



(12) 发明专利

(10) 授权公告号 CN 114492779 B

(45) 授权公告日 2024.09.27

(21) 申请号 202210141278.5

G06F 7/483 (2006.01)

(22) 申请日 2022.02.16

G06N 3/0495 (2023.01)

(65) 同一申请的已公布的文献号

申请公布号 CN 114492779 A

(56) 对比文件

US 2018230011 A1, 2018.08.16

WO 2023231794 A1, 2023.12.07

CN 116662593 A, 2023.08.29

WO 2023059215 A1, 2023.04.13

CN 116543419 A, 2023.08.04

CN 117291275 A, 2023.12.26

(43) 申请公布日 2022.05.13

(73) 专利权人 安谋科技(中国)有限公司

地址 200233 上海市闵行区田林路1016号

科技绿洲三期11号楼

He, Y等. Quantization and pruning optimization method attention mechanism. 《Journal of National university of defense technology》. 2024, 第46卷(第1期), 113-20.

(72) 发明人 韩冥生 余宗桥 鲁若荻

陈东等. 深度学习中的模型压缩技术. 《人工智能》. 2023, (第03期), 40-51.

(74) 专利代理机构 上海华诚知识产权代理有限公司 31300

专利代理师 肖华

审查员 雷冬

(51) Int. Cl.

G06N 3/063 (2023.01)

G06N 3/0464 (2023.01)

G06F 16/22 (2019.01)

G06F 7/499 (2006.01)

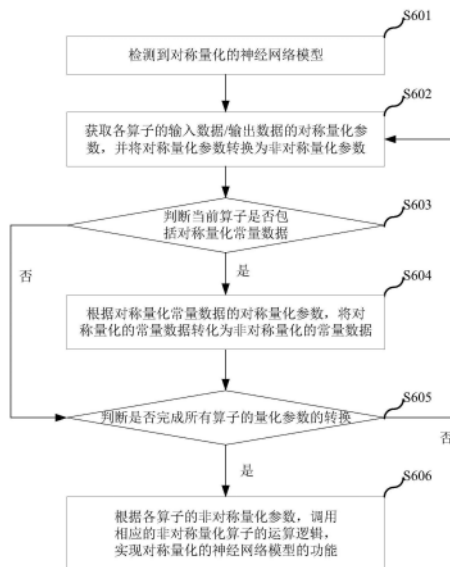
权利要求书2页 说明书19页 附图6页

(54) 发明名称

神经网络模型的运行方法、可读介质和电子设备

(57) 摘要

本申请涉及人工智能领域,公开了一种神经网络模型的运行方法、可读介质和电子设备。该方法应用于电子设备,包括:检测到第一神经网络模型,第一神经网络模型为对称量化的神经网络模型,并且第一神经网络模型中包括第一神经网络模型各算子的对称量化参数;将各算子的对称量化参数转换为非对称量化参数;利用各算子的非对称量化参数,调用预设的非对称量化的算子的运算逻辑,得到第一神经网络模型的推理结果。如此,只能运行非对称量化的神经网络模型的电子设备,可以调用预设的非对称量化的算子的运算逻辑来实现对称量化的神经网络模型的功能,增加了电子设备能够运行的神经网络模型的类型。



1. 一种神经网络模型的运行方法,应用于电子设备,其特征在于,包括:

将图像数据输入第一神经网络模型,所述第一神经网络模型为对称量化的神经网络模型,并且所述第一神经网络模型中包括所述第一神经网络模型的各算子的对称量化参数;

其中,所述对称量化参数包括:常量数据的对称量化参数,其中所述常量数据的对称量化参数包括对称量化常量数据、对称量化常量数据的对称量化缩放系数,所述对称量化常量数据还包括对称量化的查找表,所述对称量化的查找表中包括对称量化的查表索引和各对称量化的查表索引对应的对称量化的查表结果;

将各所述算子的对称量化参数转换为非对称量化参数;

其中,所述将各所述算子的对称量化参数转换为非对称量化参数,包括:

根据所述对称量化常量数据的对称量化缩放系数,确定出各所述对称量化的查表索引对应的浮点数查表索引;

根据各所述对称量化的查表索引对应的浮点数查表索引和各所述算子的运算逻辑,确定各所述浮点数查表索引对应的浮点数查表结果;

根据所述浮点数查表索引的非对称量化缩放系数和非对称量化零点得到非对称量化的查表索引、根据所述浮点数据查表结果的非对称量化缩放系数和非对称量化零点得到对称量化的查表结果,其中,所述浮点数查表索引的非对称量化缩放系数和非对称量化零点基于所述对称量化的查表索引的数据类型确定、所述浮点数查表结果的非对称量化缩放系数和非对称量化零点基于所述对称量化的查表结果的数据类型确定;

基于各所述非对称量化的查表索引和相对应的非对称量化的查表结果,得到非对称量化的查找表;

利用各所述算子的非对称量化参数,调用预设的非对称量化的算子的运算逻辑,得到所述第一神经网络模型的推理结果;

基于所述推理结果确定所述图像数据的类别。

2. 根据权利要求1所述的方法,其特征在于,所述电子设备包括第一处理器,所述第一处理器能够对非对称量化的神经网络模型进行推理,不能对对称量化的神经网络模型进行推理;并且

由所述第一处理器运行所述第一神经网络模型。

3. 根据权利要求1或2所述的方法,其特征在于,所述对称量化参数还包括以下参数中的至少一种:输入数据的对称量化缩放系数、输出数据的对称量化缩放系数。

4. 根据权利要求3所述的方法,其特征在于,所述将各所述算子的对称量化参数转换为非对称量化参数,包括:

根据所述输入数据或所述输出数据的数据类型、所述输入数据或所述输出数据的对称量化缩放系数,确定所述输入数据或所述输出数据对应的浮点数的最大值和最小值;

根据所述输入数据或所述输出数据对应的浮点数的最大值和最小值,确定所述输入数据或所述输出数据的非对称量化缩放系数、非对称量化零点。

5. 根据权利要求4所述的方法,其特征在于,根据所述输入数据或所述输出数据的数据类型、所述输入数据或所述输出数据的对称量化缩放系数,确定所述输入数据或所述输出数据对应的浮点数的最大值和最小值,包括:

根据各算子的所述输入数据或所述输出数据的数据类型,确定所述输入数据或所述输

出数据的定点数的最大值和最小值；

根据所述输入数据或所述输出数据的对称量化缩放系数以及所述输入数据或所述输出数据的定点数的最大值和最小值，确定所述输入数据或所述输出数据对应的浮点数的最大值和最小值。

6. 根据权利要求3所述的方法，其特征在于，所述对称量化常量数据包括对称量化常数、对称量化矩阵；并且，所述将各所述算子的对称量化参数转换为非对称量化参数，包括：

根据所述对称量化常量数据的数据类型和所述对称量化常量数据的对称量化缩放系数，确定所述对称量化常量数据对应的浮点数的最大值和最小值；

根据所述对称量化常量数据对应的浮点数的最大值和最小值，确定所述对称量化常量数据对应的浮点数的非对称量化缩放系数和非对称量化零点；

根据确定出的所述对称量化常量数据对应的浮点数的非对称量化缩放系数和非对称量化零点，将所述对称量化常量数据对应的浮点数常量数据转换为非对称量化常量数据，其中，所述对称量化常量数据对应的浮点数常量数据，由所述常量数据的对称量化参数确定。

7. 根据权利要求6所述的方法，其特征在于，所述根据所述对称量化常量数据的数据类型和所述对称量化常量数据的对称量化缩放系数，确定所述对称量化常量数据对应的浮点数的最大值和最小值，包括：

根据所述对称量化常量数据的数据类型，确定所述对称量化常量数据的定点数的最大值和最小值；

根据所述对称量化常量数据的对称量化缩放系数、确定出的所述对称量化常量数据的定点数的最大值和最小值，确定所述对称量化常量数据对应的浮点数的最大值和最小值。

8. 根据权利要求1所述的方法，其特征在于，还包括：

根据所述对称量化的查表索引或所述非对称量化的查表结果的数据类型，确定所述对称量化的查表索引或所述对称量化的查表结果的对应的定点数的最大值和最小值，并基于确定出的最大值和最小值，根据所述对称量化常量数据的对称量化缩放系数，确定对称量化的查表索引对应的符点数查表索引的最大值和最小值或所述非对称量化的查表结果对应的符点数查表结果的最大值和最小值；

根据确定出的对称量化的查表索引对应的符点数查表索引的最大值和最小值或所述对称量化的查表结果对应的符点数查表结果的最大值和最小值，确定所述浮点数查表索引或所述浮点数查表结果的非对称量化缩放系数和非对称量化零点。

9. 一种可读介质，其特征在于，所述可读介质中包含有指令，当所述指令被电子设备的处理器执行时使电子设备实现权利要求1至8中任一项所述的神经网络模型的运行方法。

10. 一种电子设备，其特征在于，包括：

存储器，用于存储由电子设备的一个或多个处理器执行的指令；

以及处理器，是所述电子设备的处理器之一，用于运行所述指令以使所述电子设备实现权利要求1至8中任一项所述的神经网络模型的运行方法。

神经网络模型的运行方法、可读介质和电子设备

技术领域

[0001] 本申请涉及人工智能领域,特别涉及一种神经网络模型的运行方法、可读介质和电子设备。

背景技术

[0002] 随着人工智能(artificial intelligence, AI)的迅速发展,神经网络模型在人工智能领域的应用越来越广泛。由于运行神经网络模型的运算单元,例如神经网络处理器(Neural-Network Processing Unit, NPU),通常为定点运算单元,为提高神经网络模型的运行速度,通常将神经网络模型的各算子进行量化,得到定点运算的神经网络模型,再由电子设备来运行。对神经网络模型各算子的量化包括非对称量化或对称量化,但是,为了节省开发、制造成本,部分NPU中只预设有针对性非对称量化的算子的运算逻辑,该类NPU只能运行非对称量化的神经网络模型,而无法运行对称量化的神经网络模型。

发明内容

[0003] 有鉴于此,本申请实施例提供了神经网络模型的运行方法、可读介质和电子设备。电子设备通过对称量化的神经网络模型的对称量化参数转换为非对称量化参数,即可根据得到的非对称量化参数调用预设的非对称量化的算子的运算逻辑来实现对称量化的神经网络模型的功能,增加了电子设备能够运行的神经网络模型的类型,提高了电子设备的NPU的通用性。

[0004] 第一方面,本申请实施例提供了一种神经网络模型的运行方法,应用于电子设备,该方法包括:检测到第一神经网络模型,第一神经网络模型为对称量化的神经网络模型,并且第一神经网络模型中包括第一神经网络模型各算子的对称量化参数;将各算子的对称量化参数转换为非对称量化参数;利用各算子的非对称量化参数,调用预设的非对称量化的算子的运算逻辑,得到第一神经网络模型的推理结果。

[0005] 通过本申请实施例提供的方法,若电子设备中用于运行第一神经网络模型的处理器(例如NPU)只能调用非对称量化的算子的运算逻辑来实现神经网络模型的推理,则可以通过将第一神经网络模型各算子的对称量化参数转换为非对称量化参数,并通过得到的非对称量化参数来调用非对称量化的算子的运算逻辑,来对第一神经网络模型进行推理,得到第一神经网络模型的推理结果。如此,增加了电子设备能够运行的神经网络模型的类型,提高了电子设备的NPU的通用性。此外,由于过程中无需先将对称量化的神经网络模型转换为浮点型的神经网络模型,再转换为非对称量化的神经网络模型,提高了神经网络模型的部署速度。

[0006] 在上述第一方面的一种可能实现中,上述电子设备包括第一处理器,第一处理器能够对非对称量化的神经网络模型进行推理,不能对对称量化的神经网络模型进行推理;并且由第一处理器运行第一神经网络模型。

[0007] 也即是说,电子设备的第一处理器只能够调用非对称量化的算子的运算逻辑来对

神经网络模型进行推理,通过本申请实施例提供的方法,第一处理器可以通过将第一神经网络模型的对称量化参数转换为非对称量化参数,并基于得到的非对称量化参数调用非对称量化的算子的运算逻辑来实现第一神经网络模型的功能,增加了第一处理器能够运行的神经网络模型的类型,提高了第一处理器的通用性。

[0008] 在上述第一方面的一种可能实现中,上述对称量化参数包括以下参数中的至少一种:输入数据的对称量化缩放系数、输出数据的对称量化缩放系数、常量数据的对称量化参数,其中常量数据的对称量化参数包括对称量化常量数据、对称量化常量数据的对称量化缩放系数。

[0009] 在上述第一方面的一种可能实现中,上述将各算子的对称量化参数转换为非对称量化参数,包括:根据输入数据或输出数据的数据类型、输入数据或输出数据的对称量化缩放系数,确定输入数据或输出数据对应的浮点数的最大值和最小值;根据输入数据或输出数据对应的浮点数的最大值和最小值,确定输入数据或输出数据的非对称量化缩放系数、非对称量化零点。

[0010] 例如,在一个算子的输入数据或输出数据的数据类型为INT8时,若输入数据或输出数据的对称量化缩放系数为1,则确定输入数据或输出数据的浮点数的最大值为浮点数127,最小值为浮点数-128,进而根据下文中的公式(4)和公式(5)得到输入数据或输出数据的非对称量化缩放系数为1、非对称量化零点为128。

[0011] 在上述第一方面的一种可能实现中,上述根据输入数据或输出数据的数据类型、输入数据或输出数据的对称量化缩放系数,确定输入数据或输出数据对应的浮点数的最大值和最小值,包括:根据各算子的输入数据或输出数据的数据类型,确定输入数据或输出数据的定点数的最大值和最小值;根据输入数据或输出数据的对称量化缩放系数以及输入数据或输出数据的定点数的最大值和最小值,确定输入数据或输出数据对应的浮点数的最大值和最小值。

[0012] 例如,在输入数据或输出数据的数据类型为INT8的情况下,则输入数据或输出数据的定点数的最大值为127,最小值为-128,从而可以根据下文中的公式(11)确定输入数据或输出数据的浮点数的最大值为浮点数127,最小值为浮点数-128。

[0013] 在上述第一方面的一种可能实现中,上述对称量化常量数据包括对称量化常数、对称量化矩阵;并且,将各算子的对称量化参数转换为非对称量化参数,包括:根据对称量化常量数据的数据类型和对称量化常量数据的对称量化缩放系数,确定对称量化常量数据对应的浮点数的最大值和最小值;根据对称量化常量数据对应的浮点数的最大值和最小值,确定对称量化常量数据对应的浮点数的非对称量化缩放系数和非对称量化零点;根据确定出的对称量化常量数据对应的浮点数的非对称量化缩放系数和非对称量化零点,将对称量化常量数据对应的浮点数量数据转换为非对称量化常量数据,其中,对称量化常量数据对应的浮点数量数据,由常量数据的对称量化参数确定。

[0014] 例如,某一算子的常量数据包括对称量化的常数100,该对称量化常量数据的对称量化缩放系数为2,则根据公式(11)可以得到该常数对应的浮点数量数为50;基于该常量数据的数据类型,例如INT8,可以得到常量数据对应的浮点数的最大值为63.5,最小值为-64,进而根据下文中的公式(4)和公式(5)可以得到该常量数据的非对称量化缩放系数为2,非对称量化零点为128。

[0015] 在上述第一方面的一种可能实现中,上述根据对称量化常量数据的数据类型和对称量化常量数据的对称量化缩放系数,确定对称量化常量数据对应的浮点数的最大值和最小值,包括:根据对称量化常量数据的数据类型,确定对称量化常量数据的定点数的最大值和最小值;根据对称量化常量数据的对称量化缩放系数、确定出的对称量化常量数据的定点数的最大值和最小值,确定对称量化常量数据对应的浮点数的最大值和最小值。

[0016] 例如,在对称量化常量的数据类型为INT8、对称量化常量数据的对称量化缩放系数为2的情况下,则非对称量化常量数据的定点数的最大值为127,最小值为-128,从而可以根据下文中的公式(11)确定非对称量化常量数据对应的浮点数的最大值为浮点数63.5,最小值为浮点数-64。

[0017] 在上述第一方面的一种可能实现中,上述对称量化常量数据还包括对称量化的查找表,对称量化的查找表中包括对称量化的查表索引和各对称量化的查表索引对应的对称量化的查表结果;并且,将各算子的对称量化参数转换为非对称量化参数,包括:根据对称量化常量数据的对称量化缩放系数,确定出各对称量化的查表索引对应的浮点数查表索引;根据各对称量化的查表索引对应的浮点数查表索引和各算子的运算逻辑,确定各浮点数查表索引对应的浮点数查表结果;根据浮点数查表索引的非对称量化缩放系数和非对称量化零点得到非对称量化的查表索引、根据浮点数据查表结果的非对称量化缩放系数和非对称量化零点得到对称量化的查表结果,其中,浮点数查表索引的非对称量化缩放系数和非对称量化零点基于对称量化的查表索引的数据类型确定、浮点数查表结果的非对称量化缩放系数和非对称量化零点基于对称量化的查表结果的数据类型确定;基于各非对称量化的查表索引和相对应的非对称量化的查表结果,得到非对称量化的查找表。

[0018] 例如,对于下文中的Softmax算子,电子设备可以先根据查找表LUT中的查表索引的数据类型(INT8),查表索引的对称量化缩放系数(1.1×10^{-5}),确定各查表索引对应的浮点数查表索引,例如对于查表索引[1,5,121],根据公式(11)可以得到浮点数查表索引为[2727273,4727273,8272727];再将浮点数查表索引代入到Softmax算子的运算逻辑(公式(12))中,得到浮点数查表索引对应的浮点数查表结果[0.0069,0.042,0.9511],再根据浮点数查表结果的非对称量化缩放系数(1.1×10^{-5})、非对称量化零点128,将浮点数查表结果量化为定点数查表结果[129,133,249]。

[0019] 在上述第一方面的一种可能实现中,上述方法还包括:根据对称量化的查表索引或非对称量化的查表结果的数据类型,确定对称量化的查表索引或对称量化的查表结果的对应的定点数的最大值和最小值,并基于确定出的最大值和最小值,根据对称量化常量数据的对称量化缩放系数,确定对称量化的查表索引对应的浮点数查表索引的最大值和最小值或非对称量化的查表结果对应的浮点数查表结果的最大值和最小值;根据确定出的对称量化的查表索引对应的浮点数查表索引的最大值和最小值或对称量化的查表结果对应的浮点数查表结果的最大值和最小值,确定浮点数查表索引或浮点数查表结果的非对称量化缩放系数和非对称量化零点。

[0020] 例如,在下文中的Softmax算子的对称量化的查表结果、对称量化的查表结果的数据类型为INT8的情况下,对称量化的查表结果的定点数的最大值为127,最小值为-128,根据对称量化的查表索引的对称量化缩放系数(1.1×10^{-5}),对称量化的查表结果的非对称量化缩放系数(255),基于下文中的公式(11)可以得到对称量化查表索引对应的浮点数的最大

值为11545455、最小值为-11636364,对称量化查表结果对应的浮点数的最大值为1、最小值为-1,进而根据下文中的公式(4)和公式(5)可以得到浮点数查表索引的非对称量化缩放系数为 1.1×10^{-5} 、非对称量化零点为128,浮点数查表结果的非对称量化缩放系数为127.5、非对称量化零点为128。

[0021] 第二方面,本申请实施例提供了一种可读介质,该可读介质中包含有指令,当指令被电子设备的处理器执行时使电子设备实现上述第一方面及上述第一方面的各种可能实现提供的任意一种神经网络模型的运行方法。

[0022] 第三方面,本申请实施例提供了一种电子设备,该电子设备包括:存储器,用于存储由电子设备的一个或多个处理器执行的指令;以及处理器,是电子设备的处理器之一,用于运行指令以使电子设备实现上述第一方面及上述第一方面的各种可能实现提供的任意一种神经网络模型的运行方法。

附图说明

[0023] 图1A根据本申请的一些实施例,示出了一种8位对称量化的示意图;

[0024] 图1B根据本申请的一些实施例,示出了一种8位非对称量化的示意图;

[0025] 图2根据本申请的一些实施例,示出了一种将对称量化的神经网络模型部署到电子设备100中的场景示意图;

[0026] 图3根据本申请的一些实施例,示出了一种电子设备100运行对称量化的神经网络模型的场景图;

[0027] 图4根据本申请的一些实施例,示出了一种神经网络模型10的结构示意图;

[0028] 图5根据本申请的一些实施例,示出了一种电子设备利用对称量化的神经网络模型10对图像20进行分类的过程示意图;

[0029] 图6根据本申请的一些实施例,示出了一种神经网络模型运行方法的流程示意图;

[0030] 图7根据本申请的一些实施例,示出了一种电子设备100调用非对称量化的算子的运算逻辑对图像20进行分类的过程示意图;

[0031] 图8根据本申请的一些实施例,示出了一种电子设备100的结构示意图。

具体实施方式

[0032] 本申请的说明性实施例包括但不限于神经网络模型的运行方法、可读介质和电子设备。

[0033] 为了便于理解,首先介绍本申请实施例涉及的术语。

[0034] (1) 对称量化

[0035] 对称量化,即是将浮点数转换为取值范围相对于“0”对称的定点数。假设待量化的浮点数为 x_f ,量化目标为将 x_f 进行n位对称量化,即量化后的定点数的取值范围为 $[-2^{n-1}, 2^{n-1}-1]$,对称量化过程表示为如下公式(1)。

$$[0036] \quad x_q = \text{round}\left(x_f \cdot \frac{2^{n-1} - 1}{\max(\text{abs}(\max(x_f)), \text{abs}(\min(x_f)))}\right) \quad (1)$$

[0037] 公式(1)中,abs()为求绝对值函数,max()为求最大值函数,min()为求最小值函数,round为求四舍五入函数, x_q 为定点数。此外,在公式(1)中,如下公式(2)所示 S_c 项可以称

为对称量化缩放系数。也即是说,对于一个对称量化的定数,可以根据该定点数对应的对称量化缩放系数确定该定点数对应的浮点数。

$$[0038] \quad S_c = \frac{2^{n-1} - 1}{\max(\text{abs}(\max(x_f)), \text{abs}(\min(x_f)))} \quad (2)$$

[0039] 具体地,图1A根据本申请的一些实施例,示出了一种对浮点数 x_f 进行8位对称量化的示意图。参考图1A,量化的目标为将 x_f 量化进行8位对称量化,即是将 x_f 转换为INT8型(取值范围为[-128,127]),假设 x_f 的绝对值的最大值为 $\max(|x_f|)$,则对 x_f 进行8位对称量化的过程即是映射到区间[-128,127]中。

[0040] 可以理解,在一些实施例中INT型可以表示为INT n ,其中 n 可以取4、8、16、32等,也可以取其他整数,并且INT n 型的有符号数据的取值范围为 $[-2^{n-1}, 2^{n-1}-1]$ 。即是对于一个给定的INT型数据,电子设备可以根据 n 的值确定该类数据的取值范围,并基于公式(1)和公式(2)计算出该数据对应的浮点数的取值范围。

[0041] (2) 非对称量化

[0042] 非对称量化,即是将浮点数转换为取值范围为 $[0, 2^n - 1]$ 的无符号整形(unsigned int, UINT)数,其中 n 为非对称量化的位数。假设待量化的浮点数为 x_f ,量化目标为将 x_f 进行 n 位非对称量化,即量化后的定点数的取值范围为 $[0, 2^n - 1]$,则非对称量化过程表示为如下公式(3)。

$$[0043] \quad x_q = \text{round}((x_f - \min(x_f)) \frac{2^n - 1}{\max(x_f) - \min(x_f)}) \quad (3)$$

[0044] 公式(3)中, $\max()$ 为求最大值函数, $\min()$ 为求最小值函数,round为求四舍五入函数, x_q 为定点数。此外,在公式(3)中,如下公式(4)所示 AS_c 项可以称为非对称量化缩放系数,如下公式(5)所示的 Z_p 项可以称为非对称量化零点。也即是说,对于一个非对称量化的定点数,可以根据该定点数对应的非对称量化缩放系数和非对称量化零点,确定该定点数对应的浮点数。

$$[0045] \quad AS_c = \frac{2^n - 1}{\max(x_f) - \min(x_f)} \quad (4)$$

$$[0046] \quad Z_p = \text{round}(-AS_c \cdot \min(x_f)) \quad (5)$$

[0047] 基于公式(3)至公式(5),可以得到公式(6)所示的非对称量化的另一种表示方式:

$$[0048] \quad x_q = \text{round}(x_f \cdot AS_c + Z_p) \quad (6)$$

[0049] 具体地,图1B根据本申请的一些实施例,示出了一种对浮点数 x_f 进行8位非对称量化的示意图。参考图1B,量化的目标为将 x_f 量化进行8位非对称量化,即是将 x_f 转换为UINT8型(取值范围为[0,255]),假设 x_f 的最大值为 $\max(x_f)$,最小值为 $\min(x_f)$,则对 x_f 进行8位非对称量化的过程即是映射到区间[0,255]中。

[0050] 可以理解,在一些实施例中UINT型可以表示为UINT n ,其中 n 可以取4、8、16、32等,也可以取其他整数,并且UINT n 型的无符号数据的取值范围为 $[0, 2^n - 1]$ 。即是对于一个给定的UINT型数据,电子设备可以根据 n 的值确定该类数据的取值范围,并基于公式(6)计算出该数据对应的浮点数的取值范围。

[0051] (3) 神经网络模型量化

[0052] 神经网络模型量化,即是将神经网络模型各算子中的输入数据、输出数据、常量数据从大数据类型的浮点数(例如,32位浮点数)转换为较小数据类型的定点数(例如,4/8/16位定点数),并且定点数的位数通常和运行神经网络模型的运算单元,例如NPU,所支持的定点数位数相匹配,以提高NPU运行神经网络模型的速度。

[0053] 一般地,神经网络模型的量化的过程,即是根据各算子浮点数的输入数据、浮点数的输出数据的取值范围,以及量化的定点数的类型(例如UINT8),确定各算子的浮点数输入数据、浮点数输出数据和浮点数常量数据的量化参数(例如,非对称量化的量化参数包括非对称量化缩放系数和非对称量化零点,对称量化的量化参数包括对称量化缩放系数)。也即经过量化后的算子的量化参数中,包括了该算子的输入数据的量化参数、输出数据的量化参数、量化后的常量数据以及常量数据的量化参数。而用于运行量化后的神经网络模型的电子设备中,预设有量化后算子的运算逻辑。该量化后的算子的运算逻辑,以定点数为输入、定点数为输出,并且该运算逻辑中的参数包括输入数据的量化参数、输出数据的量化参数、量化后的常量数据以及常量数据的量化参数,电子设备在运行一个量化后的算子时,根据该算子的量化参数,调用预设的运算逻辑,即可通过定点运算实现该算子的功能。

[0054] 下面结合附图介绍本申请实施例的技术方案。

[0055] 为便于理解,下面先以卷积算子为例,说明采用对称量化和非对称量化的算子的运算逻辑。

[0056] 假设 B_f 为卷积算子的浮点数输入矩阵, C_f 为浮点数卷积核, D_f 为浮点数卷积结果,则 B_f 、 C_f 、 D_f 间的关系可以表示为公式(7)。

$$[0057] \quad D_f = B_f * C_f \quad (7)$$

[0058] 公式(7)中,“*”为卷积运算符号,卷积运算的具体过程将在后文进行介绍,在此不做赘述。

[0059] 根据公式(6)可以得到 $B_f = (B_q - Z_{p_B}) / B_ASc$, $C_f = (C_q - Z_{p_C}) / C_ASc$, $D_f = (D_q - Z_{p_D}) / D_ASc$,其中: B_q 为 B_f 对应的非对称量化的定点数矩阵, B_ASc 为将 B_f 量化为 B_q 的非对称量化缩放系数, Z_{p_B} 为将 B_f 非对称量化为 B_q 的非对称量化零点; C_q 为 C_f 对应的非对称量化的定点数卷积核, C_ASc 为将 C_f 量化为 C_q 的非对称量化缩放系数, Z_{p_C} 为将 C_f 非对称量化为 C_q 的非对称量化零点; D_q 为 D_f 对应的非对称量化的定点数矩阵, D_ASc 为将 D_f 非对称量化为 D_q 的对称量化缩放系数, Z_{p_D} 为将 D_f 非对称量化为 D_q 的非对称量化零点。进而公式(7)可以表示为如下公式(8)。

$$[0060] \quad \frac{D_q - Z_{p_D}}{D_ASc} = \frac{B_q - Z_{p_B}}{B_ASc} * \frac{C_q - Z_{p_C}}{C_ASc} \quad (8)$$

[0061] 对公式(8)进行变形可以得到如下公式(9)所示的非对称量化的卷积算子的运算逻辑。

$$[0062] \quad D_q = \text{round} \left\{ (B_q - Z_{p_B}) * (C_q - Z_{p_C}) \times \frac{D_ASc}{B_ASc \times C_ASc} + Z_{p_D} \right\} \quad (9)$$

[0063] 也即是说电子设备中预设有公式(9)所示的运算逻辑,该运算逻辑的输入包括:非对称量化的输入数据 B_q ,输入数据 B_q 的非对称量化零点 Z_{p_B} 、非对称量化缩放系数 B_ASc ,非对称量化的卷积核 C_q 、卷积核 C_q 的非对称量化零点 Z_{p_C} 、非对称量化缩放系数 C_ASc ,输出数据 D_q 的非对称量化零点 Z_{p_D} 、非对称量化缩放系数 D_ASc ;输出数据为非对称

量化的卷积结果 D_q 。

[0064] 类似地,对称量化的卷积算子的运算逻辑可以表示表如下公式(10),具体推导过程可以参考对称量化的卷积算子的运算逻辑,在此不做赘述。

$$[0065] \quad D_q = B_q * C_q \times \frac{D_{Sc}}{B_{Sc} \times C_{Sc}} \quad (10)$$

[0066] 公式(10)中, B_q 为对称量化的输入数据, C_q 为对称量化的卷积核, D_q 为定点数卷积结果, B_{Sc} 为 B_q 的对称量化缩放系数, D_{Sc} 为输出数据 D_q 的对称量化参数, C_{Sc} 为卷积核的对称量化缩放系数。

[0067] 从公式(9)和公式(10)中可以看出,对于同一算子,采用非对称量化和对称量化的运算逻辑和输入参数并不相同。如前所述,部分NPU只能运行非对称量化的神经网络模型,即只能运行非对称量化的算子的运算逻辑(例如公式(9)所示的卷积算子的运算逻辑),而无法运行对称量化的神经网络模型。若将对称量化的神经网络模型部署到包括该部分NPU的电子设备中,需要先将对称量化的神经网络模型转换浮点型神经网络模型(即神经网络模型的各算子的输入数据、输出数据及常量数据都为浮点数),再将浮点型神经网络模型量化为非对称量化的神经网络后,才能部署到该电子设备中。

[0068] 例如,参考图2,电子设备100的NPU只能运行非对称量化的神经网络模型,而待运行的神经网络模型为对称量化的神经网络模型。从而需要由电子设备200将对称量化的神经网络模型转换为浮点型神经网络模型,并将浮点型神经网络模型量化为非对称量化的神经网络模型后,再将非对称量化后的神经网络模型部署到电子设备100中,由电子设备100的NPU来运行。由于将浮点数神经网络模型量化为定点数神经网络模型(即神经网络模型的各算子的输入数据、输出数据、常量数据等为定点数)的过程中占用大量计算资源,耗时较长,不利于神经网络模型的快速部署。

[0069] 为了解决上述问题,本申请实施例提供了一种神经网络模型的运行方法,电子设备100在检测到对称量化的神经网络模型后,获取该对称量化的神经网络模型的各算子的对称量化参数,并将各算子对称量化参数转换为对应的非对称量化参数,电子设备100的NPU再根据各算子的非对称量化参数调用预设的非对称量化的算子的运算逻辑,来实现该对称量化的神经网络模型的相关功能。如此,参考图3,虽然电子设备100的NPU并不能运行对称量化的神经网络模型,但电子设备100通过将对称量化的神经网络模型中的各算子的对称量化参数转换为非对称量化参数,电子设备100的NPU即可调用预设的非对称量化的算子的运算逻辑,来实现该对称量化神经网络模型的相关功能,而无需由其他电子设备将对称量化的神经网络模型转换为浮点型神经网络模型,并将浮点型神经网络模型转化为非对称量化的神经网络模型后,再由电子设备100的NPU来运行,增加了电子设备100能够运行的神经网络模型的类型,提高了NPU的通用性,提升了神经网络模型的部署速度。

[0070] 可以理解,各算子的对称量化参数包括以下参数中的至少一项:输入数据的对称量化缩放系数、输出数据的对称量化缩放系数、对称量化常量数据、对称量化常量数据的对称量化缩放系数。各算子的非对称量化参数包括以下参数中的至少一项:输入数据的非对称量化缩放系数、输入数据的非对称量化零点、输出数据的非对称量化缩放系数、输出数据的非对称量化零点、非对称量化常量数据、非对称量化常量数据的非对称量化零点、非对称量化常量数据的非对称量化缩放系数。

[0071] 可以理解,在另一些实施例中,各算子的非对称量化参数/对称量化参数也可以包括更多或更少的参数,在此不做限定。

[0072] 可以理解,预设的非对称量化的算子的运算逻辑可以由电子设备100的NPU的开发商预设于NPU中,也可以由NPU的开发商提供给电子设备100的开发商,由电子设备100的开发商预设于电子设备100的存储器中。电子设备100的NPU可以根据基于算子的非对称量化的输入数据及非对称量化的输入数据的非对称量化参数、输出数据的非对称量化参数、非对称量化常量数据调用非对称量化的算子的运算逻辑,得到算子的非对称量化的定点数输出数据。

[0073] 具体地,在一些实施例中,对上述公式(1)和公式(2)可以得到如公式(11)所示的、将对称量化的定点数转化为浮点数的计算公式。

$$[0074] \quad x_f = \frac{x_q}{S_c} \quad (11)$$

[0075] 电子设备100可以根据对称量化的算子的定点数输入数据的数据类型确定定点数输入数据的取值范围(例如若输入数据为INT8,说明取值范围为定点数[-128,127],最大值为127,最小值为-128),基于对称量化的输入数据的最大值和最小值、对称量化的输入数据的对称量化缩放系数,基于公式(11)确定输入数据对应的浮点数的最大值和最小值;再基于输入数据对应的浮点数的最大值和最小值,根据公式(4)和公式(5)确定输入数据的非对称量化缩放系数和非对称量化零点。

[0076] 类似地,电子设备100可以根据对称量化的算子的输出数据的类型确定定点数输出数据的取值范围(例如若输入数据为INT8,说明取值范围为定点数[-128,127],最大值为127,最小值为-128),并根据对称量化的输出数据的最大值和最小值、对称量化缩放系数,基于公式(11)确定输出数据对应的浮点数的最大值和最小值;再基于确定的浮点数的最大值和最小值,根据公式(4)和公式(5)确定输出数据的非对称量化缩放系数和非对称量化零点。

[0077] 类似地,电子设备100也可以根据对称量化的算子中对称量化常量数据的数据类型,确定对称量化常量数据的定点数的最大值和最小值,再基于确定的最大值和最小值、对称量化常量数据的对称量化缩放系数,根据公式(11)确定对称量化常量数据对应的浮点数的最大值和最小值;再基于确定的浮点数的最大值和最小值,根据公式(4)和公式(5)确定常量数据的非对称量化缩放系数和非对称量化零点;然后,根据公式(11)将对称量化常量数据转换对应的浮点数常量数据,再基于确定的常量数据的非对称量化缩放系数、常量数据的非对称量化零点和对称量化常量数据对应的浮点数常量数据,根据公式(6)将对称量化常量数据对应的浮点数常量数据转换为非对称量化常量数据。

[0078] 例如,假设某一量化为INT8的非对称量化的卷积算子的输入数据矩阵B_f中的各

元素的取值范围为浮点数[-1,1],浮点数卷积核C_f =
$$\begin{bmatrix} 0.1 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0.1 \end{bmatrix}$$
,则可以得到浮点数卷

积结果D_f中的各元素的取值范围为[-0.7,0.7]。

[0079] 根据公式(2),可以得到输入数据的对称量化缩放系数B_Sc=127,输出数据的对称量化缩放系数D_Sc=181.43。假设卷积核C_f的量化参数与输入数据B_f的对称量化参数

相同(即 $C_{Sc}=127$),根据公式(6)可以得到 $C_q = \begin{bmatrix} 13 & 0 & 0 \\ 0 & 64 & 0 \\ 0 & 0 & 13 \end{bmatrix}$ 。也即是该对称量化的卷积

算子的对称量化参数包括:输入数据的对称量化缩放系数 B_{Sc} ,输出数据的对称量化缩放系数 D_{Sc} ,对称量化常量数据 C_q ,以及对称量化常量数据的对称量化缩放系数 C_{Sc} 。

[0080] 电子设备在检测到上述对称量化的卷积算子时,先根据输入数据的数据类型INT8,确定定点数输入数据 D_q 的取值范围为 $[-128, 127]$,最大值为127,最小值为-128;将 $x_q=127, B_{Sc}=127$ 代入公式(11)得到输入数据对应的浮点数最大值为1;将 $x_q=0, B_{Sc}=-128$ 代入公式(11)得到输入数据对应的浮点数最小值为-1,进而根据公式(4)可以得到输入数据的非对称量化缩放系数 $B_{ASc} = (2^8 - 1) / (1 - (-1)) = 127.5$,根据公式(5)可以得到输入数据的非对称量化零点 $Z_{pB} = \text{round}(-127.5 \times (-1)) = 128$ 。

[0081] 类似地,电子设备可以根据输出数据的数据类型INT8,确定定点数输出数据 D_q 的取值范围为 $[-128, 127]$,最大值为127,最小值为-128;将 $x_q=127, D_{Sc}=181.43$ 代入公式(11)得到输出数据对应的浮点数最大值为 $127/181.43=0.7$;将 $x_q=-128, D_{Sc}=181.43$ 代入公式(11)得到输出数据对应的浮点数最小值为 $127/181.43=-0.7$,进而根据公式(4)可以得到输出数据的非对称量化缩放系数 $D_{ASc} = (2^8 - 1) / (0.7 - (-0.7)) = 182.14$,根据公式(5)可以得到输出数据的非对称量化零点 $Z_{pD} = \text{round}(-182.14 \times (-0.7)) = 128$ 。

[0082] 由于卷积核 C_q 的对称量化参数与输入数据 B_q 的对称量化参数相同,则卷积核 C_q 的非对称量化参数也应当与输入数据 B_q 的非对称量化参数相同,即是卷积核 C_q 的非对称量化缩放系数 $C_{ASc}=127.5$,卷积核的非对称量化零点 $Z_{pC}=128$ 。将 $C_{Sc}=127, C_q$ 代

入公式(11)可以得到 $C_f = \begin{bmatrix} 0.1 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0.1 \end{bmatrix}$,再将 $C_f, C_{ASc}=127.5, Z_{pC}=128$ 公式(6)可以

得到对称量化的卷积核 $C_{q'} = \begin{bmatrix} 141 & 128 & 128 \\ 128 & 192 & 128 \\ 128 & 128 & 141 \end{bmatrix}$ 。

[0083] 进而电子设备100的NPU可以根据 $B_{ASc}=127.5, D_{ASc}=181.43, C_{ASc}=127.5, Z_{pB}=128, Z_{pC}=128, Z_{pD}=128$ 以及输入到该卷积算子的定点数输入数据 B_q ,即可调用预设的公式(9)所示的非对称量化的卷积算子的运算逻辑来实现前述对称量化的卷积算子的功能,增加了电子设备100的NPU能够运行的神经网络模型的类型,提高了NPU的通用性。

[0084] 下面结合具体的神经网络模型介绍本申请实施例的技术方案。

[0085] 图4根据本申请的一些实施例,示出了一种神经网络模型10的结构示意图。如图4所示,神经网络模型10为对称量化的神经网络模型,包括输入层11、卷积算子12、全连接算子13、Softmax算子14和输出层15,用于对输入的图像数据进行分类。其中,输入层11用于对输入的图像进行预处理,将输入的图像数据转换为对称量化的输入数据,例如INT8型的数据;卷积算子12用于对对称量化的输入数据进行卷积运算,得到输入的图像对应的对称量化的特征矩阵;全连接算子13用于对对称量化的特征矩阵进行全连接运算,得到输入的图像数据属于各预设类别的得分;Softmax算子14,用于根据输入的图像数据属于各预设类别

的得分,得到输入的图像数据属于各预设类别的概率;输出层15用于根据输入的图像数据属于各预设类别的概率,确定出输入的图像数据的类别,例如以输入的图像数据为各预设类别的概率最大的类别作为该输入的图像数据的类别。

[0086] 进一步,图5根据本申请实施例,示出了电子设备利用神经网络模型10对图像20进行分类的过程示意图。

[0087] 参考图5,电子设备先利用输入层11对图像20进行预处理,得到INT8的图像矩阵H,输入层11的对称量化缩放系数为 $Sc_{out1}=0.498$ 。

[0088] 其次,利用卷积算子12将图像矩阵H分别与卷积核 K_i ($i=1,2,3$)进行卷积运算,得到三个特征矩阵 A_i ($i=1,2,3$),卷积算子12的输入数据的对称量化缩放系数为 $Sc_{in2}=0.498$ 、输出数据的对称量化缩放系数为 $Sc_{out2}=0.0417$ 、对称量化常量数据包括卷积核 K_i ($i=1,2,3$) (卷积核的对称量化参数缩放系数与卷积算子12的输入数据的对称量化缩放系数相同)。

[0089] 再利用全连接算子13对特征矩阵 A_i ($i=1,2,3$)进行全连接运算,例如分别将特征矩阵 A_i ($i=1,2,3$)与权重矩阵W做内积运算,得到图像20为预设类别(兔子/狗/猫)的得分,全连接算子13的输入数据的对称量化缩放系数 $Sc_{in3}=0.0417$ 、输出数据的非对称量化缩放系数为 $Sc_{out3}=1.1 \times 10^{-5}$ 、非对称量化常量数据包括权重矩阵W(权重矩阵的对称量化缩放系数与全连接算子13的输入数据的对称量化缩放系数相同)。

[0090] 然后,利用Softmax算子14根据图像20为预设类别的得分,从查找表LUT中获取图像20为预设类别的概率,Softmax算子14的输入数据的对称量化缩放系数 $Sc_{in4}=1.1 \times 10^{-5}$ 、输出数据的对称量化缩放系数为 $Sc_{out4}=127$ 、非对称量化常量数据包括查找表LUT。

[0091] 最后,利用输出层15根据图像20为各预设类别的概率,确定图像20的类别,例如比较图像20为各预设类别的概率,将图像20为预设类别的概率中最大的概率确定为图像20的类别(猫)。

[0092] 下面结合图4所示的神经网络模型10的结构和图5所示的神经网络模型10对图像20进行分类的过程,介绍本申请实施例的技术方案。

[0093] 具体地,图6根据本申请的一些实施例,示出了一种神经网络模型的运行方法的流程图示意图。该方法的执行主体为电子设备100,如图6所示,该流程包括如下步骤。

[0094] S601:检测到对称量化的神经网络模型。

[0095] 电子设备100在检测待运行的神经网络模型为对称量化的神经网络模型的情况下,触发本申请实施例提供的神经网络模型的运行方法。

[0096] 在一些实施例中,电子设备100可以根据待运行神经网络模型中的数据的数据类型,来确定该神经网络是否是对称量化的神经网络模型。具体地,电子设备100在检测到待运行神经网络模型中的数据的数据类型为INT(INT包括但不限于INT4、INT8、INT16、INT32等)时,例如检测到神经网络模型10中的数据的数据类型为INT8时,确定待运行的神经网络模型为对称量化的神经网络模型。

[0097] 在一些实施例,电子设备100也可以根据待运行神经网络模型各算子的量化参数来确定待运行神经网络模型是否是对称量化的神经网络模型。例如,电子设备100可以在检测到待运行神经网络模型的量化参数包括缩放系数但不包括零点时,确定该神经网络模型为对称量化的神经网络模型。

[0098] 可以理解,在另一些实施例中,电子设备100也可以通过其他方式确定待运行的神经网络模型是否为对称量化的神经网络模型,并在检测到待运行的神经网络模型是对称量化的神经网络模型的情况下,触发本申请实施例提供的神经网络模型的运行方法。

[0099] S602:获取各算子的输入数据/输出数据的对称量化参数,并将对称量化参数转换为非对称量化参数。

[0100] 即是电子设备100依次获取各算子的输入数据和输出数据的对称量化参数,并将输入数据的对称量化参数(输入数据的对称量化缩放系数)转换为输入数据的非对称量化参数(输入数据的非对称量化缩放系数、输入数据的非对称量化零点),将输出数据的对称量化参数(输出数据的对称量化缩放系数)转换为输出数据的非对称量化参数(输出数据的非对称量化缩放系数、输出数据的非对称量化零点)。

[0101] 例如,对于图4所示的神经网络模型10,输入层11的对称量化参数包括:输出数据的对称量化缩放系数 $Sc_out1=0.498$ 。由于INT8对应的取值范围为 $[-128, 127]$,从而根据公式(11)可以得到输出数据对应的浮点数的最大值为浮点数 $127/0.498=255$,最小值为浮点数 $-128/0.498=-257$,电子设备100可以基于公式(4),确定输入层11的输出数据的非对称量化缩放系数 $ASc_out1=(2^8-1)/(255-(-257))=0.498$,基于公式(5)确定输入层11的输出数据的非对称量化零点 $Z_{p_out1}=\text{round}(-0.498 \times (-257))=128$ 。

[0102] 又例如,对于图4所示的神经网络模型10,卷积层12的对称量化参数包括:输入数据的对称量化缩放系数 $Sc_in2=0.498$ 、输出数据的对称量化缩放系数 $Sc_out2=0.0417$ 。由于INT8对应的取值范围为 $[-128, 127]$,从而根据公式(11)可以得到输入数据对应的浮点数的最大值为浮点数 $127/0.498=255$ 、最小值为浮点数 $-128/0.498=-257$,输出数据对应的浮点数的最大值为浮点数 $127/0.0417=3046$,最小值为浮点数 $-128/0.0417=-3070$ 。电子设备100可以基于公式(4),确定卷积算子12的输入数据的非对称量化缩放系数 $ASc_in2=(2^8-1)/(255-(-257))=0.498$ 、输出数据的非对称量化缩放系数 $ASc_out2=(2^8-1)/(3046-(-3070))=0.0417$,基于公式(5)确定卷积算子12的输入数据的非对称量化零点 $Z_{p_in2}=\text{round}(-0.498 \times (-257))=128$ 、输出数据的非对称量化零点 $Z_{p_out2}=\text{round}(-0.0417 \times (-3070))=128$ 。类似地,可以得到全连接层13的输入数据的非对称量化缩放系数 $ASc_in3=0.0417$ 、输入数据的非对称量化零点 $Z_{p_in3}=128$,输出数据的非对称量化缩放系数 $ASc_out3=1.1 \times 10^{-5}$,输出数据的非对称量化零点 $Z_{p_out3}=128$ 。

[0103] 再例如,对于图4所示的神经网络模型10,Softmax算子14的非对称量化参数包括:输入数据的对称量化缩放系数 $Sc_in4=1.1 \times 10^{-5}$ 、输出数据的对称量化缩放系数 $Sc_out4=127$ 。由于INT8对应的取值范围为 $[-128, 127]$,从而根据公式(11)可以得到输入数据对应的浮点数的最大值为浮点数为 $127/(1.1 \times 10^{-5})=11545455$ 、最小值为浮点数 $-128/(1.1 \times 10^{-5})=-11636364$,输出数据对应的浮点数的最大值为浮点数 $127/127=1$,最小值为浮点数 $-128/127=-1$ 。可以基于公式(4),确定Softmax算子14的输入数据的非对称量化缩放系数 $ASc_in4=(2^8-1)/(11545455-(-11636364))=1.1 \times 10^{-5}$ 、输出数据的非对称量化缩放系数 $ASc_out4=(2^8-1)/(1-(-1))=127.5$,基于公式(5)确定Softmax算子14的输入数据的非对称量化零点 $Z_{p_in4}=\text{round}(1.1 \times 10^{-5} \times (-11636364))=128$ 、输出数据的非对称量化零点 $Z_{p_out4}=\text{round}(-0.0417 \times (-3070))=128$ 。

[0104] S603:判断当前算子是否包括对称量化常量数据。

[0105] 电子设备100判断当前算子是否包括对称量化常量数据,如果有,说明需要将对称量化常量数据转换为非对称量化常量数据,转至步骤S604;否则,说明不需要将对称量化常量数据转换为非对称量化常量数据,转至步骤S605。

[0106] 例如,对于前述神经网络模型10,在当前算子为输入层11或输出层15时,当前算子不存在对称量化常量数据,转至步骤S605;对于前述神经网络模型10,卷积算子12存在对称量化常量数据卷积核 K_i ($i=1,2,3$)、全连接算子13存在对称量化常量数据权重矩阵 W , Softmax算子14存在对称量化常量数据查找表LUT,在当前算子为卷积算子12、全连接算子13或Softmax算子14时,电子设备100可以确定出当前算子包括对称量化常量数据,转至步骤S604。

[0107] S604:根据对称量化常量数据的对称量化参数,将对称量化常量数据转化为非对称量化常量数据。

[0108] 电子设备100在当前算子存在对称量化常量数据的情况下,根据当前算子的对称量化常量数据的对称量化参数,先根据对称量化常量数据的定点数的数据类型,确定对称量化常量数据的定点数的最大值和最小值,从而根据公式(11)确定对称量化常量数据对应的浮点数的最大值和最小值;再根据公式(4)和公式(5)确定将浮点数的常量数据转化为非对称量化常量数据的非对称量化缩放系数和非对称量化零点;然后根据公式(11)将对称量化常量数据转化为对应的浮点数的常量数据,再根据公式(6)和确定出的将浮点数的常量数据转化为非对称量化常量数据的非对称量化缩放系数和非对称量化零点,将浮点数的常量数据转换为非对称量化常量数据。

[0109] 例如,在当前算子为前述卷积算子12时,对称量化常量数据包括卷积核 K_i ($i=1,2,3$)。由于卷积核 K_i 的对称量化参数与卷积算子12的输入数据的对称量化参数相同,则卷积核 K_i 的非对称量化缩放系数与卷积算子12的输入数据的非对称量化缩放系数相同(均为 ASc_in2)、卷积核 K_i 的非对称量化零点与卷积算子12的输入数据的非对称量化零点相同(均为 Z_p_in2)。电子设备100可以先根据公式(11),将 K_i 转换为对应的浮点数 $K_{i_f}=K_i/Sc_in2$;再根据公式(6),将 K_{i_f} 转换为非对称量化的卷积核 $K'_i=round(K_{i_f} \times ASc_in2 + Z_p_in2)$ 。

具体地,参考图7,假设 $K_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$,则 $K'_1 = \begin{bmatrix} 129 & 128 & 128 \\ 128 & 130 & 128 \\ 128 & 128 & 130 \end{bmatrix}$ 。

[0110] 又例如,在当前算子为前述全连接算子13时,对称量化常量数据包括权重矩阵 W 。由于权重矩阵 W 的对称量化参数与全连接算子13的输入数据的对称量化参数相同,则权重矩阵 W 的非对称量化缩放系数与全连接算子13的输入数据的非对称量化缩放系数相同(即 ASc_in3)、权重矩阵 W 的非对称量化零点与全连接算子13的输入数据的非对称量化零点相同(即 Z_p_in3)。电子设备100可以先根据公式(11),将 W 转换为对应的浮点数 $W_f=W/Sc_in3$;再根据公式(6),将 W_f 转换为非对称量化的权重矩阵 $W'=round(W_f \times ASc_in3 + Z_p_in3)$ 。

[0111] 具体地,参考图7,假设

$$[0112] \quad W = \begin{bmatrix} 5 & 5 & 10 & 9 & 5 & 5 \\ 0 & 0 & 2 & 9 & 10 & 0 \\ 0 & 0 & 2 & 4 & 5 & 5 \\ 2 & 10 & 9 & 0 & 6 & 9 \\ 2 & 6 & 0 & 0 & 0 & 9 \\ 1 & 5 & 7 & 5 & 5 & 0 \end{bmatrix},$$

[0113] 则

$$[0114] \quad W' = \text{round}\left(\frac{W}{1.1 \times 10^{-5}} \times 1.1 \times 10^{-5} + 128\right) = \begin{bmatrix} 133 & 133 & 138 & 137 & 133 & 133 \\ 128 & 128 & 130 & 137 & 138 & 128 \\ 128 & 128 & 130 & 132 & 133 & 133 \\ 130 & 138 & 137 & 128 & 134 & 137 \\ 130 & 134 & 128 & 128 & 128 & 137 \\ 129 & 133 & 135 & 133 & 133 & 128 \end{bmatrix}.$$

[0115] 再例如,在当前算子为前述Softmax算子14时,对称量化常量数据包括查找表LUT。Softmax运算可以表示为如下公式(12)。

$$[0116] \quad P_{k_f} = \frac{e^{\frac{in_{k_f}}{1159091}}}{\sum_{j=1}^3 e^{\frac{in_{j_f}}{1159091}}} \quad (k=1, 2, 3) \quad (12)$$

[0117] 公式(12)中, in_{k_f} 表示Softmax算子的浮点数输入数据,即是输入的图像数据在第k类别的浮点数得分; P_{k_f} 是Softmax算子的浮点数输出结果,表示输入的图像数据为第k类的概率,其中, $k=1$ 表示类别为兔子, $k=2$ 表示类别为狗, $k=3$ 表示类别为猫。从公式(12)可知, P_{k_f} 的取值范围为[0,1]。

[0118] 查找表LUT用于存储不同的定点数输入数据 in_{k_q} 对应的Softmax运算的定数结果。例如,假设全连接算子13的输出数据的对称量化缩放系数为 1.1×10^{-5} , $in_{1_q}=30$, $in_{2_q}=52$, $in_{3_q}=91$,则可以得到 $in_{1_f}=2727273$, $in_{2_f}=4727273$, $in_{3_f}=8272727$,将 $in_{1_f}=2727273$, $in_{2_f}=4727273$, $in_{3_f}=8272727$ 代入前述公式(12)可以得到 $P_{1_f}=0.0069$, $P_{2_f}=0.042$, $P_{3_f}=0.9511$,再将 $P_{1_f}=0.0069$, $P_{2_f}=0.042$, $P_{3_f}=0.9511$ 进行8位对称量化得到: $P_{1_q}=1$, $P_{2_q}=5$, $P_{3_q}=121$,也即说,在查找表LUT中,存储有查表索引[30,52,91](对应[in1,in2,in3])对应的查表结果为[1,5,121](对应 P_1,P_2,P_3)。其他的查表索引对应的查表结果可以通过类似的方法得到,在此不做赘述。

[0119] 电子设备100在检测到上述查找表LUT时,先根据公式(11)将查找表LUT的查表索引转换为浮点数查表索引,将浮点数查表索引转化为对称量化的定点数查表索引,并将浮点数查表索引代入前述公式(12)得到浮点数 P_{k_f} ,再将浮点数 P_{k_f} 进行非对称量化得到的非对称量化查表结果作为前述非对称量化的定点数查表索引在新的查找表LUT'的查表结果。例如,将查表索引[30,52,91]转换为浮点数得到 $in_{1_f}=2727273$, $in_{2_f}=4727273$, $in_{3_f}=8272727$,将 $in_{1_f}=2727273$, $in_{2_f}=4727273$, $in_{3_f}=8272727$ 在非对称量化缩放系数为 ASc_in4 、非对称量化零点为 Z_p_in4 的情况下进行8位非对称量化,得到新的查表索引[158,180,219],并将 $in_{1_f}=2727273$, $in_{2_f}=4727273$, $in_{3_f}=8272727$ 代入前述公式(12)

可以得到 $P_{1_f}=0.0069, P_{2_f}=0.042, P_{3_f}=0.9511$, 再将 $P_{1_f}=0.0069, P_{2_f}=0.042, P_{3_f}=0.9511$ 在非对称量化缩放系数为 ASc_out4 、非对称量化零点为 Z_{p_out4} 进行8位非对称量化, 得到 $P_{1_q}=129, P_{2_q}=133, P_{3_q}=249$, 即是在图7所示的查找表LUT'中, 查表索引[158, 180, 219] (对应 $[in1', in2', in3']$) 对应的查表结果为[129, 133, 249] (对应 P_1', P_2', P_3')。其他的查表索引对应的结果可以通过类似的方法得到, 在此不做赘述。

[0120] 可以理解, 以上将卷积算子、全连接算子和Softmax算子中对称量化常量数据转换为非对称量化常量数据只是一种示例, 对于其他算子 (包括但不限于池化算子、激活算子、排序算子、归一化算子等) 中的对称量化常量数据, 可以使用类似的方法转换为非对称量化常量数据, 在此不做赘述。

[0121] S605: 判断是否完成所有算子的量化参数的转换。

[0122] 电子设备100判断是否完成所有算子的转换, 如果完成, 则转至步骤S606; 否则转至步骤S602进行下一个算子的量化参数的转换。

[0123] S606: 根据各算子的非对称量化参数, 调用相应的非对称量化算子的运算逻辑, 实现对称量化的神经网络模型的功能。

[0124] 电子设备100在完成所有算子的量化参数的转换后, 根据各算子的非对称量化参数, 通过NPU调用相对应的非对称量化算子的运算逻辑, 实现对称量化的神经网络模型的功能。

[0125] 具体地, 参考图7, 电子设备100将神经网络模型10的各算子的对称量化参数转换为非对称量化参数后, 各算子可以表示为图7所示的输入层11', 卷积算子12', 全连接算子13', Softmax算子14', 和输出层15'。

[0126] 电子设备100的NPU可以先在非对称量化缩放系数 $ASc_out1=0.0498$ 的情况下, 基于公式(1)将图像20量化为图像矩阵 H' 。

[0127] 其次, 电子设备100的NPU调用非对称量化的卷积算子的运算逻辑, 例如前述公式(9)所示的运算逻辑, 在卷积核为 K_i' ($i=1, 2, 3$) 的情况下, 将图像矩阵 H' 分别与 K_i' 进行卷积, 得到定点数特征矩阵 A_i' ($i=1, 2, 3$)。即是NPU获取非对称量化的输入数据 B_q (例如前述 H')、输入数据的非对称量化缩放系数 B_ASc (例如由前述步骤S602得到的 ASc_in2)、输入数据的非对称量化零点 (例如由前述步骤S602得到的 Z_{p_in2})、非对称量化的卷积核 C_q (例如由前述步骤S603得到的卷积核 K_i')、卷积核的非对称量化缩放系数 C_Sc (例如前述 ASc_in2)、卷积核的非对称量化零点 (例如由前述步骤S602得到的 Z_{p_in2})、输出数据的非对称量化缩放系数 D_ASc (例如由前述步骤S602得到的 ASc_out2)、输出数据的非对称量化零点 (例如由前述步骤S602得到的 Z_{p_out2}), 再根据前述公式(9)得到 D_q 。例如, 在 $B_q=H', C_q=K_i'$ 的情况下, 可以得到前述特征矩阵 A_i' 。

[0128] 可以理解, 由于NPU中并没有可以直接实现除法运算的电路, 在一些实施例中, 公式(9)中的除法运算可以通过乘法移位来实现, 以提高NPU运行卷积算子的速度。例如, 假设 $B_ASc \times C_ASc = 0.498^2 = 0.248$, 可以将0.248表示为 $1/1 \times 2^{-2}$, 从而将 $(B_q - Z_{p_B}) * (C_q - Z_{p_C}) \times D_ASc$ 的结果对应的二进制数向右移-2位再乘以1, 即可得到 $(B_q - Z_{p_B}) * (C_q - Z_{p_C}) \times D_ASc / (B_ASc \times C_ASc)$ 的运算结果, 最后将该运算结果加上 Z_{p_D} 即可得到 D_q 。

[0129] 然后, 电子设备100的NPU调用非对称量化的全连接算子的运算逻辑 (例如下方公式(17)所示的运算逻辑), 分别将特征矩阵 A_i' 与权重矩阵 W' 作全连接运算, 例如做内积, 得

到图像20为各预设类别的得分 $in1'$ 、 $in2'$ 和 $in3'$ ；再调用非对称量化的Softmax算子的运算逻辑,即是以 $[in1', in2', in3']$ 为查表索引从查找表LUT'中查找得到图像20属于各预设类别的概率;最后再调用非对称量化的输出层的运算逻辑,即是将图像20属于各预设类别的概率中最大的概率对应的预设类别作为图像20的类别,例如将图像20的类别确定为猫。其中,非对称量化全连接算子的运算逻辑的推导过程将在下文进行介绍,在此不做赘述。

[0130] 可以理解,非对称量化的算子的运算逻辑可以由电子设备100的NPU的开发商预先设置于NPU中,也可以由NPU的开发商提供给电子设备100的开发商,由电子设备100的开发商预设于电子设备100的存储器中。

[0131] 可以理解,上述步骤S601至步骤S605中的各步骤可以全部由电子设备100的CPU来完成,也可以全部由电子设备100的NPU来完成,还可以由电子设备100的CPU和NPU分别完成部分步骤,在此不做限定。

[0132] 可以理解,上述步骤S601至步骤S605的运行顺序只是一种示例,在另一些实施例中,可以调整部分步骤的运行顺序,也可以合并或拆分部分步骤,本申请实施例不做限定。

[0133] 通过本申请实施例提供的方法,电子设备100的NPU可以通过调用预设的非对称量化的算子来实现该对称量化神经网络模型的相关功能,而无需由其他电子设备将对称量化的神经网络模型转换为浮点型神经网络模型,并将浮点型神经网络模型转化为非对称量化的神经网络模型后,再电子设备100的NPU来运行,增加了电子设备100能够运行的神经网络模型的类型,提高了NPU的通用性,提升了神经网络模型的部署速度。

[0134] 下面介绍卷积计算的具体过程和对称量化的全连接算子的运算逻辑。

[0135] 首先介绍卷积运算的计算过程。

[0136] 假设矩阵B的大小为 $M \times M$,卷积核C的大小为 $N \times N$,卷积步长为 k ,则矩阵B与卷积核C的卷积结果D可以表示为:

$$[0137] \quad D(m, n) = \sum_{i=1}^N \sum_{j=1}^N B(k(m-1)+i, k(n-1)+j) \times C(i, j) \quad (13)$$

[0138] 在公式(13)中, $D(m, n)$ 为矩阵D第 m 行第 n 列的元素; m, n 满足以下关系式:

$$[0139] \quad 1 \leq m \leq \lfloor \frac{(M-N)}{k} + 1 \rfloor, \quad 1 \leq n \leq \lfloor \frac{(M-N)}{k} + 1 \rfloor。$$

[0140] 其中 $\lfloor X \rfloor$ 为向下取整运算,即 $\lfloor X \rfloor$ 为小于 X 的最大整数。由于 $M-N < M$ 且 k 为正整数,可见 $\lfloor \frac{(M-N)}{k} + 1 \rfloor \leq M$,也即是说卷积结果D的大小总是小于或等于矩阵B的大小。

[0141] 为确保卷积结果对应的矩阵的大小与输入数据的大小相同,避免丢失数据矩阵边缘的数据特征,通常在卷积计算的过程中在输入矩阵的第一行前和最后一行后填充值为0的行以及在输入矩阵第一列之前和最后一列之后填充值为0的行或列,即在输入矩阵的四周填充值为0的行或列。设在输入矩阵B的四周各填充数量为 P 的值为0的行或列,此时,输入矩阵B的大小变为 $(M+2P) \times (M+2P)$ 。此时,公式(1)中的 m, n 满足以下关系式:

$$[0142] \quad 1 \leq m \leq \lfloor \frac{(M-N+2 \times P)}{k} + 1 \rfloor, \quad 1 \leq n \leq \lfloor \frac{(M-N+2 \times P)}{k} + 1 \rfloor。$$

[0143] 令 $\frac{(M-N+2 \times P)}{k} + 1 = M$,即可计算得到 P 的值,例如,在卷积核大小为 $N=3$,步长 k

=1, 则P=1。

[0144] 下面介绍非对称量化的全连接算子的运算逻辑。

[0145] 全连接算子为对输入数据进行加权计算的算子, 输入矩阵E和权重矩阵W的全连接计算结果F可以表示为如下公式(14)。

$$[0146] \quad F = \sum_{i=1}^M \sum_{j=1}^N E(i, j) \times W(i, j) \quad (14)$$

[0147] 其中, E(i, j) 为输入矩阵的第i行第j列的元素, W(i, j) 为权重矩阵的第i行第j列的元素, 输入矩阵E和权重矩阵W的大小均为M×N。

[0148] 假设E_f为全连接算子的浮点数输入矩阵, W_f为浮点数权重矩阵, F_f为浮点数全连接计算结果, 基于公式(14), E_f、W_f、F_f间的关系可以表示为如下公式(15)。

$$[0149] \quad F_f = \sum_{i=1}^M \sum_{j=1}^N E(i, j)_f \times W(i, j)_f \quad (15)$$

[0150] 根据公式(6)可以得到E_f = (E_q - Z_p_E) / E_ASc, W_f = (W_q - Z_p_W) / W_ASc, F_f = (F_q - Z_p_F) / F_ASc, 其中: E_q为E_f对应的非对称量化的定点数矩阵, E_ASc为将E_f量化为E_q的非对称量化缩放系数, Z_p_E为将E_f量化为E_q的非对称量化零点; W_q为W_f对应的非对称量化的定点数权重矩阵, W_ASc为将W_f量化为W_q的非对称量化缩放系数, Z_p_W为将W_f量化为W_q的非对称量化零点; F_q为F_f对应的非对称量化的定点数, F_ASc为将F_f量化为F_q的非对称量化缩放系数, Z_p_F为将F_f量化为F_q的非对称量化零点。进而公式(15)可以表示为如下公式(16)。

$$[0151] \quad \frac{F_q - Z_p_F}{F_ASc} = \sum_{i=1}^M \sum_{j=1}^N \frac{E_q(i, j) - Z_p_E}{E_ASc} \times \frac{W_q(i, j) - Z_p_W}{W_ASc} \quad (16)$$

[0152] 对公式(16)进行变形可以得到如下公式(17)所示的非对称量化的全连接算子的运算逻辑。

$$[0153] \quad F_q = \text{round} \left\{ \frac{F_ASc \sum_{i=1}^M \sum_{j=1}^N [E_q(i, j) - Z_p_E] \times [W_q(i, j) - Z_p_W]}{E_ASc \times W_ASc} + Z_p_F \right\} \quad (17)$$

[0154] NPU在执行公式(17)所示的全连接算子的运算逻辑时, 获取非对称量化的输入数据E_q(例如前述A_i' (i=1, 2, 3))及输入数据的非对称量化缩放系数E_ASc(例如由前述步骤S602得到的AS_{c_in3})、输入数据的非对称量化零点Z_p_E(例如由前述步骤S602得到的Z_{p_in3})、非对称量化的权重矩阵W_q(例如由前述步骤S603得到的权重矩阵W')、权重矩阵的非对称量化缩放系数W_ASc(例如前述AS_{c_in3})、权重矩阵的非对称量化零点Z_p_W(例如由前述步骤S602得到的Z_{p_in3})、输出数据的非对称量化缩放系数F_ASc(例如由前述步骤S602得到的AS_{c_out3})、输出数据的非对称量化零点Z_p_F(例如由前述步骤S602得到的Z_{p_out3}), 再根据前述公式(17)得到A_i'和W'的全连接计算结果, 例如在E_q=A₁'的情况下, 可以得到前述in1'。

[0155] 可以理解, 由于NPU中并没有可以直接实现除法运算的电路, 在一些实施例中, 公式(17)中的除法运算可以通过移位和乘法来实现, 以提高NPU运行全连接算子的速度。例如, 假设E_ASc×W_ASc=0.0417²=0.00174, 可以将0.00174表示为1/9×2⁻⁶, 从而将

$F_ASc \sum_{i=1}^M \sum_{j=1}^N [E_q(i,j) - Z_p_E] \times [W_q(i,j) - Z_p_W]$ 的结果对应的二进制数向右移-6位

(即向左移6位)再乘以9,即可得到 $\frac{F_ASc \sum_{i=1}^M \sum_{j=1}^N [E_q(i,j) - Z_p_E] \times [W_q(i,j) - Z_p_W]}{E_ASc \times W_SAC}$ 的

运算结果,最后将该运算结果加上 Z_p_F 即可得到全连接计算结果 F_q 。

[0156] 可以理解,在另一些实施例中全连接算子也可以采用其他运算逻辑,本申请实施例不做限定。

[0157] 可以理解,对于其他的非对称量化的算子,运算逻辑可以通过相似的方法得到,在此不做赘述。

[0158] 进一步,图8根据本申请的一些实施例,示出了一种电子设备100的结构示意图。如图8所示,电子设备100包括一个或多个处理器101A、NPU 101B系统内存102、非易失性存储器(Non-Volatile Memory, NVM) 103、通信接口104、输入/输出(I/O)设备105、以及用于耦接处理器101A、系统内存102、非易失性存储器103、通信接口104和输入/输出(I/O)设备105的系统控制逻辑106。其中:

[0159] 处理器101A可以包括一个或多个处理单元,例如,可以包括中央处理器CPU(Central Processing Unit)、图像处理GPU(Graphics Processing Unit)、数字信号处理器DSP(Digital Signal Processor)、微处理器MCU(Micro-programmed Control Unit)、AI(Artificial Intelligence,人工智能)处理器或可编程逻辑器件FPGA(Field Programmable Gate Array)的处理模块或处理电路可以包括一个或多个单核或多核处理器。

[0160] 神经网络处理器101B可以用于调用预设的非对称量化的算子的运算逻辑,实现神经网络模型的推理。神经网络处理器101B可以是独立的处理器,也可以集成于处理器101A内部。在一些实施例中,NPU可以用于运行本申请实施例提供的神经网络模型的运行方法对应的指令。

[0161] 系统内存102是易失性存储器,例如随机存取存储器(Random-Access Memory, RAM),双倍数据率同步动态随机存取存储器(Double Data Rate Synchronous Dynamic Random Access Memory, DDR SDRAM)等。系统内存用于临时存储数据和/或指令,例如,在一些实施例中,系统内存102可以用于存储上述神经网络模型10的相关指令、非对称/对称量化参数、非对称/对称量化常量数据等,也可以用于存储预设的非对称量化的算子的运算逻辑。

[0162] 非易失性存储器103可以包括用于存储数据和/或指令的一个或多个有形的、非暂时性的计算机可读介质。在一些实施例中,非易失性存储器103可以包括闪存等任意合适的非易失性存储器和/或任意合适的非易失性存储设备,例如硬盘驱动器(Hard Disk Drive, HDD)、光盘(Compact Disc, CD)、数字通用光盘(Digital Versatile Disc, DVD)、固态硬盘(Solid-State Drive, SSD)等。在一些实施例中,非易失性存储器103也可以是可移动存储介质,例如安全数字(Secure Digital, SD)存储卡等。在另一些实施例中,非易失性存储器103可以用于存储上述神经网络模型10的相关指令、非对称/对称量化参数、非对称/对称量化常量数据等,也可以用于存储预设的非对称量化的算子的运算逻辑。

[0163] 特别地,系统内存102和非易失性存储器103可以分别包括:指令107的临时副本和永久副本。指令107可以包括:由处理器101A和/或神经网络处理器101B中的至少一个执行时使电子设备100实现本申请各实施例提供的神经网络模型的运行方法。

[0164] 通信接口104可以包括收发器,用于为电子设备100提供有线或无线通信接口,进而通过一个或多个网络与任意其他合适的设备进行通信。在一些实施例中,通信接口104可以集成于电子设备100的其他组件,例如通信接口104可以集成于处理器101A中。在一些实施例中,电子设备100可以通过通信接口104和其他设备通信,例如,电子设备100可以通过通信接口104从其他电子设备获取待运行的神经网络模型。

[0165] 输入/输出(I/O)设备105可以包括输入设备如键盘、鼠标等,输出设备如显示器等,用户可以通过输入/输出(I/O)设备105与电子设备100进行交互。

[0166] 系统控制逻辑106可以包括任意合适的接口控制器,以电子设备100的其他模块提供任意合适的接口。例如在一些实施例中,系统控制逻辑106可以包括一个或多个存储器控制器,以提供连接到系统内存102和非易失性存储器103的接口。

[0167] 在一些实施例中,处理器101A中的至少一个可以与用于系统控制逻辑106的一个或多个控制器的逻辑封装在一起,以形成系统封装(System in Package, SiP)。在另一些实施例中,处理器101A中的至少一个还可以与用于系统控制逻辑106的一个或多个控制器的逻辑集成在同一芯片上,以形成片上系统(System-on-Chip, SoC)。

[0168] 可以理解,电子设备100可以是能够运行神经网络模型的任意电子设备,包括但不限于手机、可穿戴设备(如智能手表等)、平板电脑、桌面型、膝上型、手持计算机、笔记本电脑、超级移动个人计算机(ultra-mobile personal computer, UMPC)、上网本,以及蜂窝电话、个人数字助理(personal digital assistant, PDA)、增强现实(augmented reality, AR)/虚拟现实(virtual reality, VR)设备等,本申请实施例不做限定。

[0169] 可以理解,图8所示的电子设备100的结构只是一种示例,在另一些实施例中,电子设备100可以包括比图示更多或更少的部件,或者组合某些部件,或者拆分某些部件,或者不同的部件布置。图示的部件可以以硬件,软件或软件和硬件的组合实现。

[0170] 本申请公开的机制的各实施例可以被实现在硬件、软件、固件或这些实现方法的组合中。本申请的实施例可实现为在可编程系统上执行的计算机程序或程序代码,该可编程系统包括至少一个处理器、存储系统(包括易失性和非易失性存储器和/或存储元件)、至少一个输入设备以及至少一个输出设备。

[0171] 可将程序代码应用于输入指令,以执行本申请描述的各功能并生成输出信息。可以按已知方式将输出信息应用于一个或多个输出设备。为了本申请的目的,处理系统包括具有诸如例如数字信号处理器(Digital Signal Processor, DSP)、微控制器、专用集成电路(Application Specific Integrated Circuit, ASIC)或微处理器之类的处理器的任何系统。

[0172] 程序代码可以用高级程序化语言或面向对象的编程语言来实现,以便与处理系统通信。在需要时,也可用汇编语言或机器语言来实现程序代码。事实上,本申请中描述的机制不限于任何特定编程语言的范围。在任一情形下,该语言可以是编译语言或解释语言。

[0173] 在一些情况下,所公开的实施例可以以硬件、固件、软件或其任何组合来实现。所公开的实施例还可以被实现为由一个或多个暂时或非暂时性机器可读(例如,计算机可读)

存储介质承载或存储在其上的指令,其可以由一个或多个处理器读取和执行。例如,指令可以通过网络或通过其他计算机可读介质分发。因此,机器可读介质可以包括用于以机器(例如,计算机)可读的形式存储或传输信息的任何机制,包括但不限于,软盘、光盘、光碟、只读存储器(CD-ROMs)、磁光盘、只读存储器(Read Only Memory,ROM)、随机存取存储器(Random Access Memory, RAM)、可擦除可编程只读存储器(Erasable Programmable Read Only Memory, EPROM)、电可擦除可编程只读存储器(Electrically Erasable Programmable Read-Only Memory, EEPROM)、磁卡或光卡、闪存、或用于利用因特网以电、光、声或其他形式的传播信号来传输信息(例如,载波、红外信号数字信号等)的有形的机器可读存储器。因此,机器可读介质包括适合于以机器(例如计算机)可读的形式存储或传输电子指令或信息的任何类型的机器可读介质。

[0174] 在附图中,可以以特定布置和/或顺序示出一些结构或方法特征。然而,应该理解,可能不需要这样的特定布置和/或排序。而是,在一些实施例中,这些特征可以以不同于说明性附图中所示的方式和/或顺序来布置。另外,在特定图中包括结构或方法特征并不意味着暗示在所有实施例中都需要这样的特征,并且在一些实施例中,可以不包括这些特征或者可以与其他特征组合。

[0175] 需要说明的是,本申请各设备实施例中提到的各单元/模块都是逻辑单元/模块,在物理上,一个逻辑单元/模块可以是一个物理单元/模块,也可以是一个物理单元/模块的一部分,还可以以多个物理单元/模块的组合实现,这些逻辑单元/模块本身的物理实现方式并不是最重要的,这些逻辑单元/模块所实现的功能的组合才是解决本申请所提出的技术问题的关键。此外,为了突出本申请的创新部分,本申请上述各设备实施例并没有将与解决本申请所提出的技术问题关系不太密切的单元/模块引入,这并不表明上述设备实施例并不存在其它的单元/模块。

[0176] 需要说明的是,在本专利的示例和说明书中,术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含,从而使得包括一系列要素的过程、方法、物品或者设备不仅包括那些要素,而且还包括没有明确列出的其他要素,或者是还包括为这种过程、方法、物品或者设备所固有的要素。在没有更多限制的情况下,由语句“包括一个”限定的要素,并不排除在包括所述要素的过程、方法、物品或者设备中还存在另外的相同要素。

[0177] 虽然通过参照本申请的某些优选实施例,已经对本申请进行了图示和描述,但本领域的普通技术人员应该明白,可以在形式上和细节上对其作各种改变,而不偏离本申请的精神和范围。

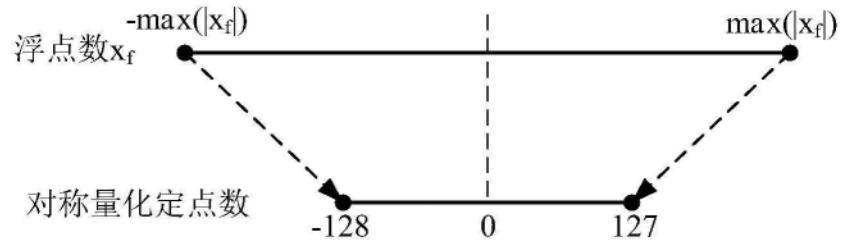


图1A

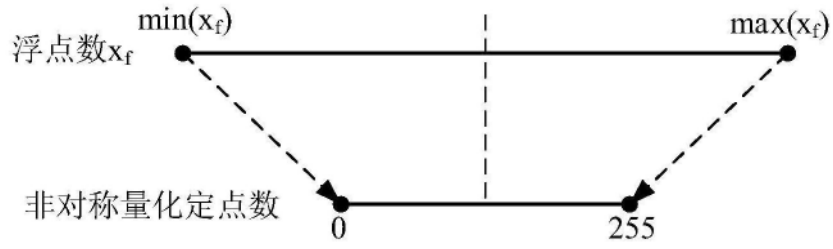


图1B

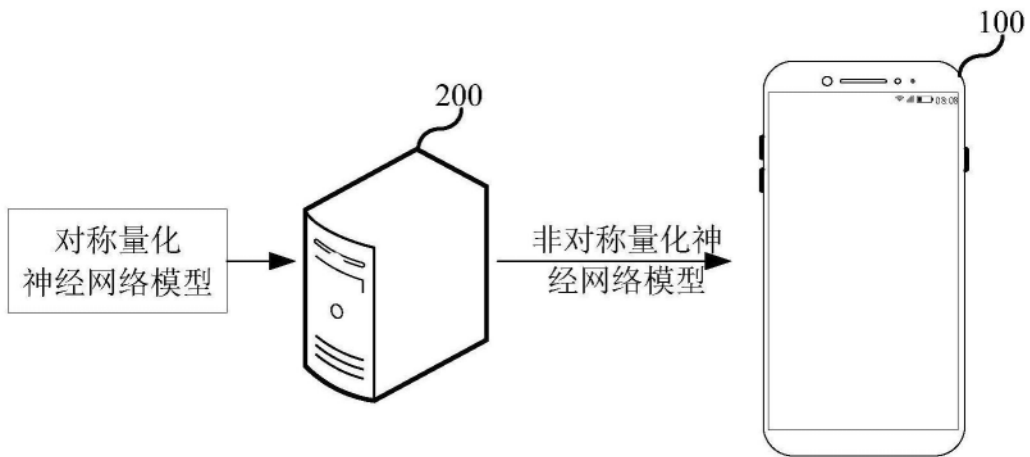


图2

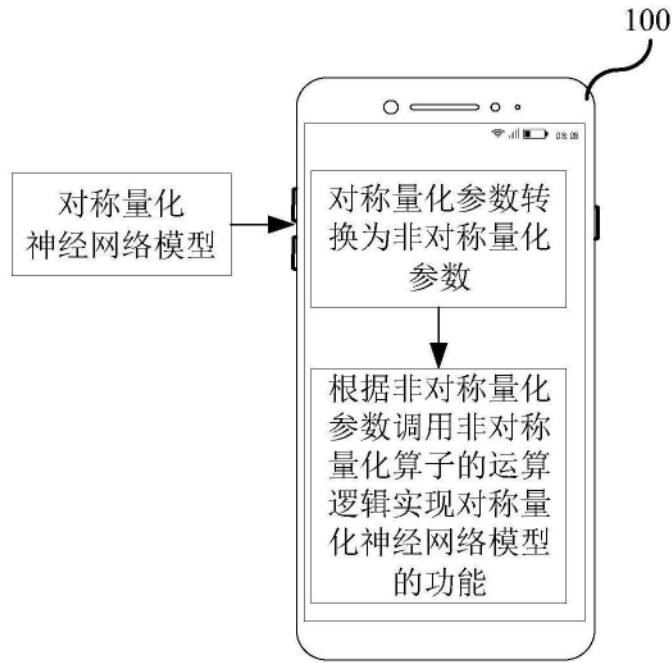


图3

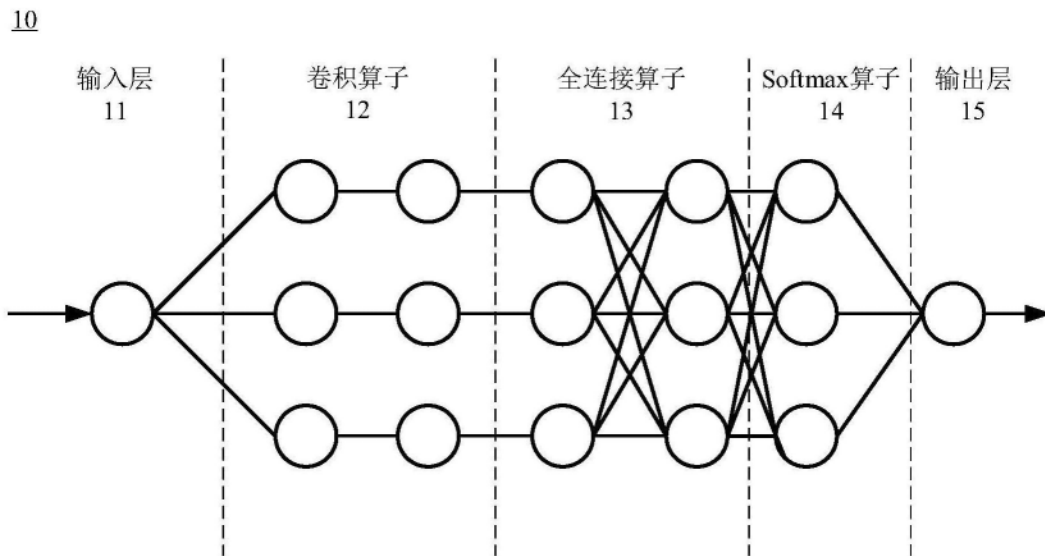


图4

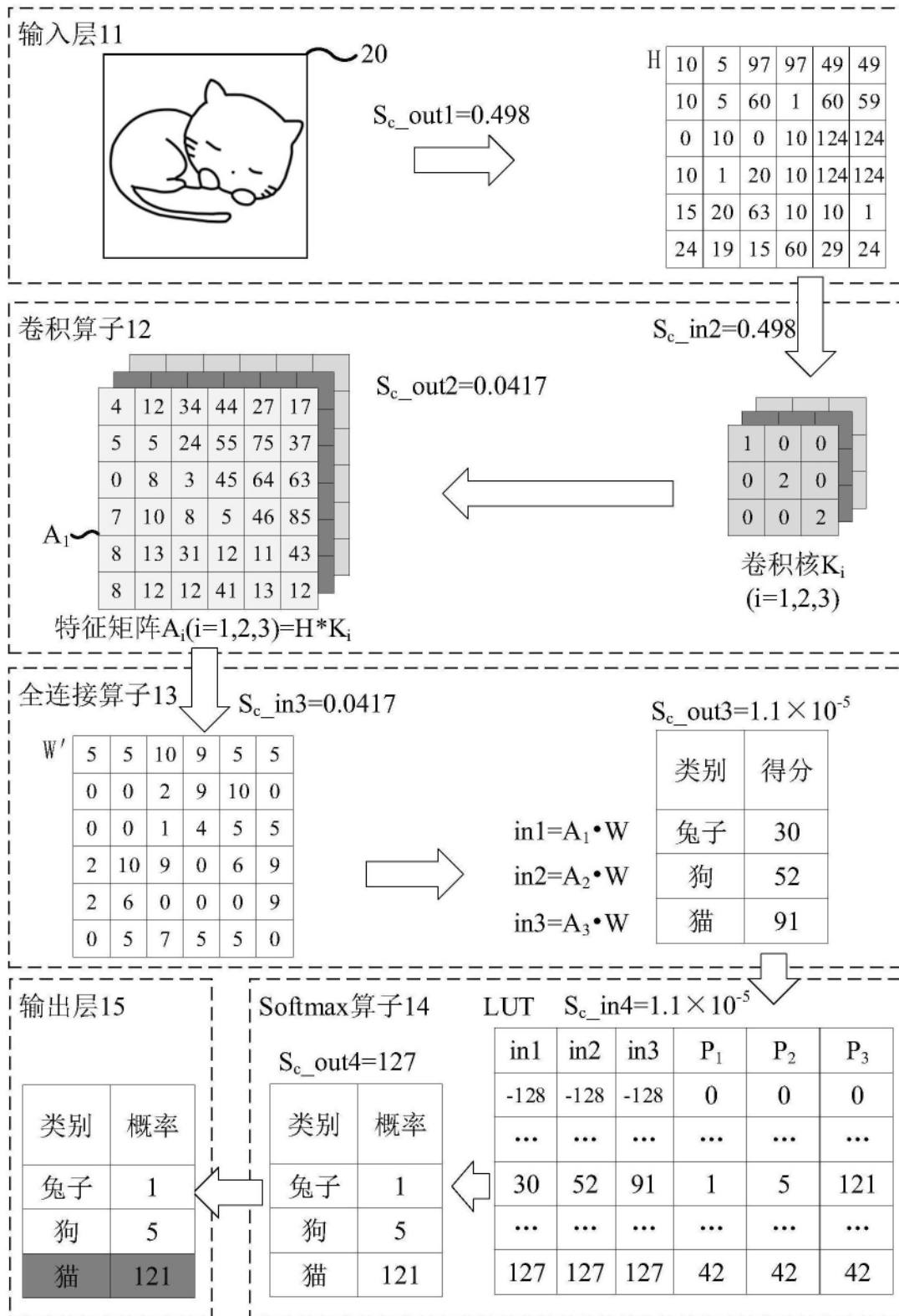


图5

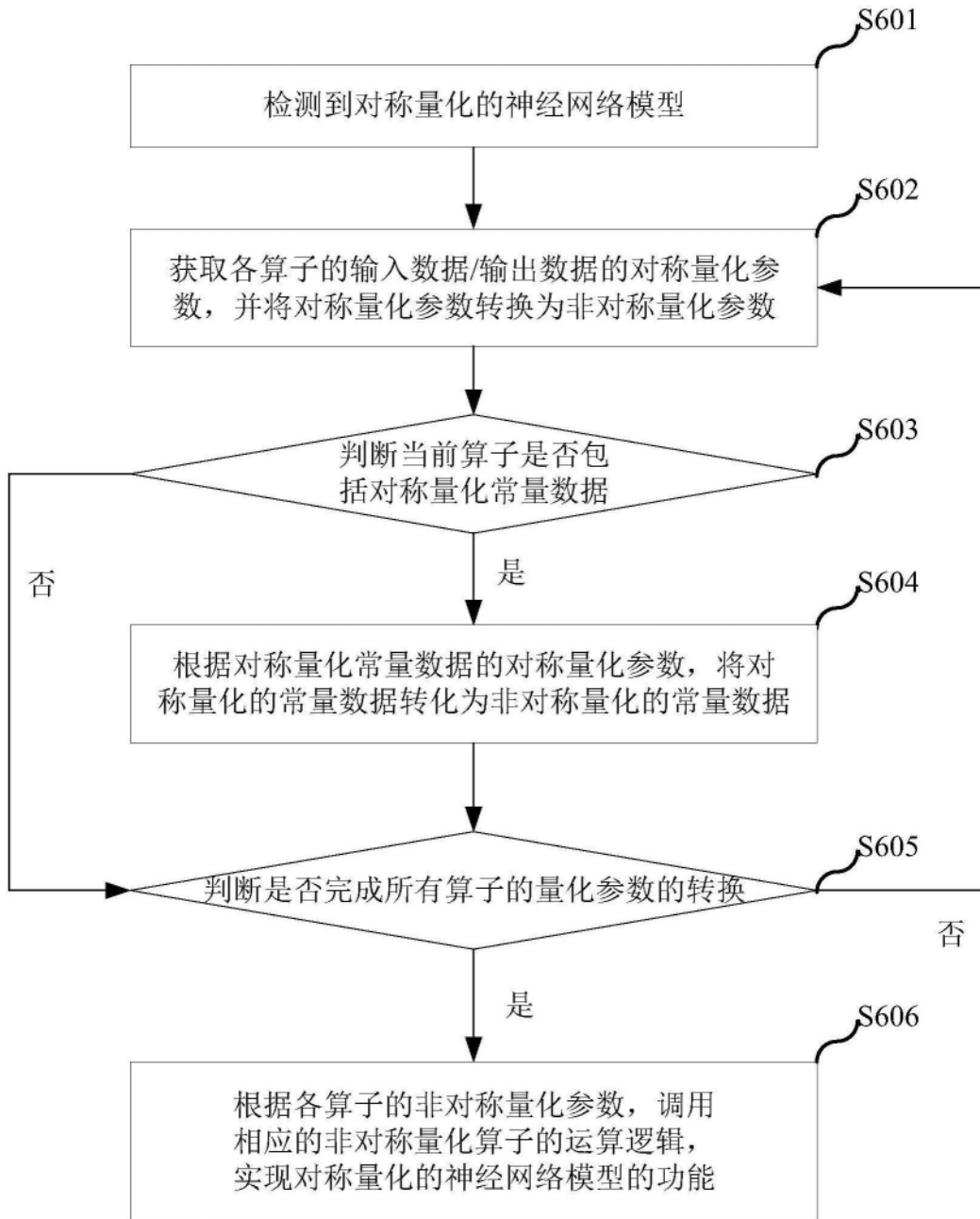


图6

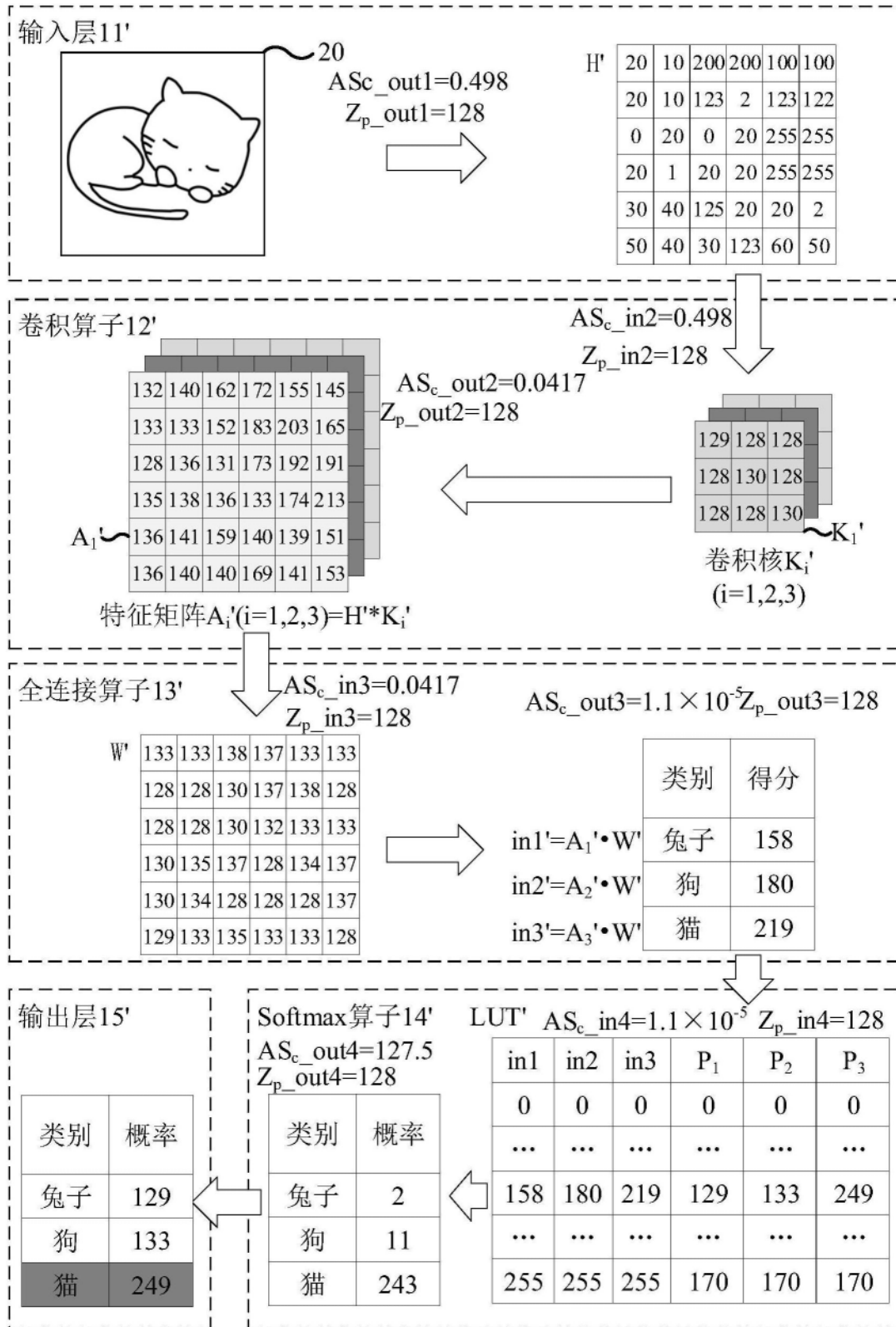


图7

100

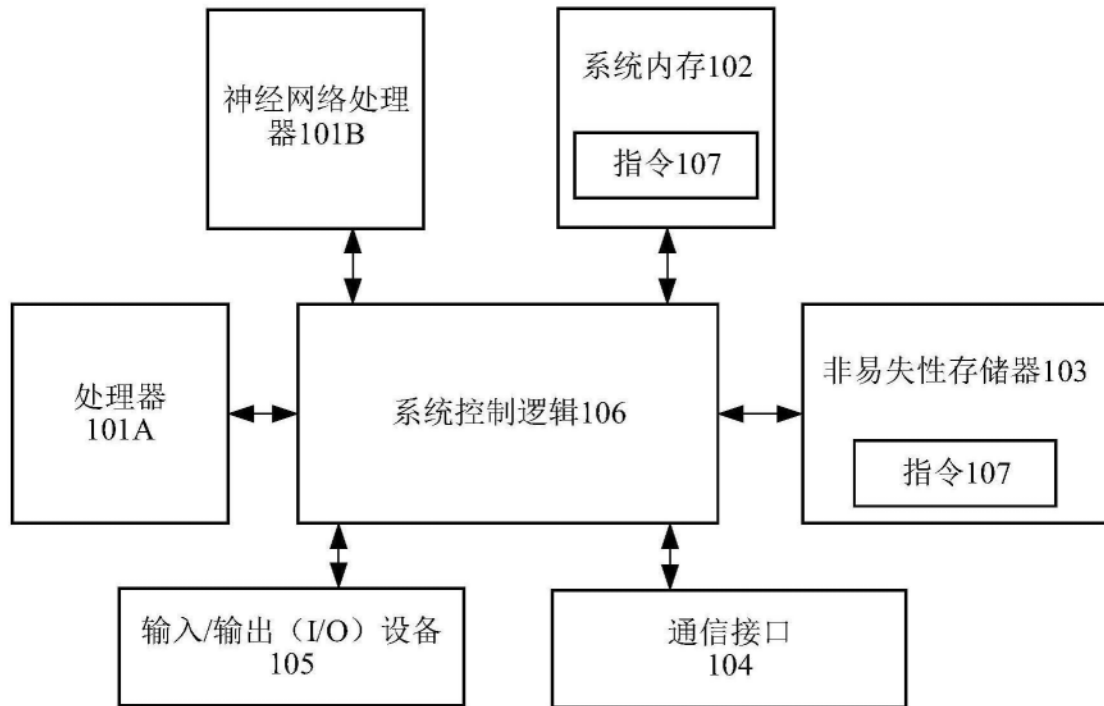


图8