的输出单元150的架构，来确定用于任务神经网络的网络架构。即，任务神经网络包括输出单元150的多个实例。基于任务神经网络内的实例的位置，输出单元150的实例内的卷积运算的过滤器的数量可以不同。在一些情况下，任务神经网络包括输出单元150的多个实例的堆叠。在一些情况下，除了输出单元的堆叠之外，任务神经网络还包括一个或多个其他神经网络层，

例如输出层和/或一种或多种其他类型的层。例如，任务神经网络可以包括卷积神经网络层，其后是输出单元的多个实例的堆叠，其后是全局池化神经网络层，其后是softmax分类神经网络层。下面参照图3更详细地描述任务神经网络的示例架构。

[0025] 一般，单元是被配置为接收单元输入并生成单元输出的全卷积神经网络。在一些实施方式中，单元输出可以具有与单元输入相同的尺寸，例如，相同的高度(H)、宽度(W)和深度(F)。例如，单元可以接收特征图作为输入，并生成具有与输入的特征图相同尺寸的输出特征图。在另一些实施方式中，单元输出可以具有与单元输入的尺寸不同的尺寸。例如，当单元是步幅为2的全卷积神经网络时，假定单元输入是H×W×F张量，则单元输出可以是H'×W'×F'张量，其中H'＝H/2，W'＝W/2，F'＝2F。

[0026] 在一些情况下，单元包括B个运算块，其中B是预定的正整数。例如，B可以是三、五或十。单元中的每个运算块接收一个或多个相应的输入隐藏状态，并将一个或多个运算应用于输入隐藏状态以生成相应的输出隐藏状态。

[0027] 在一些实施方式中，B个运算块中的每个被配置为将第一运算应用于到该运算块的第一输入隐藏状态，以生成第一输出。该运算块被配置为将第二运算应用于到该运算块的第二输入隐藏状态以生成第二输出。然后，该运算块被配置为将组合运算应用于第一输出和第二输出，以生成对于该运算块的输出隐藏状态。可以通过与运算块相关联的超参数集合来定义第一输入隐藏状态、第二输入隐藏状态、第一运算、第二运算和组合运算。例如，与运算块相对应的超参数集合包括以下超参数：表示哪个隐藏状态用作第一输入隐藏状态的第一超参数、表示哪个隐藏状态用作第二输入隐藏状态的第二超参数、表示哪个运算用作第一运算的第三超参数、表示哪个运算用作第二运算的第四超参数、以及表示哪个运算用作组合运算以组合第一运算和第二运算的输出的第五超参数。

[0028] 下面参照图2更详细地描述单元的示例架构。

[0029] 为了确定输出单元150的架构，NAS系统100包括具有多个性能预测参数(在本说明书中也称为“预测参数”)的性能预测神经网络110(也称为“预测器110”)。预测器110是包括一个或多个循环神经网络层的循环神经网络。例如，预测器110可以是长短期记忆(LSTM)神经网络或门控循环单元(GRU)神经网络。

[0030] 一般，预测器110被配置为接收指定候选单元的数据并根据预测参数处理该数据以生成性能预测，该性能预测表征具有该候选单元的神经网络对于特定的机器学习任务的训练之后将执行得如何。指定候选单元的数据是定义候选单元的嵌入序列(例如，多组超参数的嵌入，其中每组超参数定义了候选单元中包括的相应运算块)。在本说明书中使用的嵌入是超参数的数字表示，例如，向量或数值的其他有序集合。嵌入可以被预先确定或作为训练预测器的部分被学习。

[0031] 性能预测可以是例如对训练后的神经网络的准确度的预测。作为另一示例，性能预测可以包括预测的平均准确度以及准确度的预测的标准偏差或方差两者。

[0032] 特别地，作为确定在整个任务神经网络的网络架构中重复的输出单元150的架构的部分，NAS系统100获得指定用于输出单元150的候选单元的当前集合的数据102。在一些情况下，候选单元的当前集合是候选单元的初始集合。在另一些情况下，NAS系统100从先前的迭代获得单元，然后通过扩展每个先前的单元，例如通过将相应的一个或多个运算块添加到每个先前的单元，来生成候选单元的当前集合。

[0033] 对于当前集合中的每个候选单元，预测器110接收指定该候选单元的数据，并根据性能预测参数的当前值使用性能预测神经网络110来处理该数据，以生成针对每个候选单元的性能预测。

[0034] 然后，通过基于针对当前集合中的候选单元的性能预测来选择当前集合中的一个或多个候选单元，NAS系统110生成候选单元的更新集合(updatedset)112。即，NAS系统110基于由性能预测神经网络110生成的预测来修剪当前集合，以生成更新集合。例如，NAS系统110从当前集合中选择具有最佳性能预测的K个候选单元，以包括在更新集合112中，其中，K是预定的整数。

[0035] 为了更新预测器110的性能预测参数的值，NAS系统110包括训练引擎120和预测参数更新引擎130。一般，训练引擎120和预测参数更新引擎130将被实现为在一个或多个位置中的一个或多个计算机上安装的一个或多个软件模块或组件。在一些情况下，一个或多个计算机将专用于特定的引擎；在其他情况下，可以在相同的一个计算机或多个计算机上安装并运行多个引擎。

[0036] 对于更新集合中的每个候选单元，训练引擎120被配置为生成具有该候选单元的任务神经网络的实例，并训练该实例以执行特定的机器学习任务。例如，训练引擎120根据任务神经网络的预定模板架构来生成任务神经网络的实例。例如，任务神经网络的模板架构包括第一神经网络层(例如卷积层)，其后是单元的N个实例的堆叠，其后是输出子网络(例如，包括softmax神经网络层的输出子网络)。

[0037] 为了训练任务神经网络的实例，训练引擎120获得用于关于特定机器学习任务来训练实例的训练数据和用于评估任务神经网络的训练后的实例关于该特定机器学习任务的性能的验证集。

[0038] 训练引擎120可以接收用于以各种方式中的任何一种训练实例的数据。例如，在一些实施方式中，训练引擎120例如通过使用NAS系统100可用的应用编程接口(API)经由数据通信网络从NAS系统100的远程用户接收作为上载(upload)的训练数据。

[0039] 训练引擎120评估每个训练后的实例关于特定机器学习任务的性能，以确定该训练后的实例的实际性能122。例如，实际性能可以是如通过适当的准确度度量所测量的训练后的实例关于验证集的准确度。例如，当任务是分类任务时，准确度可以是分类错误率，或者当任务是回归任务时，准确度可以是交并比(intersection over union)差异度量。作为另一示例，实际性能可以是针对实例训练的最后两个、五个或十个时期(epoch)中的每个，实例的准确度的平均值或最大值。

[0040] 预测参数更新引擎130使用训练后的实例的实际性能来调整性能预测神经网络110的性能预测参数的值。特别地，预测参数更新引擎130通过使用常规监督学习技术(例如随机梯度下降(SGD)技术)训练预测器110准确地预测候选单元的实际性能来调整预测参数的值。

[0041] 通过使用预测器110来为当前集合中的每个候选单元生成性能预测，NAS系统110考虑当前集合中的所有候选单元。然而，NAS系统110仅需要实际训练少量候选单元，即，基于由预测器110生成的性能预测选择以包括在更新集合中的那些候选单元。因此，NAS系统110定义一种特定的技术实施方式，其比依赖于通过实际训练具有候选单元的网络来评估大量候选单元的性能的现有系统在计算上更有效率(即，消耗更少的计算资源)。这是因为

训练任务神经网络的实例与仅使用预测器110预测其实际性能相比，在计算上要昂贵得多。此外，在一些实施方式中，可以并行地训练和评估由预测器选择以包括在更新集合中的候选单元，因而允许NAS系统100比传统系统更快地确定输出单元。

[0042] 在更新预测器110的预测参数之后，NAS系统100扩展更新集合中的候选单元以生成包括多个新候选单元的新集合。特别地，NAS系统100通过针对更新集合中的每个候选单元将具有超参数的相应集合的新运算块添加到该候选单元来扩展更新集合中的候选单元。

[0043] 一般，假定更新集合具有N个候选单元，其中每个候选单元具有b个运算块，则NAS系统100针对更新集合中的每个特定的候选单元生成所有可能的单元的子集，其中每个可能的单元具有b+1个运算块(即，通过将新的第(b+1)个运算块添加到该特定的候选单元)。新集合是具有b+1个运算块的所有可能单元的子集的组合。

[0044] 在一些实施方式中，可以通过5个超参数(I_1,I_2,O_1,O_2,C)来指定新的第(b+1)个运算块，其中$I_1, I_2 \in \mathcal{I}_{b+1}$指定新运算块的输入，$\mathcal{I}_{b+1}$是新运算块的可能输入的集合；$O_1, O_2 \in \mathcal{O}$指定分别应用于输入$I_1$和$I_2$的运算，其中$\mathcal{O}$是预定的运算空间；$C \in \mathcal{C}$指定如何组合$O_1$和$O_2$以为新运算块生成块输出$H^c_{b+1}$，其中$\mathcal{C}$是可能的组合运算符的集合。

[0045] 在这些实施方式中，第(b+1)个运算块的可能结构的搜索空间为B_{b+1}，其大小为$|B_{b+1}| = |\mathcal{I}_{b+1}|^2 \times |\mathcal{O}|^2 \times |\mathcal{C}|^2$，其中$|\mathcal{I}_{b+1}|=2+(b+1)-1$，$|\mathcal{O}|$是运算空间中的运算的数量，$|\mathcal{C}|$是集合$\mathcal{C}$中的组合运算符的数量。因此，新集合中候选单元的数量为$N \times ||B_{b+1}|$个单元。

[0046] 然后，NAS系统100将候选单元的新集合设置成候选单元的当前集合，并重复上述过程，直到候选单元具有预定的最大数量的运算块。

[0047] 当每个候选单元中的运算块的数量等于预定的最大数量的运算块时，NAS系统100选择与具有最佳实际性能的训练后的实例相对应的新候选单元作为任务神经网络的输出单元150。

[0048] 在一些实施方式中，一旦确定了输出单元150，系统100将指定输出单元的架构的数据例如经由网络提供给用户设备。代替提供指定架构的数据或除了提供指定架构的数据之外，系统100例如从头开始或微调作为训练更大的神经网络的结果而生成的参数值来训练具有确定的输出单元150的神经网络，然后使用训练后的神经网络来处理由用户例如通过NAS系统100所提供的API接收的请求。

[0049] 尽管本说明书描述了搜索在整个任务神经网络中重复多次的单元的可能架构的空间，但是在另一些实施方式中，NAS系统100例如通过除了一个或多个预定输出层以及可选地一个或多个预定的输入层以外的整个任务神经网络的可能架构来搜索不重复的一部分架构。

[0050] 图2示出了可以用于构建任务神经网络的示例单元200的架构。

[0051] 单元200是被配置为处理单元输入(例如H×W×F张量)以生成单元输出(例如，H'×W'×F'张量)的全卷积神经网络。

[0052] 在一些实施方式中，例如，当单元200是步幅为1的完全卷积神经网络时，单元输出可以具有与单元输入相同的尺寸(例如，H'＝H，W'＝W，以及F'＝F)。在另一些实施方式中，单元输出可以具有与单元输入的尺寸不同的尺寸。例如，当单元是步幅为2的全卷积神经网

络时，假定单元输入是H×W×F张量，则单元输出可以是H'×W'×F'张量，其中H'＝H/2，W'＝W/2，以及F'＝2F。

[0053] 单元200包括多个运算块(B个块)。例如，如图2所示，单元200包括5个块：块202、204、206、208和210。

[0054] 单元200中的每个块b可以由5个超参数(I_1,I_2,O_1,O_2,C)指定，其中$I_1, I_2 \in \mathcal{I}_b$指定到块b的输入；$O_1, O_2 \in \mathcal{O}$分别指定应用于输入$I_1$和$I_1$的运算，其中$\mathcal{O}$是运算空间；以及$C \in \mathcal{C}$指定如何组合$O_1$和$O_2$来生成对于块b的块输出$H_b^c$，其中$\mathcal{C}$是可能的组合运算符的集合。

[0055] 可能输入的集合$\mathcal{I}_b$是单元200中所有先前块的集合$\{H_b^c, \ldots, H_{b-1}^c\}$加上前一个单元的输出$H_B^{c-1}$加上前一个单元之前的单元的输出$H_B^{c-2}$。

[0056] 运算空间$\mathcal{O}$可以包括但不限于以下运算：3×3深度可分离卷积、5×5深度可分离卷积、7×7深度可分离卷积、1×7其后是7×1卷积、标识、3×3平均池化、3×3最大池化以及3×3扩张卷积。

[0057] 在一些实施方式中，可能的组合运算符$\mathcal{C}$的集合包括加法运算和连接(concatenation)运算。

[0058] 在一些实施方式中，可能的组合运算符$\mathcal{C}$的集合仅包括加法运算。在这些实施方式中，单元200的每个块b可以由4个超参数(I_1,I_2,O_1,O_2)指定。

[0059] 在每个块b生成块输出之后，所有块的块输出被组合，例如被连接、求和或平均，以生成单元200的单元输出H^c。

[0060] 图3示出了示例任务神经网络300的架构。任务神经网络300被配置为接收网络输入302并生成对于输入302的网络输出320。

[0061] 任务神经网络300包括单元实例的堆叠306。堆叠306包括单元的多个实例，所述多个实例一个接一个地堆叠。堆叠306中的单元实例可以具有相同的结构但是具有不同的参数值。堆叠306中的单元实例内的卷积运算的过滤器的数量可以基于堆叠内实例的位置而不同。例如，在一个实施方式中，单元实例308是步幅为2的单元，单元实例310是步幅为1的单元。在这样的实施方式中，单元实例308具有两倍于单元实例310所具有的过滤器。

[0062] 堆叠306中的第一单元实例308被配置为接收第一单元输入并处理第一单元输入以生成第一单元输出。

[0063] 在一些情况下，第一单元输入是任务神经网络的网络输入302。

[0064] 在另一些情况下，网络输入302是图像，并且任务神经网络300可以在单元堆叠306之前包括卷积神经网络层304，以便减少与处理图像相关联的计算成本。例如，卷积神经网络层304是步幅为2的3×3卷积过滤器层。在这些情况下，卷积神经网络层304被配置为处理网络输入302以生成中间输出从而作为第一单元输入被提供给单元实例308。

[0065] 在第一单元实例之后的每个单元实例(例如，单元实例310-312)被配置为接收前一个单元实例的单元输出作为输入并生成相应的单元输出，该相应的单元输出作为输入被馈送到下一个单元实例。堆叠306的输出是最后一个单元实例314的单元输出。

[0066] 任务神经网络300包括在单元实例的堆叠306之后的子网络316。子网络316被配置

为接收单元实例的堆叠306的输出作为输入并处理堆叠306的输出以生成网络输出320。作为示例，子网络316包括全局池化神经网络层，其后是softmax分类神经网络层。

[0067] 图4是用于确定在整个任务神经网络中重复的单元的架构的示例过程400的流程图。为了方便起见，将过程400描述为由位于一个或多个位置的一个或多个计算机的系统执行。例如，根据本说明书适当编程的神经架构搜索系统(例如图1的神经架构搜索系统100)可以执行过程400。

[0068] 该系统获得指定用于构建单元神经网络的输出单元的候选单元的当前集合的数据(步骤402)。

[0069] 在一些情况下，候选单元的当前集合是候选单元的初始集合。在另一些情况下，系统从先前的迭代获得单元，然后通过扩展每个先前的单元(例如，通过将相应的一个或多个运算块添加到每个先前的单元)来生成候选单元的当前集合。

[0070] 系统使用具有多个性能预测参数的性能预测神经网络来处理指定候选单元的数据(步骤404)。性能预测神经网络被配置为根据性能预测参数的当前值来处理指定候选单元的数据，以生成性能预测，该性能预测表征具有候选单元的神经网络在关于特定机器学习任务训练之后将执行得有多好。

[0071] 系统通过基于当前集合中的候选单元的性能预测选择当前集合中的一个或多个候选单元来生成候选单元的更新集合(步骤406)。即，系统基于由性能预测神经网络生成的预测来修剪当前集合，以生成更新集合。例如，系统从当前集合中选择具有最佳性能预测的K个候选单元以包括在更新集合中，其中K是预定的整数。

[0072] 系统针对当前集合中的每个候选单元如下迭代地执行步骤408-412。

[0073] 系统生成具有候选单元的任务神经网络的实例(步骤408)。例如，系统根据任务神经网络的预定模板架构来生成任务神经网络的实例。例如，任务神经网络的模板架构包括第一神经网络层(例如卷积层)，其后是单元的N个实例的堆叠，其后是输出子网络(例如，包括softmax神经网络层的输出子网络)。

[0074] 系统训练实例以执行特定的机器学习任务(步骤410)。

[0075] 为了训练任务神经网络的实例，系统获得关于特定的机器学习任务的用于训练实例的训练数据和用于评估任务神经网络的训练后的实例关于该特定的机器学习任务的性能的验证集。然后，系统使用常规机器学习训练技术关于训练数据来训练实例。

[0076] 然后，系统例如通过测量训练后的实例关于验证数据集的准确度来评估每个训练后的实例关于特定机器学习任务的性能以确定训练后的实例的实际性能(步骤412)。

[0077] 一旦系统针对当前集合中的所有候选单元重复了步骤408-412，系统使用训练后的实例的实际性能来调整性能预测神经网络的性能预测参数的值(步骤414)。

[0078] 特别地，系统通过训练性能预测神经网络来调整预测参数的值，以使用常规监督学习技术(例如随机梯度下降(SGD)技术)来准确地预测候选单元的实际性能。

[0079] 然后，系统确定更新集合中的每个候选单元中的运算块的数量是否小于单元中允许的运算块的预定最大数量(步骤416)。

[0080] 当新集合中的每个新候选单元中的运算块的数量小于单元中允许的运算块的预定最大数量时，系统扩展更新集合中的候选单元以生成候选单元的新集合。特别地，系统通过针对更新集合中的每个候选单元将具有超参数的相应集合的相应的新运算块添加到该

候选单元来扩展更新集合中的候选单元。然后，系统将候选单元的该新集合设置为候选单元的当前集合，并重复步骤402-416，直到每个候选单元中的运算块的数量等于运算块的最大数量。

[0081] 当更新集合中的每个候选单元中的运算块的数量等于运算块的预定最大数量时，系统选择与具有最佳实际性能的训练后的实例相对应的新候选单元作为在整个任务神经网络的架构中重复的输出单元(步骤418)。

[0082] 本说明书结合系统和计算机程序组件使用术语“配置”。使一个或多个计算机的系统被配置为执行特定的操作或动作意思是该系统已在其上安装了在操作中使该系统执行所述操作或动作的软件、固件、硬件或其组合。使一个或多个计算机程序被配置为执行特定的操作或动作意思是该一个或多个程序包括指令，该指令当由数据处理装置运行时使该装置执行所述操作或动作。

[0083] 本说明书中描述的主题和功能操作的实施例可以在数字电子电路中、在有形地体现的计算机软件或固件中、在计算机硬件(包括本说明书中公开的结构及其结构等同物)中、或在它们中的一个或多个的组合中实现。本说明书中描述的主题的实施例可以被实现为一个或多个计算机程序，即，在有形的非暂时性存储介质上编码的计算机程序指令的一个或多个模块，以由数据处理装置运行或控制数据处理装置的操作。该计算机存储介质可以是机器可读存储设备、机器可读存储基板、随机或串行访问存储器设备或它们中的一个或多个的组合。可替代地或另外地，程序指令可以被编码在人为生成的传播信号(例如机器生成的电信号、光信号或电磁信号)上，生成该人为生成的传播信号来对信息进行编码用于到合适的接收器装置的传输以由数据处理装置运行。

[0084] 术语“数据处理装置”是指数据处理硬件，并涵盖用于处理数据的各种装置、设备和机器，包括例如一个可编程处理器、一个计算机或多个处理器或多个计算机。该装置还可以是或可以进一步包括专用逻辑电路，例如，FPGA(现场可编程门阵列)或ASIC(专用集成电路)。除了硬件之外，该装置还可以可选地包括为计算机程序创建运行环境的代码，例如，构成处理器固件、协议栈、数据库管理系统、操作系统或它们中的一个或多个的组合的代码。

[0085] 也可称为或描述为程序、软件、软件应用、app、模块、软件模块、脚本或代码的计算机程序可以以任何形式的编程语言来编写，包括编译或解释语言、或声明性或过程型语言；它可以以任何形式进行部署，包括作为独立程序或作为模块、组件、子例程或适用于计算环境的其他单元。程序可以但不必对应于文件系统中的文件。程序可以存储在文件的保存其他程序或数据(例如存储在标记语言文档中的一个或多个脚本)的部分中、在专用于所讨论的程序的单个文件中或在多个协调文件(例如，存储一个或多个模块、子程序或部分代码的文件)中。可以将计算机程序部署为在位于一个站点上或分布在多个站点上并通过数据通信网络互连的一个计算机或多个计算机上运行。

[0086] 在本说明书中，术语“数据库”被广泛地用于指代任何数据集合：该数据不需要以任何特定的方式来结构化或者根本不需要结构化，并且可以将其存储在一个或多个位置中的存储设备中。因此，例如，索引数据库可以包括多个数据集合，每个数据集合可以被不同地组织和访问。

[0087] 类似地，在本说明书中，术语“引擎”广泛地用于指代被编程以执行一个或多个特定功能的基于软件的系统、子系统或过程。一般，引擎将被实现为安装在一个或多个位置中

的一个或多个计算机上的一个或多个软件模块或组件。在一些情况下，一个或多个计算机将专用于特定引擎；在另一些情况下，可以在相同的一个计算机或多个计算机上安装并运行多个引擎。

[0088] 本说明书中描述的过程和逻辑流程可以由运行一个或多个计算机程序以通过对输入数据进行操作并生成输出来执行功能的一个或多个可编程计算机执行。所述过程和逻辑流程还可以由专用逻辑电路(例如FPGA或ASIC)执行，或由专用逻辑电路和一个或多个编程计算机的组合执行。

[0089] 适合于运行计算机程序的计算机可以基于通用微处理器或专用微处理器或它们两者，或者基于任何其他种类的中央处理单元。一般，中央处理单元将从只读存储器或随机存取存储器或它们两者接收指令和数据。计算机的基本元件是用于执行或运行指令的中央处理单元以及用于存储指令和数据的一个或多个存储器设备。中央处理单元和存储器可以由专用逻辑电路补充或并入专用逻辑电路中。一般，计算机还将包括用于存储数据的一个或多个大容量存储设备(例如，磁盘、磁光盘或光盘)，或可操作地联接以从所述一个或多个大容量存储设备接收数据或将数据传输到所述一个或多个大容量存储设备，或包括所述一个或多个大容量存储设备并且可操作地联接以从所述一个或多个大容量存储设备接收数据或将数据传输到所述一个或多个大容量存储设备。然而，计算机不必具有这样的设备。此外，计算机可以嵌入另一个设备中，仅举几例，例如，移动电话、个人数字助理(PDA)、移动音频或视频播放器、游戏主机、全球定位系统(GPS)接收器或便携式存储设备(例如通用串行总线(USB)闪存驱动器)。

[0090] 适合于存储计算机程序指令和数据的计算机可读介质包括所有形式的非易失性存储器、介质和存储器设备，包括例如半导体存储器设备(例如EPROM、EEPROM和闪存设备)；磁盘(例如内部硬盘或可移动磁盘)；磁光盘；以及CD ROM和DVD-ROM磁盘。

[0091] 为了提供与用户的交互，本说明书中描述的主题的实施例可以在具有用于向用户显示信息的显示器设备(例如CRT(阴极射线管)或LCD(液晶显示器)监视器)以及用户可通过其向计算机提供输入的键盘和定点设备(例如鼠标或轨迹球)的计算机上实现。其他种类的设备也可以用于提供与用户的交互；例如，提供给用户的反馈可以是任何形式的感觉反馈，例如视觉反馈、听觉反馈或触觉反馈；并且可以以任何形式接收来自用户的输入，包括声音输入、语音输入或触觉输入。另外，计算机可以通过向用户使用的设备发送文档以及从用户使用的设备接收文档(例如，通过响应从网络浏览器收到的请求而将网页发送到用户设备上的网络浏览器)来与用户交互。而且，计算机可以通过将文本消息或其他形式的消息发送到个人设备(例如运行消息收发应用的智能手机)并转而从用户接收回答消息来与用户交互。

[0092] 用于实现机器学习模型的数据处理装置还可以包括例如专用硬件加速器单元，用于处理机器学习训练或生产(即推理)工作量的通用部分和计算密集部分。

[0093] 可以使用机器学习框架(例如TensorFlow框架、Microsoft Cognitive Toolkit框架、Apache Singa框架或Apache MXNet框架)来实现和部署机器学习模型。

[0094] 本说明书中描述的主题的实施例可以在包括后端组件(例如，作为数据服务器)的计算系统、或包括中间件组件(例如应用服务器)的计算系统、或包括前端组件(例如具有图形用户界面、网页浏览器、或用户可通过其与本说明书中描述的主题的实现方式交互的app

的客户端计算机)的计算系统、或包括一个或多个这样的后端组件、中间件组件或前端组件的任何组合的计算系统中实现。系统的组件可以通过数字数据通信的任何形式或介质(例如通信网络)互连。通信网络的示例包括局域网(LAN)和广域网(WAN)(例如因特网)。

[0095] 计算系统可以包括客户端和服务器。客户端和服务器一般彼此远离,并且通常通过通信网络交互。客户端和服务器之间的关系是由于计算机程序在各个计算机上运行并彼此具有客户端-服务器关系而产生。在一些实施例中,例如出于向充当客户端的与设备交互的用户显示数据并从该用户接收用户输入的目的,服务器将数据(例如HTML页面)发送给用户设备。在用户设备处生成的数据(例如用户交互的结果)可以在服务器处从该设备接收。

[0096] 虽然本说明书包含许多特定的实现细节,但是这些细节不应被解释为对任何发明的范围或所要求保护的内容的范围的限制,而应被解释为对特定发明的特定实施例可能特有的特征的描述。在分开的实施例的上下文中在本说明书中描述的某些特征也可以在单个实施例中组合地实现。相反,在单个实施例的上下文中描述的各种特征也可以分开地在多个实施例中或以任何合适的子组合来实现。此外,尽管以上可能将特征描述为以某些组合起作用,甚至最初也这样要求进行保护,但在一些情况下,可以从组合中去除来自所要求保护的组合的一个或多个特征,并且所要求保护的组合可以涉及子组合或子组合的变体。

[0097] 类似地,虽然按特定的顺序在附图中示出了操作并在权利要求中陈述了所述操作,但这不应被理解为要求按示出的特定顺序或按先后顺序执行这些操作、或要求执行所有图示的操作以实现期望的结果。在某些情况下,多任务和并行处理可以是有利的。此外,上述实施例中的各种系统模块和组件的分离不应被理解为在所有实施例中都要求这样的分离,并且应理解,所描述的程序组件和系统一般可以被一起集成在单个软件产品中或被打包到多个软件产品中。

[0098] 已经描述了主题的特定实施例。其他实施例在所附权利要求的范围内。例如,权利要求中陈述的动作可以按不同的顺序执行并且仍然实现期望的结果。作为一个示例,附图中绘出的过程不一定要求示出的特定顺序或先后顺序来实现期望的结果。在一些情况下,多任务和并行处理可以是有利的。

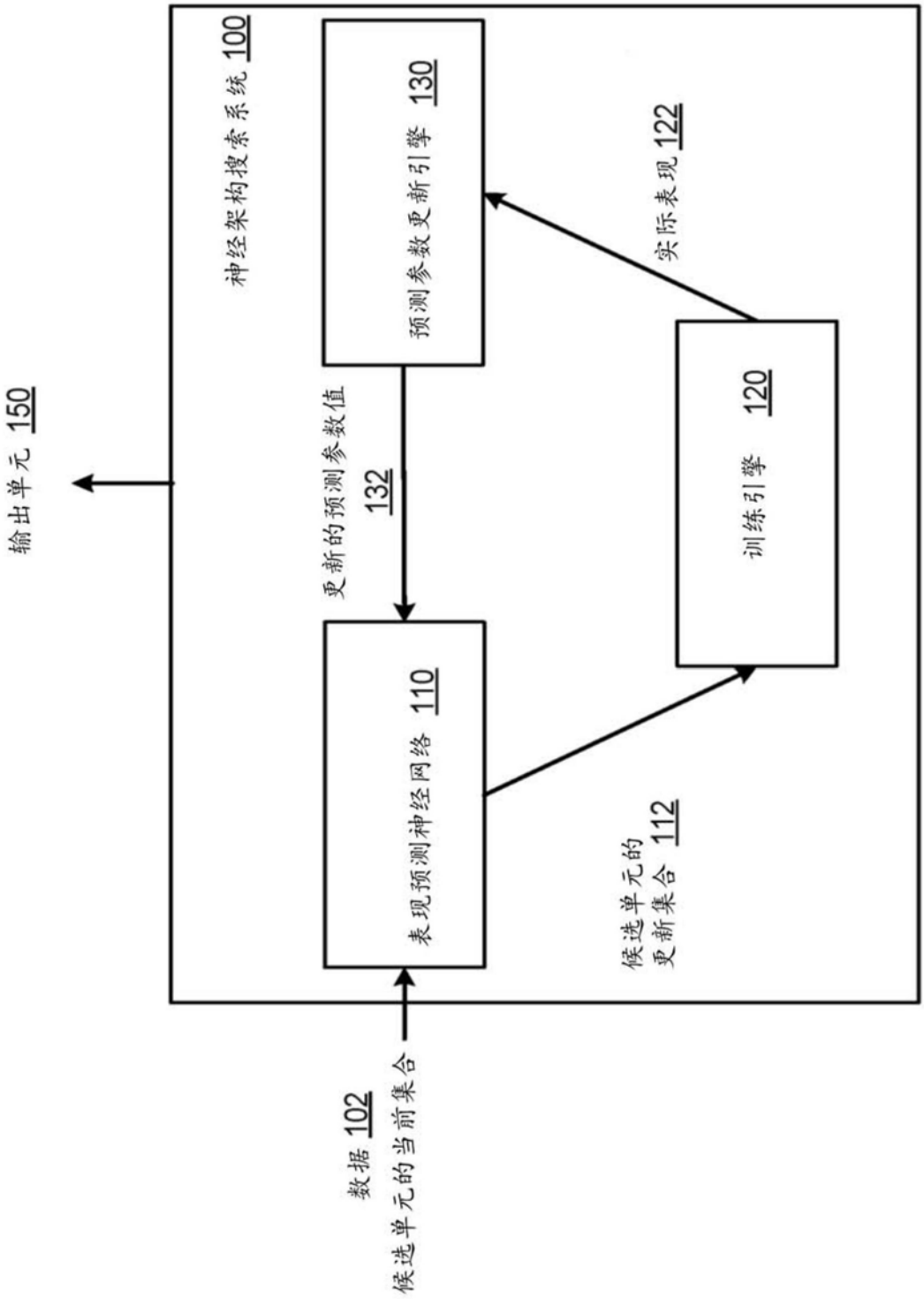

神经架构搜索系统 100
预测参数更新引擎 130
实际表现 122
输出单元 150
更新的预测参数值 132
训练引擎 120
表现预测神经网络 110
候选单元的更新集合 112
数据 102
候选单元的当前集合

图1

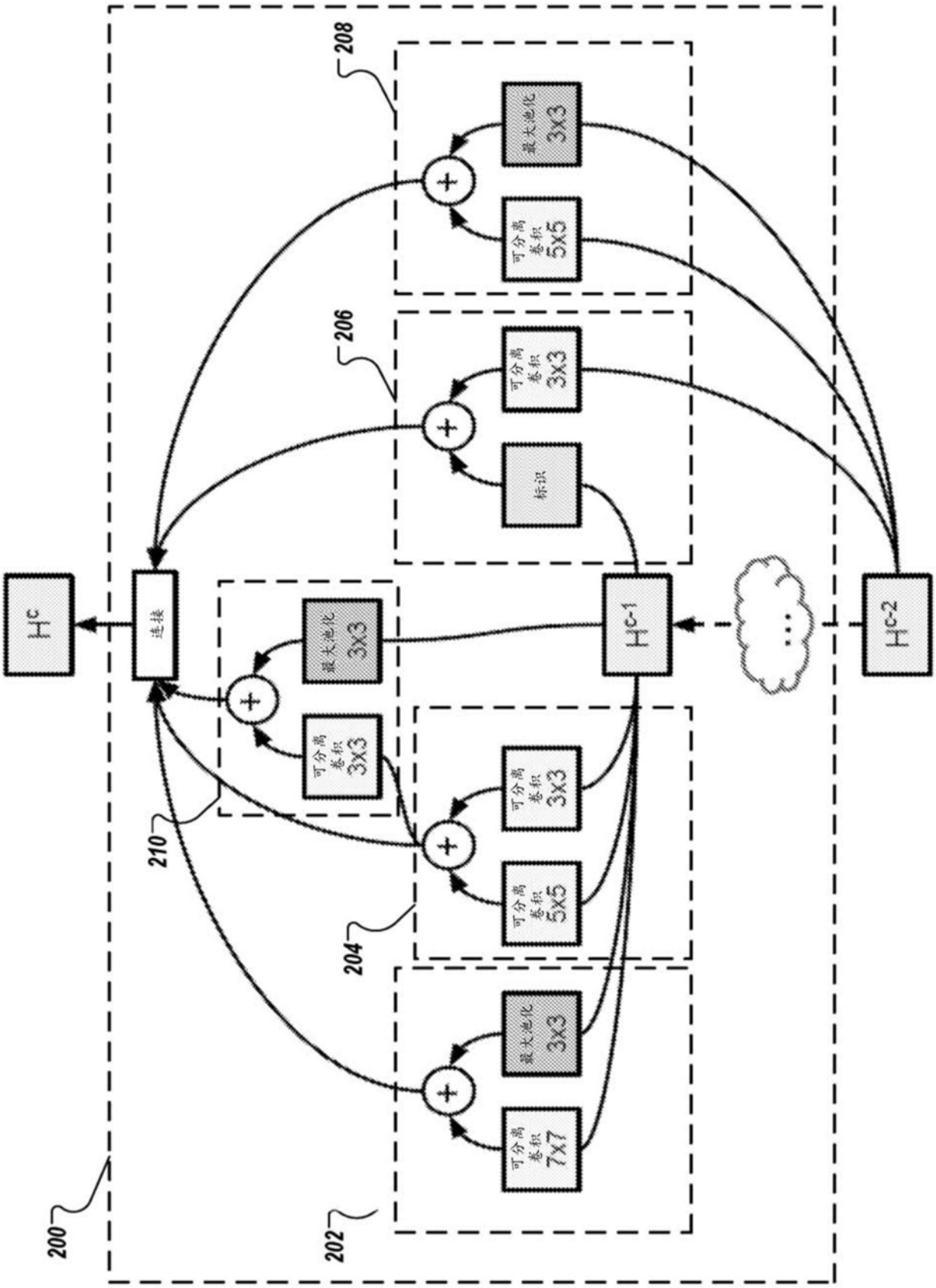

208
206
204
202
200
210
H^c
连接
H^{c-1}
H^{c-2}
最大池化 3x3
可分离卷积 5x5
可分离卷积 3x3
标识
可分离卷积 7x7

图2

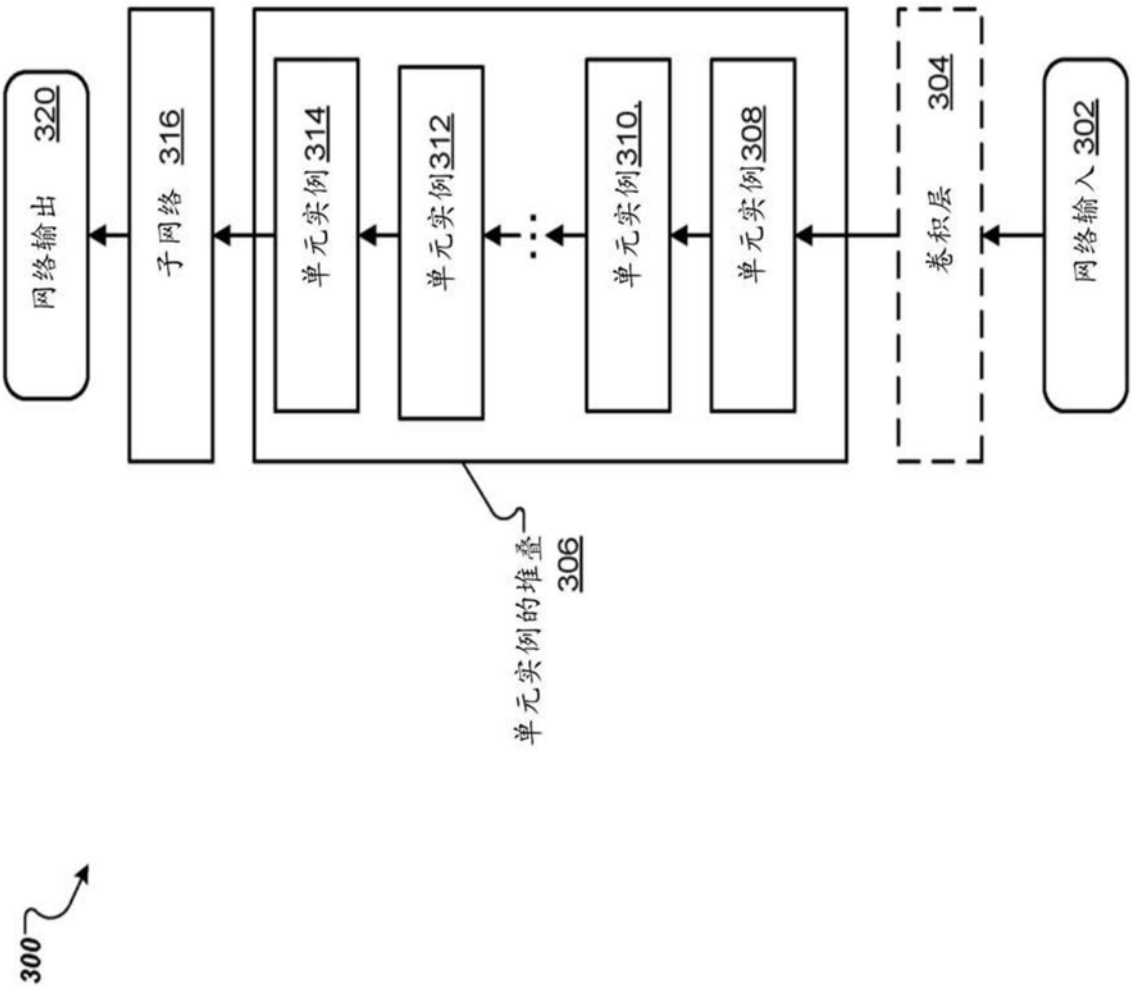

网络输出 320
子网络 316
单元实例 314
单元实例312
单元实例310
单元实例308
卷积层 304
网络输入 302
单元实例的堆叠 306
300

图3

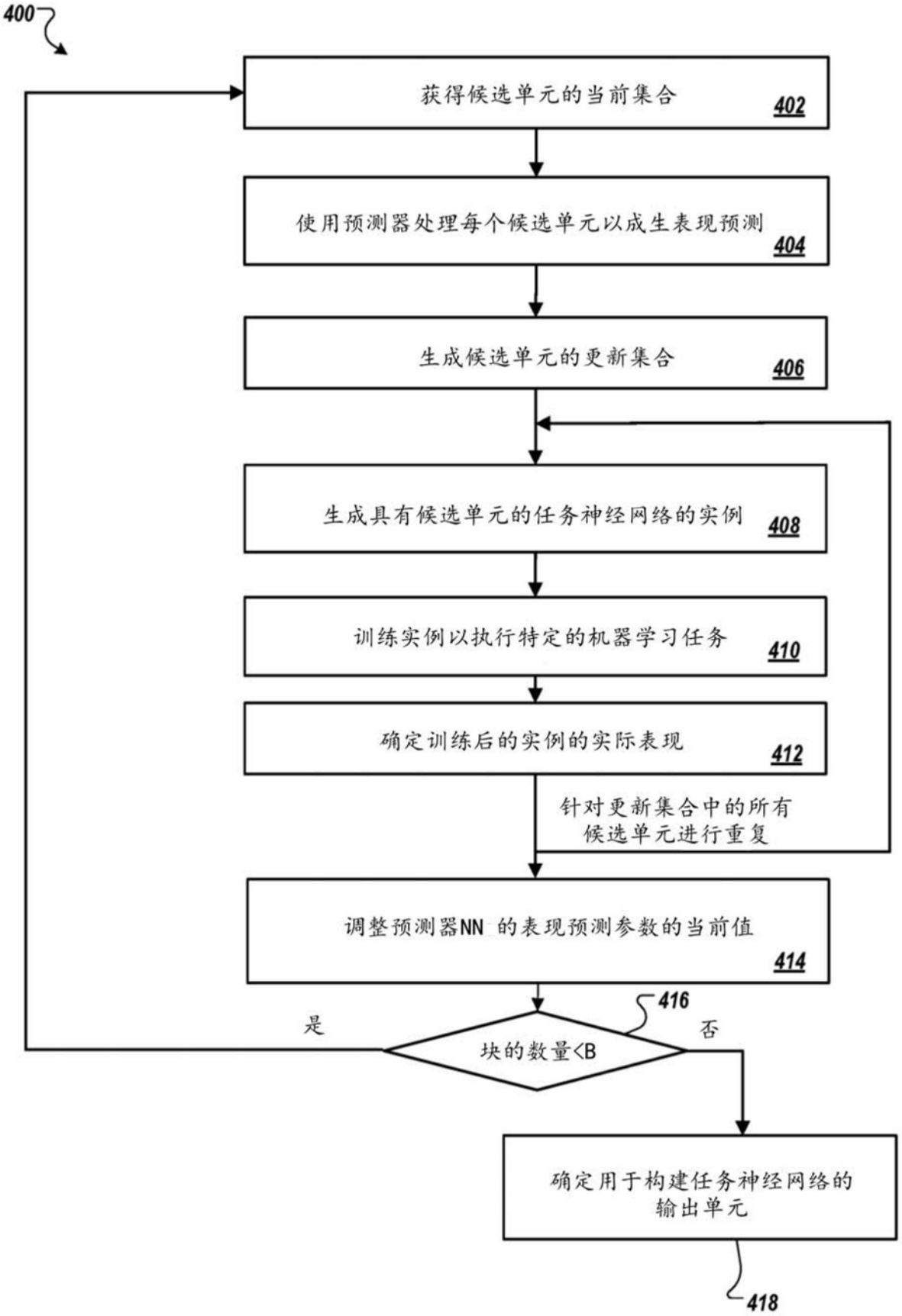

400
获得候选单元的当前集合
402
使用预测器处理每个候选单元以成生表现预测
404
生成候选单元的更新集合
406
生成具有候选单元的任务神经网络的实例
408
训练实例以执行特定的机器学习任务
410
确定训练后的实例的实际表现
412
针对更新集合中的所有候选单元进行重复
调整预测器NN 的表现预测参数的当前值
414
416
是
块的数量<B
否
确定用于构建任务神经网络的输出单元
418

图4