

(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国际局

(43) 国际公布日
2019 年 4 月 25 日 (25.04.2019)



(10) 国际公布号
WO 2019/075604 A1

- (51) 国际专利分类号:
G06N 3/08 (2006.01) **G06F 9/318** (2006.01)
G06F 17/11 (2006.01)
- (21) 国际申请号: PCT/CN2017/106333
- (22) 国际申请日: 2017 年 10 月 16 日 (16.10.2017)
- (25) 申请语言: 中文
- (26) 公布语言: 中文
- (71) 申请人: 深圳市大疆创新科技有限公司 (SZ DJI TECHNOLOGY CO., LTD.) [CN/CN]; 中国广东省深圳市南山区高新区南区粤兴一道9号香港科大深圳产学研大楼6楼, Guangdong 518057 (CN)。
- (72) 发明人: 李似锦 (LI, Sijin); 中国广东省深圳市南山区高新区南区粤兴一道9号香港科大深圳产学研大楼6楼, Guangdong 518057 (CN)。 杨康 (YANG, Kang); 中国广东省深圳市南山区高新区南区粤兴

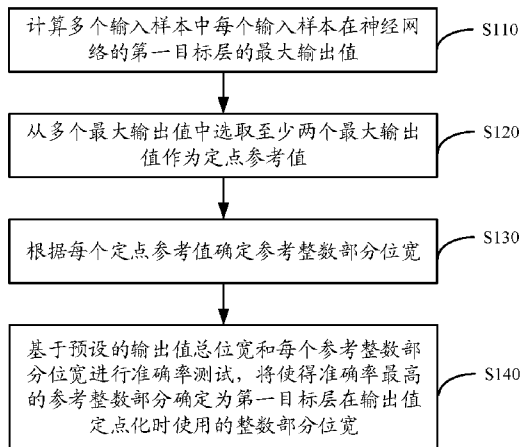
一道9号香港科大深圳产学研大楼6楼, Guangdong 518057 (CN)。 林蔓虹 (LIN, Manhong); 中国广东省深圳市南山区高新区南区粤兴一道9号香港科大深圳产学研大楼6楼, Guangdong 518057 (CN)。 颜钊 (YAN, Zhao); 中国广东省深圳市南山区高新区南区粤兴一道9号香港科大深圳产学研大楼6楼, Guangdong 518057 (CN)。

(74) 代理人: 北京龙双利达知识产权代理有限公司 (LONGSUN LEAD IP LTD.); 中国北京市海淀区北清路68号院3号楼101, Beijing 100094 (CN)。

(81) 指定国 (除另有指明, 要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK,

(54) Title: DATA FIXED-POINT METHOD AND DEVICE

(54) 发明名称: 数据定点化方法和装置



- S110 Calculate the maximum output value of each input sample in a plurality of input samples in a first target layer of a neural network
- S120 Select at least two maximum output values from among a plurality of maximum output values as fixed-point reference values
- S130 Determine a reference integer part bit width according to each fixed-point reference value
- S140 Conduct an accuracy rate test based on a pre-set total output value bit width and each reference integer part bit width, so that the reference integer part with the highest accuracy rate is determined to be the integer part bit width used by the first target layer during output value fixed-point processing

图 2

(57) Abstract: A data fixed-point method and device, a data processing method and device, and a data alignment method and device. The data fixed-point method comprises: calculating the maximum output value of each input sample in a plurality of input samples in a first target layer of a neural network (S110); selecting at least two maximum output values from among a plurality of maximum output values as fixed-point reference values (S120); determining a reference integer part bit width according to each fixed-point reference value (S130); and conducting an accuracy rate test based on a pre-set total output value bit width and each reference integer part bit width, so that the reference integer part with the highest accuracy rate is determined to be the integer part bit width used by the first target layer during the output value fixed-point processing (S140). According to the method, a plurality of values are selected from among a plurality of maximum output values of a first target layer as fixed-point reference values, a reference integer part bit width corresponding to each fixed-point reference value is determined, and the optimal integer part bit width is determined based on an accuracy rate test,

LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX,
MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL,
PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL,
SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG,
US, UZ, VC, VN, ZA, ZM, ZW。

(84) 指定国 (除另有指明, 要求每一种可提供的地区
保护): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ,
NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM,
AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG,
CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU,
IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT,
RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI,
CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

根据细则4.17的声明:

- 关于申请人有权申请并被授予专利 (细则
4.17(ii))

本国际公布:

- 包括国际检索报告 (条约第21条(3))。

such that a network can, after fixed-point processing is carried out, transfer more useful information and can also maintain a higher accuracy according to the optimal integer part bit width, thereby improving the expressive power and accuracy rate of the network.

(57) 摘要: 一种数据定点化方法和装置、数据处理方法和装置以及数据对齐方法和装置, 其中数据定点化方法包括: 计算多个输入样本中每个输入样本在神经网络的第一目标层的最大输出值 (S110); 从多个最大输出值中选取至少两个最大输出值作为定点参考值 (S120); 根据每个定点参考值确定参考整数部分位宽 (S130); 基于预设的输出值总位宽和每个参考整数部分位宽进行准确率测试, 将使得准确率最高的参考整数部分确定为第一目标层在输出值定点化时使用的整数部分位宽 (S140)。该方法通过从第一目标层的多个最大输出值中选取多个值作为定点参考值, 确定每个定点参考值对应的参考整数部分位宽, 并基于准确率测试确定最优的整数部分位宽, 根据最优的整数部分位宽使得定点化后网络可以在传递更多的有用信息同时保持较高的精度, 提高网络的表达能力和准确率。

数据定点化方法和装置

版权申明

- 5 本专利文件披露的内容包含受版权保护的材料。该版权为版权所有人所有。版权所有人不反对任何人复制专利与商标局的官方记录和档案中所存在的该专利文件或者该专利披露。

技术领域

- 10 本申请涉及数据处理领域，尤其涉及数据定点化方法和装置。

背景技术

- 当前主流的神经网络计算框架中，基本都是利用浮点数进行训练计算的。其中，神经网络的反向传播过程中，梯度的计算需要基于浮点数表示，
15 以保证足够的精度；神经网络的前向传播过程的各层，尤其是卷积层和全连接层的权重系数和各层的输出值也均以浮点数表示。但是，前向传播过程中，基于浮点数的运算相比基于定点数的运算的逻辑设计更复杂，会消耗更多的硬件资源，功耗也更高。基于定点数的硬件逻辑设计相较于基于浮点数的硬件逻辑设计更友好。
- 20 业界的相关公司通常通过最小化数值误差将训练时用浮点数表示的各层输出值和权重系数转换为定点数表示。即，为输出值设置优化目标函数，根据优化目标函数，在给定位宽的条件下，找到使得输出值定点截断后得到的数与浮点数误差最小时的小数位位宽。权重系数的定点化也以类似的原理实现。然而，以优化目标函数的误差最小确定定点位置，得到的定点化结果
25 可能很糟糕。仍以输出值为例，其主要原因在于，输出值中最重要的信息往往是由数值比较大的输出值决定的，而其占比通常比较小。以该定点化方法得到的定点位置进行定点化时，虽然截断率比较低，但往往会把最有用的高位信息剔除掉，从而影响网络的表达能力，造成网络的准确率下降。

30 发明内容

本申请提供了一种数据定点化方法和装置，使得定点化后网络可以在传

递更多的有用信息同时保持较高的精度，从而提高网络的表达能力和准确率。

第一方面，提供了一种数据定点化方法，包括：计算多个输入样本中每个所述输入样本在神经网络的第一目标层的最大输出值；从多个所述最大输出值中选取至少两个所述最大输出值作为定点参考值；根据每个所述定点参考值确定参考整数部分位宽；基于预设的输出值总位宽和每个所述参考整数部分位宽进行准确率测试，将使得准确率最高的参考整数部分确定为所述第一目标层在输出值定点化时使用的整数部分位宽。

第一方面的数据定点化方法通过从第一目标层的多个最大输出值中选取多个值作为定点参考值，确定每个定点参考值对应的参考整数部分位宽，并基于准确率测试确定最优的整数部分位宽，根据最优的整数部分位宽使得定点化后网络可以在传递更多的有用信息同时保持较高的精度，可以提高网络的表达能力和准确率。

第二方面，提供了一种数据定点化方法，包括：计算输入样本在神经网络的第一目标层的参考输出值；确定为输出值预设的输出值总位宽和预设的第一符号位宽；根据所述参考输出值的大小确定输出值整数部分位宽；根据所述输出值总位宽、所述第一符号位宽和所述输出值整数部分位宽，确定输出值小数部分位宽，其中，所述第一符号位宽、所述输出值整数部分位宽和所述输出值小数部分位宽作为所述第一目标层在输出值定点化时使用的符号位宽、整数部分位宽和小数部分位宽。

第二方面的数据定点化方法在输出值定点化时考虑符号位，这样使得确定的定点化方案更优，提高增加网络准确率的可能性。

第三方面，提供了一种数据处理方法，包括：对神经网络的至少两层进行合并预处理；基于合并预处理后的神经网络进行神经网络运算。

第三方面的数据处理方法，对神经网络的至少两层进行合并预处理，基于合并预处理后的神经网络进行运算，可以节省计算资源，提高系统效率。

第四方面，提供了一种数据对齐方法，包括：从神经网络中确定需数据对齐的多个层；根据所述多个层中每层输出值定点化应使用的整数部分位宽，确定所述多个层输出值定点化最终使用的整数部分位宽，其中，所述多个层中任意两层输出值定点化最终使用的整数部分位宽相等。

第四方面的数据对齐方法可以在确定定点化方案时解决一些层有输入数据对齐需求的问题，可以减少硬件资源的使用，提高系统效率。

第五方面，提供了一种数据定点化装置，包括：前向传播计算模块，用于计算多个输入样本中每个所述输入样本在神经网络的第一目标层的最大输出值；定点参考选取模块，用于从所述前向传播计算模块得到的多个所述最大输出值中选取至少两个所述最大输出值作为定点参考值；参考位宽确定模块，用于根据所述定点参考选取模块选取的每个所述定点参考值确定参考整数部分位宽；准确率测试模块，用于基于预设的输出值总位宽和所述参考位宽确定模块确定的每个所述参考整数部分位宽进行准确率测试，将使得准确率最高的参考整数部分确定为所述第一目标层在输出值定点化时使用的整数部分位宽。

第六方面，提供了一种数据定点化装置，包括：前向传播计算模块，用于计算输入样本在神经网络的第一目标层的参考输出值；确定模块，用于确定为输出值预设的输出值总位宽和预设的第一符号位宽；输出值位宽确定模块，用于根据所述前向传播计算模块得到的所述参考输出值的大小确定输出值整数部分位宽；根据所述确定模块确定的所述输出值总位宽、所述第一符号位宽和所述输出值整数部分位宽，确定输出值小数部分位宽，其中，所述第一符号位宽、所述输出值整数部分位宽和所述输出值小数部分位宽作为所述第一目标层在输出值定点化时使用的符号位宽、整数部分位宽和小数部分位宽。

第七方面，提供了一种数据处理装置，包括：预处理模块，用于对神经网络的至少两层进行合并预处理；运算模块，用于基于所述预处理模块合并预处理后的神经网络进行神经网络运算。

第八方面，提供了一种数据对齐装置，包括：第一确定模块，用于从神经网络中确定需数据对齐的多个层；第二确定模块，用于根据所述多个层中每层输出值定点化应使用的整数部分位宽，确定所述多个层输出值定点化最终使用的整数部分位宽，其中，所述多个层中任意两层输出值定点化最终使用的整数部分位宽相等。

第九方面，提供了一种数据定点化方法，包括：计算多个输入样本中每个所述输入样本在神经网络的第一目标层的最大输出值；从多个所述最大输出值中选取一个所述最大输出值作为定点参考值；根据所述定点参考

值确定参考整数部分位宽，作为所述第一目标层在输出值定点化时使用的整数部分位宽。

第十方面，提供了一种数据定点化装置，包括：前向传播计算模块，用于计算多个输入样本中每个所述输入样本在神经网络的第一目标层的最大输出值；定点参考选取模块，用于从所述前向传播计算模块得到的多个所述最大输出值中选取一个所述最大输出值作为定点参考值；位宽确定模块，用于根据所述定点参考选取模块选取的所述定点参考值确定参考整数部分位宽，作为所述第一目标层在输出值定点化时使用的整数部分位宽。

附图说明

图 1 是深度卷积神经网络的框架示意图。

图 2 是本申请一个实施例的数据定点化方法的示意性流程图。

图 3A、图 3B 和图 3C 是本申请实施例的合并预处理的几种情况的示意图；图 3D 是卷积层之后是 BN 层的一种层连接方式的示意图。

图 4 是本申请一个实施例的选取定点参考值的示意图。

图 5 是 Concatenation 层的工作原理的示意图。

图 6 是本申请一个实施例的后处理的示意图。

图 7 是本申请一个实施例的数据定点化方法的示意性流程图。

图 8 是本申请一个实施例的数据定点化方法的示意性流程图。

图 9 是本申请一个实施例的数据处理方法的示意性流程图。

图 10 是本申请一个实施例的数据对齐方法的示意性流程图。

图 11 是本申请一个实施例的数据定点化装置的示意性框图。

图 12 是本申请一个实施例的数据定点化装置的示意性框图。

图 13 是本申请一个实施例的数据处理装置的示意性框图。

图 14 是本申请一个实施例的数据对齐装置的示意性框图。

具体实施方式

下面将结合附图，对本申请实施例中的技术方案进行描述。

除非另有定义，本文所使用的所有的技术和科学术语与属于本申请的技术领域的技术人员通常理解的含义相同。本文中在本申请的说明书中所使用的术语只是为了描述具体的实施例的目的，不是旨在于限制本申请。

首先介绍本申请实施例涉及的相关技术及概念。

神经网络（以深度卷积神经网络（Deep Convolutional Neural Network, DCNN）为例）：

图 1 是深度卷积神经网络的框架示意图。深度卷积神经网络的输入值（由输入层输入），经隐藏层进行卷积（convolution）、转置卷积（transposed convolution or deconvolution）、归一化（Batch Normalization, BN）、缩放（Scale）、全连接（fully connected）、拼接（Concatenation）、池化（pooling）、元素智能加法（element-wise addition）和激活（activation）等运算后，得到输出值（由输出层输出）。本申请实施例的神经网络的隐藏层可能涉及的运算不仅限于上述运算。

深度卷积神经网络的隐藏层可以包括级联的多层。每层的输入为上层的输出，为特征图（feature map），每层对输入的一组或多组特征图进行前述描述的至少一种运算，得到该层的输出。每层的输出也是特征图。一般情况下，各层以实现的功能命名，例如实现卷积运算的层称作卷积层。此外，隐藏层还可以包括转置卷积层、BN 层、Scale 层、池化层、全连接层、Concatenation 层、元素智能加法层和激活层等等，此处不进行一一列举。各层的具体运算流程可以参考现有的技术，本文不进行赘述。

应理解，每层（包括输入层和输出层）可以有一个输入和/或一个输出，也可以有多个输入和/或多个输出。在视觉领域的分类和检测任务中，特征图的宽高往往是逐层递减的（例如图 1 所示的输入、特征图#1、特征图#2、特征图#3 和输出的宽高是逐层递减的）；而在语义分割任务中，特征图的宽高在递减到一定深度后，有可能会通过转置卷积运算或上采样（upsampling）运算，再逐层递增。

通常情况下，卷积层的后面会紧接着一层激活层，常见的激活层有线性整流函数（Rectified Linear Unit, ReLU）层、S 型（sigmoid）层和双曲正切（tanh）层等。在 BN 层被提出以后，越来越多的神经网络在卷积之后会先进行 BN 处理，然后再进行激活计算。

当前，需要较多权重参数用于运算的层有：卷积层、全连接层、转置卷积层和 BN 层。

浮点数和定点数：

浮点数包括单精度浮点数（32 位）和双精度浮点数（64 位）。浮点数表

达为符号位、整数部分和小数部分。bw 为定点数总位宽，s 为符号位（通常置于最左位），fl 为小数部分位宽， x_i 是各位（也称为尾数（mantissa）位）的数值。一个定点数的实值可以表示为：

$$x = (-1)^s \times 2^{-fl} \times \sum_{i=0}^{bw-2} 2^i \times x_i$$

- 5 例如，一个定点数为 01000101，位宽为 8 位，最高位（0）为符号位，小数部分位宽 fl 为 3。那么这个定点数代表的实值为：

$$x = (-1)^0 \times 2^{-3} \times (2^0 + 2^2 + 2^6) = 8.625。$$

一种现有的定点化方法：

- 10 数据的定点化主要包括权重系数的定点化，以及卷积层或全连接层输出值的定点化。一种定点化方法是通过最小化数值误差实现的。

对于每层的权重系数定点化，可以有一个优化目标函数。权重系数的该优化目标函数是在给定总位宽的条件下，找到使得权重系数定点截断后得到的数与浮点数误差最小时的小数位位宽。

- 15 对于卷积层或全连接层的输出值定点化，也可以有一个优化目标函数。其定点化的原理同权重系数的定点化原理类似。

- 20 以优化目标函数的误差最小确定定点位置，得到的定点化结果可能很差。仍以输出值为例，其主要原因在于，输出值中最重要的信息往往是由数值比较大的输出值决定的，而其占比通常比较小。以该定点化方法得到的定点位置进行定点化时，虽然截断率比较低，但往往会把最有用的高位信息剔除掉，从而造成网络的准确率下降。

现有的定点化方法都没有考虑除卷积层和全连接层以外其他层的定点化处理，特别是激活层、池化层和 BN 层，这些层中都有可能涉及浮点运算，因此都需要考虑定点化处理。

- 25 现有的定点化方法没有考虑 element-wise addition 层和 Concatenation 层等层输入的数据小数点对齐的问题。这会造成在数据定点化之后运算的过程中，需对数据进行移位操作，使得操作过程较为复杂。

针对上述问题，本申请实施例提供了一种数据定点化方法 100，图 2 是该数据定点化方法 100 的示意性流程图。方法 100 包括：

- 30 S110，计算多个输入样本中每个输入样本在神经网络的第一目标层的最大输出值。

S120, 从多个最大输出值中选取至少两个最大输出值作为定点参考值。

S130, 根据每个定点参考值确定参考整数部分位宽。

S140, 基于预设的输出值总位宽和每个参考整数部分位宽进行准确率测试, 将使得准确率最高的参考整数部分确定为第一目标层在输出值定点化时使用的整数部分位宽。

本申请实施例通过从第一目标层的多个最大输出值中选取多个值作为定点参考值, 确定每个定点参考值对应的参考整数部分位宽, 并基于准确率测试确定最优的整数部分位宽, 根据最优的整数部分位宽使得定点化后网络可以在传递更多的有用信息同时保持较高的精度, 提高网络的表达能力和准确率。

应理解, 本申请实施例确定参考整数部分位宽之后, 基于预设的输出值总位宽就能够得到参考小数部分位宽。或者在其他实施例中, 可以先得到参考小数部分位宽, 再得到参考整数部分位宽, 本申请实施例对此不作限定。

在一些实施例中, 数据定点化后可以存在符号位(例如, 符号位的位宽为第一符号位宽)。第一符号位宽、参考整数部分位宽和参考小数部分位宽的和等于预设的输出值总位宽。

还应理解, 确定定点化方案后在定点化时, 根据待定点化数据的正负确定第一符号位; 根据待定点化数据的数值(大小)确定定点化后的整数部分和小数部分, 本申请实施例在此不进行赘述。

本申请实施例的第一目标层可以包括卷积层、转置卷积层、BN 层、Scale 层、池化层、全连接层、Concatenation 层、element-wise addition 层和激活层中的一层或至少两层合并后的层。即, 本申请实施例的数据定点化方法可以应用于神经网络的隐藏层的任一层或多层。

对应于第一目标层为至少两层合并后的层的情况, 数据定点化方法还可以包括: 对神经网络的至少两层进行合并预处理, 得到合并后形成的第一目标层。该处理可以认为是数据定点化方法的预处理部分。

神经网络的训练阶段完成后, 推理(inference)阶段的卷积层、BN 层和 Scale 层的参数是固定的。通过计算推导可以知道, 其实 BN 层和 Scale 层的参数是可以合并到卷积层的参数里, 这样神经网络的知识产权核(Intellectual Property core, IP 核)就不需要专门为 BN 层和 Scale 层设计专用电路了。

早期的神经网络中，卷积层之后为激活层。为了防止网络过拟合、加快收敛速度、增强网络泛化能力等，在卷积层之后激活层之前可以引入 BN 层。BN 层的输入包括 $B = \{x_1, \dots, x_m\} = \{x_i\}$ 以及参数 γ 和 β ，其中， x_i 既是卷积层的输出又是 BN 层的输入，参数 γ 和 β 在训练阶段进行计算，在推理阶段均为常数。BN 层的输出为 $\{y_i = BN_{\gamma, \beta}(x_i)\}$ 。

其中，

$$y_i \leftarrow \gamma \hat{x}_i + \beta \equiv BN_{\gamma, \beta}(x_i)$$

$$\hat{x}_i \leftarrow \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}}$$

$$\mu_B \leftarrow \frac{1}{m} \sum_{i=1}^m x_i$$

$$\sigma_B^2 \leftarrow \frac{1}{m} \sum_{i=1}^m (x_i - \mu_B)^2$$

因此 $\hat{x}_i$ 和 y_i 的计算可以化简为：

$$\hat{x}_i = \frac{x_i - \mu_B}{\alpha}$$

$$y_i = \gamma \frac{x_i - \mu_B}{\alpha} + \beta = \frac{\gamma}{\alpha} x_i + \beta - \frac{\gamma \times \mu_B}{\alpha} = ax_i + b$$

x_i 是卷积层的输出，令 X 为卷积层的输入， W 为权重系数矩阵， $\bar{b}$ 为偏置值则有：

$$x_i = WX + \bar{b}$$

$$y_i = aWX + \bar{b} + b = \tilde{W}X + \tilde{b}$$

由此，卷积层和 BN 层的合并完成。

Scale 层本身就是要计算 $y_i = ax_i + b$ ，参考 BN 层与卷积层的合并，也可将 Scale 层和卷积层合并。在 Caffe 框架下，BN 层的输出是 $\hat{x}_i$ ，因此基于 Caffe 框架设计的神经网络，通常会在 BN 层后加入 Scale 层以实现完整的归一化。

由此，对神经网络的至少两层进行合并预处理，得到合并后形成的第一目标层，可以包括：对神经网络的卷积层和 BN 层进行合并预处理，得到第一目标层；或，对神经网络的卷积层和 Scale 层进行合并预处理，得到第一目标层；或，对神经网络的卷积层、BN 层和 Scale 层进行合并预处理，得到第一目标层。

图 3A、图 3B 和图 3C 是本申请实施例的合并预处理的几种情况的示意图。图 3D 是最简单的卷积层之后是 BN 层的一种层连接方式。

如图 3A 所示，未进行合并预处理之前，卷积层后是 BN 层，再之后是激活层，将卷积层和 BN 层合并为第一目标层，其后为激活层，得到类似于图 3D 的两层结构。

应理解，某些 IP 核支持 Scale 层的处理，那么合并预处理中卷积层与 BN 层的合并，可以替换为卷积层与 Scale 层的合并。如图 3B 所示，未进行合并预处理之前，卷积层后是 Scale 层，再之后是激活层，将卷积层和 Scale 层合并为第一目标层，其后为激活层，得到类似于图 3D 的两层结构。

如图 3C 所示，未进行合并预处理之前，卷积层后是 BN 层，继而为 Scale 层，再之后是激活层，将卷积层、BN 层和 Scale 层合并为第一目标层，其后为激活层，得到类似于图 3D 的两层结构。

应理解，合并预处理后，S110 中的最大输出值为多个输入样本中每个输入样本在合并后形成的第一目标层的最大输出值。

通过数据定点化方法 100 的步骤 S110 至 S140，可以确定第一目标层的输出值的定点化位置。

S110，计算多个输入样本中每个输入样本在神经网络的第一目标层的最大输出值。具体地，多个输入样本构成输入数据集。对输入数据集的多个，例如 M 个样本进行前向传播计算，对待定点化的第一目标层记录对每个样本的最大输出值，得到 M 个最大值。其中，M 为大于或等于 2 的正整数。需要注意的是，前向传播计算时为保证计算精度，权重系数仍然可以使用浮点数。

S120，从多个最大输出值中选取至少两个最大输出值作为定点参考值，可以包括：对多个最大输出值进行排序，按照预设的选取参数从多个最大输出值中选取至少两个最大输出值作为定点参考值。应理解，选取参数可以在预设范围内。

具体而言，对多个最大输出值（例如，M 个最大输出值）进行排序，例如升序排列或者降序排列，或者依照某预设规则排序。排好序后，按照预设的选取参数（例如选取参数为选取排序后特定位置的值）从 M 个最大输出值中选取 N 个最大输出值。其中，N 为小于或等于 M 的正整数。

图 4 是本申请一个实施例的选取定点参考值的示意图。在该具体的例子

中, 将 M 个最大输出值按从小到大的顺序排列, 选取参数为 $a(j)$, 选取第 $a(j) \times M$ 个最大输出值作为定点参考值, 其中, j 的取值为 $1, \dots, N$, $a(j)$ 大于或等于 0 且小于或等于 1。例如, N 可以等于 10, $a(1), \dots, a(10)$ 分别为 0.5, 0.6, 0.7, 0.8, 0.9, 0.92, 0.94, 0.96, 0.98, 1。

5 在一些实施例中, 选取参数 $a(j)$ 可以为选取最大值和次大值。在又一些实施例中, 选取参数 $a(j)$ 的值可以为均匀的取值, 例如, 为 0.1, 0.2, 0.3, ..., 1 等, 这里对选取定点参考值的方法不做限定。

S130, 根据每个定点参考值确定参考整数部分位宽, 可以包括: 根据定点参考值的大小确定参考整数部分位宽。在一些实施例中, 方法 100 还可以包括: 确定预设的第一符号位宽和输出值总位宽; 根据第一符号位宽、输出值总位宽和参考整数部分位宽, 确定参考小数部分位宽。在本申请实施例中, 第一符号位和参考整数部分可以认为是参考非小数部分。换言之, 参考非小数部分位宽包括第一符号位宽 (通常而言第一符号位宽为 1) 和参考整数部分位宽。具体而言, 例如 N 个定点参考值中的第 j 个定点参考值为 O_j 。bwo 为对输出值预设的输出值总位宽。根据定点参考值 O_j 的大小确定参考非小数部分位宽, 例如参考非小数部分位宽 $iwoj = \text{ceil}(\log_2(O_j) + 1)$, 则定点参考值 O_j 对应的参考小数部分位宽 $fwoj = \text{bwo} - iwoj$, 其中, j 的取值为 $1, \dots, N$, $\text{ceil}()$ 表示向上取整。应理解, 参考非小数部分位宽中包括第一符号位宽 (第一符号位宽为 1) 和参考整数部分位宽 $iwoj - 1$ 。

20 在另一些实施例中, 数据定点化后不存在符号位。S130, 根据每个定点参考值确定参考整数部分位宽, 可以包括: 根据定点参考值的大小确定参考整数部分位宽。具体而言, 在本申请实施例中, 例如 N 个定点参考值中的第 j 个定点参考值为 O_j 。bwo 为对输出值预设的输出值总位宽。根据定点参考值 O_j 的大小确定参考整数部分位宽, 例如参考整数部分位宽 $iwoj = \text{ceil}(\log_2(O_j))$, 则定点参考值 O_j 对应的参考小数部分位宽 $fwoj = \text{bwo} - iwoj$, 其中, j 的取值为 $1, \dots, N$, $\text{ceil}()$ 表示向上取整。

S140, 基于预设的输出值总位宽和每个参考整数部分位宽进行准确率测试, 将使得准确率最高的参考整数部分确定为第一目标层在输出值定点化时使用的整数部分位宽。

30 具体而言, 第一目标层有 N 个可能定点化方案, 预测准确率损失最少的一个定点化方案。在图 4 的例子中, 在 $a(j)$ 等于 0.98 时, 即定点参考值为 127

时, 预测准确率损失最少。以存在符号位的情况为例, 则该第一目标层的非小数部分位宽 $iwoj$ 等于 8 (1 位符号位, 7 位整数位)。如果输出值总位宽为 16 位, 则小数部分位宽等于 $16-8=8$ 。

以上说明了确定输出值定点化方案的过程, 数据定点化方法还可以包括确定权重系数定点化方案的过程, 步骤包括: 确定为权重系数预设的权重定点总位宽和预设的第二符号位宽; 根据第一目标层中最大权重系数的大小确定权重整数部分位宽; 根据权重定点总位宽、第二符号位宽和权重整数部分位宽, 确定权重小数部分位宽, 其中, 将第二符号位宽、权重整数部分位宽和权重小数部分位宽确定为第一目标层在权重系数定点化时使用的符号位宽、整数部分位宽和小数部分位宽。

确定权重系数定点化方案的过程与确定输出值定点化方案的过程有类似之处。区别在于, 直接从第一目标层找到最大权重系数, 根据最大权重系数的大小确定权重非小数部分位宽即可。在一个具体的例子中, 为权重系数预设的权重定点总位宽可以为 bww 。计算第一目标层中最大权重系数 w 对应的权重非小数部分位宽 $iww=\text{ceil}(\log_2(w)+1)$, 其中包括第二符号位宽和权重整数部分位宽。从而, 最大权重系数 w 对应的权重小数部分位宽 $fwf=bww-iww$ 。将第二符号位宽 (通常为 1 位)、权重整数部分位宽 $iww-1$ 和权重小数部分位宽 fwf 确定为第一目标层在权重系数定点化时使用的符号位宽、整数部分位宽和小数部分位宽。

应理解, 如果存在合并预处理, 最大权重系数为对神经网络的至少两层进行合并预处理后形成的第一目标层的权重系数的最大值。

可选地, 本申请实施例可以包括后处理, 以解决一些层有输入数据小数点对齐需求的问题。因此, 需要其上的至少两层 (例如, 包括第一目标层和第二目标层) 的输出值小数点对齐, 数据定点化方法 100 还可以包括: 确定神经网络的第二目标层在输出值定点化时使用的整数部分位宽, 使第二目标层在输出值定点化时使用的整数部分位宽和第一目标层在输出值定点化时使用的整数部分位宽相等。

在系统预设的输出值总位宽一定的情况下, 由于第二目标层在输出值定点化时使用的整数部分位宽和第一目标层在输出值定点化时使用的整数部分位宽相等。应理解, 第二目标层在输出值定点化时使用的小数部分位宽和第一目标层在输出值定点化时使用的小数部分位宽也相等。

在第一目标层和第二目标层按照各自的输出值定点化方案确定出的定点化位置不同,即整数部分位宽不同时,确定神经网络的第二目标层在输出值定点化时使用的整数部分位宽,包括:将第一目标层和第二目标层输出值定点化应使用的整数部分位宽中的最大值,确定为第一目标层和第二目标层输出值定点化最终使用的整数部分位宽。例如,按照各自的输出值定点化方案,第一目标层的非小数部分位宽为 7 (第一符号位宽为 1, 整数部分位宽为 6), 第二目标层的非小数部分位宽为 5 (第一符号位宽为 1, 整数部分位宽为 4)。为了保证整数部分不被截断,第一目标层和第二目标层输出值定点化最终使用的非小数部分位宽可以为 7。非小数部分位宽 7 中,可以包括 1 位符号位 6 为整数位,如果预设的输出值总位宽为 16,则小数部分位宽为 9。

可选地,第一目标层的输出值与第二目标层的输出值在拼接 Concatenation 层和/或 element-wise addition 层进行后续处理。当然根据 IP 核支持的层种类不同,小数点对齐后的输出值也可以在其他层进行后续处理,本申请实施例对此不作限定。

具体而言,后处理主要是针对 Concatenation 层和 element-wise addition 层,使这两层的输入值(即输入特征图)的小数点位置是对齐的。Concatenation 层实现的功能是把两组输入特征图合并到一起,达到融合特征的效果,在计算机里可以理解为两块离散的内存块拼接成连续的内存块。图 5 示出了 Concatenation 层的工作原理的示意图。element-wise addition 层实现的功能是对两组输入特征图进行点加运算,计算残差特征图。由于这两层输入的两组特征图定点化后的小数点位置有可能是不一致的,因此,这两层需要对两组输入特征图的数值进行小数点对齐处理。虽然输入特征图的数值的小数点对齐可以由硬件通过移位实现,但这样做会浪费一定的硬件资源。Concatenation 层或 element-wise addition 层输入的两组特征图是某两层(例如,包括第一目标层和第二目标层)输出的特征图,在该两层输出的时候会做一次定点化处理,因此只需该两层的输出值保持小数点对齐即可。本申请实施例的后处理可以减少硬件资源的使用,提高系统效率。

图 6 是本申请一个实施例的后处理的示意图。现有的处理方案中,数据格式为 Q5.10 的特征图经卷积运算后,得到数据格式为 Q4.11 的特征图;数据格式为 Q4.11 的特征图经卷积运算后,得到数据格式为 Q6.9 的特征图;得到的数据格式为 Q4.11 的特征图经移位后将数据格式转换为数据格式

Q6.9, 与得到的数据格式为 Q6.9 的特征图作为 Concatenation 层的输入, 经 Concatenation 层运算后得到数据格式为 Q6.9 的特征图 (Concatenation 层的输出)。如图 6 所示, 本申请一个实施例的方案则为数据格式为 Q5.10 的特征图经卷积运算以及结合后处理 (确定数据格式应为 Q6.9) 后, 得到数据格式为 Q6.9 的特征图; 数据格式为 Q4.11 的特征图经卷积运算以及结合后处理后, 得到数据格式为 Q6.9 的特征图; 得到的两个数据格式为 Q6.9 的特征图作为 Concatenation 层的输入, 经 Concatenation 层运算后得到数据格式为 Q6.9 的特征图 (Concatenation 层的输出)。

应理解, 图 6 的方案仅是本申请的一个具体的实施例。在其他实施例中仍以上述例子为例, 后处理可以选取以数据格式 Q4.11 进行对齐, 即以保证小数位位宽最大为标准对齐; 或者其他实施例还可以以其他标准选取对齐的位宽, 本申请实施例对此不作限定。

图 7 是本申请一个实施例的数据定点化方法的示意性流程图。如图 7 所示, 确定数据定点化方案需要获得神经网络的结构, 各层的权重系数以及用于确定定点化方案的输入数据集。神经网络的结构是指神经网络所包括的层的类型。根据神经网络的结构, 执行 S210 的合并预处理。之后可以执行 S220, 确定各层的权重系数的定点化方案。根据输入数据集得到在各层的输出值, 执行 S230 各层输出值的定点化, 结合 S240 准确率测试的结果, 确定各层的权重系数的定点化方案。最后, 可以执行 S250 后处理。根据 S210 至 S250 的结果, 输出各层权重系数和输出值的定点化参数, 例如, 非小数部分位宽, 或者非小数部分位宽和小数部分位宽, 或者整数部分位宽和小数部分位宽, 或者符号位宽、整数部分位宽和小数部分位宽等等。

本申请实施例还提供了一种数据定点化方法, 图 8 是本申请一个实施例的数据定点化方法 300 的示意性流程图。数据定点化方法 300 可以包括:

S310, 计算输入样本在神经网络的第一目标层的参考输出值。

S320, 确定为输出值预设的输出值总位宽和预设的第一符号位宽。

S330, 根据参考输出值的大小确定输出值整数部分位宽。

S340, 根据输出值总位宽、第一符号位宽和输出值整数部分位宽, 确定输出值小数部分位宽, 其中, 第一符号位宽、输出值整数部分位宽和输出值小数部分位宽作为第一目标层在输出值定点化时使用的符号位宽、整数部分位宽和小数部分位宽。

本申请实施例的数据定点化方法在输出值定点化时考虑符号位，这样使得确定的定点化方案更优，提高增加网络准确率的可能性。

应理解，本申请实施例的参考输出值可以为一个值，也可以为多个输入样本产生的多个参考输出值。参考输出值可以为输入样本在第一目标层的最大输出值，也可以是次大输出值或者除最大输出值以外的其他值。根据准确率测试，从多个参考输出值（例如多个最大输出值）对应的定点化方案中确定出最优的定点化方案。具体过程已在前文的实施例中进行了描述，此处不再赘述。

可选地，以参考输出值为最大输出值为例，可以根据最大输出值 O 的大小确定非小数部分位宽，例如非小数部分位宽 $iwo = \text{ceil}(\log_2(O) + 1)$ ，则小数部分位宽 $fwo = bwo - iwo$ ， $\text{ceil}()$ 表示向上取整。应理解，参考非小数部分位宽中可以包括第一符号位宽（通常而言第一符号位宽为 1）和整数部分位宽 $iwo - 1$ ；参考非小数部分位宽也可以没有符号位，仅包括整数部分位宽 iwo 。

可选地，作为一个实施例，数据定点化方法 300 还可以包括：确定为权重系数预设的权重定点总位宽和预设的第二符号位宽；根据第一目标层中最大权重系数的大小确定权重整数部分位宽；根据权重定点总位宽、第二符号位宽和权重整数部分位宽，确定权重小数部分位宽，其中，将第二符号位宽、权重整数部分位宽和权重小数部分位宽确定为第一目标层在权重系数定点化时使用的符号位宽、整数部分位宽和小数部分位宽。

可选地，作为一个实施例，数据定点化方法 300 还可以包括：对神经网络的至少两层进行合并预处理，得到合并后形成的第一目标层。

可选地，作为一个实施例，参考输出值为多个输入样本中每个输入样本在合并后形成的第一目标层的参考输出值。

可选地，作为一个实施例，最大权重系数为对神经网络的至少两层进行合并预处理后形成的第一目标层的权重系数的最大值。

可选地，作为一个实施例，对神经网络的至少两层进行合并预处理，得到合并后形成的第一目标层，可以包括：对神经网络的卷积层和归一化层进行合并预处理，得到第一目标层；或，对神经网络的卷积层和缩放层进行合并预处理，得到第一目标层；或，对神经网络的卷积层、归一化层和缩放层进行合并预处理，得到第一目标层。

可选地，作为一个实施例，第一目标层可以包括卷积层、转置卷积层、

归一化层、缩放层、池化层、全连接层、拼接层、元素智能加法层和激活层中的一层或至少两层合并后的层。

可选地，作为一个实施例，数据定点化方法 300 还可以包括：确定神经网络的第二目标层在输出值定点化时使用的整数部分位宽，使第二目标层在输出值定点化时使用的整数部分位宽和第一目标层在输出值定点化时使用的整数部分位宽相等。

可选地，作为一个实施例，第二目标层在输出值定点化时使用的小数部分位宽和第一目标层在输出值定点化时使用的小数部分位宽相等。

可选地，作为一个实施例，第一目标层的输出值与第二目标层的输出值在拼接层和/或元素智能加法层进行后续处理。

上述可选的各实施例的具体流程可以参考前文的说明，此处不再赘述。

本申请实施例还提供了一种数据处理方法。图 9 是本申请实施例的数据处理方法 400 的示意性流程图。数据处理方法 400 可以包括：

S410，对神经网络的至少两层进行合并预处理。

15 S420，基于合并预处理后的神经网络进行神经网络运算。

本申请实施例的数据处理方法，对神经网络的至少两层进行合并预处理，基于合并预处理后的神经网络进行运算，可以节省计算资源，提高系统效率。

可选地，作为一个实施例，S410 对神经网络的至少两层进行合并预处理，可以包括：对神经网络的卷积层和归一化层进行合并预处理；或，对神经网络的卷积层和缩放层进行合并预处理；或，对神经网络的卷积层、归一化层和缩放层进行合并预处理。

可选地，作为一个实施例，数据处理方法 400 还可以包括：确定至少两层进行合并预处理后形成的第一目标层的权重系数。

25 可选地，作为一个实施例，S420 基于合并预处理后的神经网络进行神经网络运算，包括：对至少两层进行合并预处理后形成的第一目标层进行定点化计算。

可选地，作为一个实施例，对至少两层进行合并预处理后形成的第一目标层进行定点化计算，可以包括：根据前文描述的数据定点化方法 100 或 200 确定第一目标层定点化使用的整数部分位宽。

上述可选的各实施例的具体流程可以参考前文的说明，此处不再赘述。

本申请实施例还提供了一种数据对齐方法。图 10 是本申请实施例的数据对齐方法 500 的示意性流程图。数据对齐方法 500 可以包括：

S510，从神经网络中确定需数据对齐的多个层。

5 S520，根据多个层中每层输出值定点化应使用的整数部分位宽，确定多个层输出值定点化最终使用的整数部分位宽，其中，多个层中任意两层输出值定点化最终使用的整数部分位宽相等。

本申请实施例的数据对齐方法可以在确定定点化方案时解决一些层有输入数据小数点对齐需求的问题，可以减少硬件资源的使用，提高系统效率。

10 可选地，作为一个实施例，数据对齐方法 500 还可以包括：根据前文描述的数据定点化方法 100 或 200 确定多个层中每层输出值定点化应使用的整数部分位宽。

可选地，作为一个实施例，多个层中任意两层输出值定点化最终使用的小数部分位宽相等。

15 可选地，作为一个实施例，S520 确定多个层输出值定点化最终使用的整数部分位宽，可以包括：将多个层中定点化应使用的所有整数部分位宽中的最大值，确定为多个层输出值定点化最终使用的整数部分位宽。

20 应理解，S520 确定多个层输出值定点化最终使用的整数部分位宽，也可以包括：将多个层中定点化应使用的所有整数部分位宽中的最小值，确定为多个层输出值定点化最终使用的整数部分位宽；或者按照其他标准或预设规则确定最终使用的整数部分位宽，本申请实施例对此不作限定。

25 本申请实施例还提供了一种数据定点化方法。该数据定点化方法包括计算多个输入样本中每个输入样本在神经网络的第一目标层的最大输出值。从多个最大输出值中选取一个最大输出值作为定点参考值。根据定点参考值确定参考整数部分位宽，作为第一目标层在输出值定点化时使用的整数部分位宽。

30 应理解，从多个最大输出值中选取一个最大输出值作为定点参考值可以按照预设规则选取。例如，从多个最大输出值中选取数值最大的一个最大输出值作为定点参考值；或者从多个最大输出值中选取数值次大的一个最大输出值作为定点参考值；或者从多个最大输出值中选取数值处于中间某位置的一个最大输出值作为定点参考值；或者对多个最大输出值进行排序，按照预设的选取参数从多个最大输出值中选取一个最大输出值作为定点参

考值，等等，本申请实施例对具体的选取方法不作限定。

可选地，作为一个实施例，根据定点参考值确定参考整数部分位宽，作为第一目标层在输出值定点化时使用的整数部分位宽，包括：根据定点参考值确定参考整数部分位宽；基于预设的输出值总位宽和参考整数部分位宽进行准确率测试，在准确率不小于预设阈值时，将参考整数部分位宽作为第一目标层在输出值定点化时使用的整数部分位宽。

在一个具体的例子中，例如预设阈值为 85%，当从多个最大输出值中选取数值次大的最大输出值作为定点参考值对应的参考整数部分位宽使得准确率不小于 85% 时，则将该参考整数部分位宽作为第一目标层在输出值定点化时使用的整数部分位宽；当从多个最大输出值中选取数值次大的最大输出值作为定点参考值对应的参考整数部分位宽使得准确率小于 85% 时，则将数值最大的最大输出值作为定点参考值重新计算参考整数部分位宽，当该参考整数部分位宽使得准确率不小于 85% 时，将该参考整数部分位宽作为第一目标层在输出值定点化时使用的整数部分位宽。应理解，这仅是确定第一目标层在输出值定点化时使用的整数部分位宽的一个具体的例子，而非对本申请实施例的限定。

以上详细说明了本申请实施例的方法，下面详细说明了本申请实施例的装置。

图 11 是本申请一个实施例的数据定点化装置 600 的示意性框图。数据定点化装置 600 包括：

前向传播计算模块 610，用于计算多个输入样本中每个输入样本在神经网络的第一目标层的最大输出值。

定点参考选取模块 620，用于从前向传播计算模块 610 得到的多个最大输出值中选取至少两个最大输出值作为定点参考值。

参考位宽确定模块 630，用于根据定点参考选取模块 620 选取的每个定点参考值确定参考整数部分位宽。

准确率测试模块 640，用于基于预设的输出值总位宽和参考位宽确定模块 630 确定的每个参考整数部分位宽进行准确率测试，将使得准确率最高的参考整数部分确定为第一目标层在输出值定点化时使用的整数部分位宽。

本申请实施例的数据定点化装置 600 通过从第一目标层的多个最大输出值中选取多个值作为定点参考值，确定每个定点参考值对应的参考整数部分

位宽，并基于准确率测试确定最优的整数部分位宽，根据最优的整数部分位宽使得定点化后网络可以在传递更多的有用信息同时保持较高的精度，提高网络的表达能力和准确率。

5 可选地，作为一个实施例，定点参考选取模块 620 从多个最大输出值中选取至少两个最大输出值作为定点参考值，可以包括：定点参考选取模块 620 对多个最大输出值进行排序，按照预设的选取参数从多个最大输出值中选取至少两个最大输出值作为定点参考值。

10 可选地，作为一个实施例，参考位宽确定模块 630 根据每个定点参考值确定参考整数部分位宽，包括：参考位宽确定模块 630 根据定点参考值的大小确定参考整数部分位宽；参考位宽确定模块 630 还用于：确定预设的第一符号位宽和输出值总位宽；根据第一符号位宽、输出值总位宽和参考整数部分位宽，确定参考小数部分位宽。

15 可选地，作为一个实施例，数据定点化装置 600 还可以包括权重位宽确定模块，用于：确定为权重系数预设的权重定点总位宽和预设的第二符号位宽；根据第一目标层中最大权重系数的大小确定权重整数部分位宽；根据权重定点总位宽、第二符号位宽和权重整数部分位宽，确定权重小数部分位宽，其中，将第二符号位宽、权重整数部分位宽和权重小数部分位宽确定为第一目标层在权重系数定点化时使用的符号位宽、整数部分位宽和小数部分位宽。

20 可选地，作为一个实施例，数据定点化装置 600 还可以包括预处理模块，用于：对神经网络的至少两层进行合并预处理，得到合并后形成的第一目标层。

可选地，作为一个实施例，最大输出值为多个输入样本中每个输入样本在合并后形成的第一目标层的最大输出值。

25 可选地，作为一个实施例，最大权重系数为对神经网络的至少两层进行合并预处理后形成的第一目标层的权重系数的最大值。

30 可选地，作为一个实施例，预处理模块对神经网络的至少两层进行合并预处理，得到合并后形成的第一目标层，包括：预处理模块对神经网络的卷积层和归一化层进行合并预处理，得到第一目标层；或，预处理模块对神经网络的卷积层和缩放层进行合并预处理，得到第一目标层；或，预处理模块对神经网络的卷积层、归一化层和缩放层进行合并预处理，得到第一目标层。

可选地，作为一个实施例，第一目标层为卷积层、转置卷积层、归一化层、缩放层、池化层、全连接层、拼接层、元素智能加法层和激活层中的一层或至少两层合并后的层。

可选地，作为一个实施例，数据定点化装置 600 还包括对齐模块，用于：
5 确定神经网络的第二目标层在输出值定点化时使用的整数部分位宽，使第二目标层在输出值定点化时使用的整数部分位宽和第一目标层在输出值定点化时使用的整数部分位宽相等。

可选地，作为一个实施例，第二目标层在输出值定点化时使用的小数部分位宽和第一目标层在输出值定点化时使用的小数部分位宽相等。

10 可选地，作为一个实施例，对齐模块确定神经网络的第二目标层在输出值定点化时使用的整数部分位宽，包括：对齐模块将第一目标层和第二目标层输出值定点化应使用的整数部分位宽中的最大值，确定为第一目标层和第二目标层输出值定点化最终使用的整数部分位宽。

可选地，作为一个实施例，第一目标层的输出值与第二目标层的输出值
15 在拼接层和/或元素智能加法层进行后续处理。

图 12 是本申请一个实施例的数据定点化装置 700 的示意性框图。数据定点化装置 700 包括：

前向传播计算模块 710，用于计算输入样本在神经网络的第一目标层的参考输出值。

20 确定模块 720，用于确定为输出值预设的输出值总位宽和预设的第一符号位宽。

输出值位宽确定模块 730，用于根据前向传播计算模块 710 得到的参考输出值的大小确定输出值整数部分位宽；根据确定模块 720 确定的输出值总位宽、第一符号位宽和输出值整数部分位宽，确定输出值小数部分位宽，其中，第一符号位宽、输出值整数部分位宽和输出值小数部分位宽作为第一目标层在输出值定点化时使用的符号位宽、整数部分位宽和小数部分位宽。
25

本申请实施例的数据定点化装置在输出值定点化时考虑符号位，这样使得确定的定点化方案更优，提高增加网络准确率的可能性。

可选地，作为一个实施例，参考输出值可以为输入样本在第一目标层的最大输出值。
30

可选地，作为一个实施例，数据定点化装置 700 还可以包括权重位宽确

定模块，用于：确定为权重系数预设的权重定点总位宽和预设的第二符号位宽；根据第一目标层中最大权重系数的大小确定权重整数部分位宽；根据权重定点总位宽、第二符号位宽和权重整数部分位宽，确定权重小数部分位宽，其中，将第二符号位宽、权重整数部分位宽和权重小数部分位宽确定
5 为第一目标层在权重系数定点化时使用的符号位宽、整数部分位宽和小数部分位宽。

可选地，作为一个实施例，数据定点化装置 700 还可以包括预处理模块，用于：对神经网络的至少两层进行合并预处理，得到合并后形成的第一目标层。

10 可选地，作为一个实施例，参考输出值为多个输入样本中每个输入样本在合并后形成的第一目标层的参考输出值。

可选地，作为一个实施例，最大权重系数为对神经网络的至少两层进行合并预处理后形成的第一目标层的权重系数的最大值。

15 可选地，作为一个实施例，预处理模块对对神经网络的至少两层进行合并预处理，得到合并后形成的第一目标层，包括：预处理模块对对神经网络的卷积层和归一化层进行合并预处理，得到第一目标层；或，预处理模块对对神经网络的卷积层和缩放层进行合并预处理，得到第一目标层；或，预处理模块对对神经网络的卷积层、归一化层和缩放层进行合并预处理，得到第一目标层。

20 可选地，作为一个实施例，第一目标层为卷积层、转置卷积层、归一化层、缩放层、池化层、全连接层、拼接层、元素智能加法层和激活层中的一层或至少两层合并后的层。

25 可选地，作为一个实施例，数据定点化装置 700 还可以包括对齐模块，用于：对齐模块确定神经网络的第二目标层在输出值定点化时使用的整数部分位宽，使第二目标层在输出值定点化时使用的整数部分位宽和第一目标层在输出值定点化时使用的整数部分位宽相等。

可选地，作为一个实施例，第二目标层在输出值定点化时使用的小数部分位宽和第一目标层在输出值定点化时使用的小数部分位宽相等。

30 可选地，作为一个实施例，第一目标层的输出值与第二目标层的输出值在拼接层和/或元素智能加法层进行后续处理。

图 13 是本申请一个实施例的数据处理装置 800 的示意性框图。该数据

处理装置 800 包括:

预处理模块 810, 用于对神经网络的至少两层进行合并预处理;

运算模块 820, 用于基于预处理模块 810 合并预处理后的神经网络进行神经网络运算。

- 5 本申请实施例的数据处理装置, 对神经网络的至少两层进行合并预处理, 基于合并预处理后的神经网络进行运算, 可以节省计算资源, 提高系统效率。

10 可选地, 作为一个实施例, 预处理模块 810 对神经网络的至少两层进行合并预处理, 可以包括: 预处理模块对神经网络的卷积层和归一化层进行合并预处理; 或, 预处理模块对神经网络的卷积层和缩放层进行合并预处理; 或, 预处理模块对神经网络的卷积层、归一化层和缩放层进行合并预处理。

可选地, 作为一个实施例, 数据处理装置 800 还可以包括确定模块, 用于: 确定至少两层进行合并预处理后形成的第一目标层的权重系数。

15 可选地, 作为一个实施例, 运算模块 820 基于合并预处理后的神经网络进行神经网络运算, 可以包括: 运算模块 820 对至少两层进行合并预处理后形成的第一目标层进行定点化计算。

可选地, 作为一个实施例, 运算模块 820 对至少两层进行合并预处理后形成的第一目标层进行定点化计算, 可以包括: 运算模块 820 根据前文描述的数据定点化方法 100 或 200 确定第一目标层定点化使用的整数部分位宽。

20 图 14 是本申请一个实施例的数据对齐装置 900 的示意性框图。数据对齐装置 900 包括:

第一确定模块 910, 用于从神经网络中确定需数据对齐的多个层。

25 第二确定模块 920, 用于根据多个层中每层输出值定点化应使用的整数部分位宽, 确定多个层输出值定点化最终使用的整数部分位宽, 其中, 多个层中任意两层输出值定点化最终使用的整数部分位宽相等。

本申请实施例的数据对齐装置可以在确定定点化方案时解决一些层有输入数据对齐需求的问题, 可以减少硬件资源的使用, 提高系统效率。

30 可选地, 作为一个实施例, 数据对齐装置 900 还可以包括第三确定模块, 用于: 根据前文描述的数据定点化方法 100 或 200 确定多个层中每层输出值定点化应使用的整数部分位宽。

可选地, 作为一个实施例, 多个层中任意两层输出值定点化最终使用的

小数部分位宽相等。

可选地，作为一个实施例，第二确定模块 920 确定多个层输出值定点化最终使用的整数部分位宽，包括：第二确定模块将多个层中定点化应使用的所有整数部分位宽中的最大值，确定为多个层输出值定点化最终使用的整数部分位宽。

本申请实施例还提供了一种数据定点化装置。该数据定点化装置包括：前向传播计算模块，用于计算多个输入样本中每个输入样本在神经网络的第一目标层的最大输出值；定点参考选取模块，用于从前向传播计算模块得到的多个最大输出值中选取一个最大输出值作为定点参考值；和位宽确定模块，用于根据定点参考选取模块选取的定点参考值确定参考整数部分位宽，作为第一目标层在输出值定点化时使用的整数部分位宽。

可选地，作为一个实施例，位宽确定模块根据定点参考值确定参考整数部分位宽，作为第一目标层在输出值定点化时使用的整数部分位宽，可以包括：位宽确定模块根据定点参考值确定参考整数部分位宽；位宽确定模块基于预设的输出值总位宽和参考整数部分位宽进行准确率测试，在准确率不小于预设阈值时，将参考整数部分位宽作为第一目标层在输出值定点化时使用的整数部分位宽。

应理解，本申请各实施例的装置可以基于存储器和处理器实现，各存储器用于存储用于执行本申请个实施例的方法的指令，处理器执行上述指令，使得装置执行本申请各实施例的方法。

应理解，本申请实施例中提及的处理器可以是中央处理单元（Central Processing Unit, CPU），还可以是其他通用处理器、数字信号处理器（Digital Signal Processor, DSP）、专用集成电路（Application Specific Integrated Circuit, ASIC）、现成可编程门阵列（Field Programmable Gate Array, FPGA）或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。

还应理解，本申请实施例中提及的存储器可以是易失性存储器或非易失性存储器，或可包括易失性和非易失性存储器两者。其中，非易失性存储器可以是只读存储器（Read-Only Memory, ROM）、可编程只读存储器（Programmable ROM, PROM）、可擦除可编程只读存储器（Erasable PROM, EPROM）、电可擦除可编程只读存储器（Electrically EPROM, EEPROM）

或闪存。易失性存储器可以是随机存取存储器 (Random Access Memory, RAM), 其用作外部高速缓存。通过示例性但不是限制性说明, 许多形式的 RAM 可用, 例如静态随机存取存储器 (Static RAM, SRAM)、动态随机存取存储器 (Dynamic RAM, DRAM)、同步动态随机存取存储器 (Synchronous DRAM, SDRAM)、双倍数据速率同步动态随机存取存储器 (Double Data Rate SDRAM, DDR SDRAM)、增强型同步动态随机存取存储器 (Enhanced SDRAM, ESDRAM)、同步连接动态随机存取存储器 (Synchlink DRAM, SLDRAM) 和直接内存总线随机存取存储器 (Direct Rambus RAM, DR RAM)。

需要说明的是, 当处理器为通用处理器、DSP、ASIC、FPGA 或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件时, 存储器 (存储模块) 集成在处理器中。

应注意, 本文描述的存储器旨在包括但不限于这些和任意其它适合类型的存储器。

本申请实施例还提供一种计算机可读存储介质, 其上存储有指令, 当指令在计算机上运行时, 使得计算机执行上述各方法实施例的方法。

本申请实施例还提供一种计算设备, 该计算设备包括上述计算机可读存储介质。

本申请实施例可以应用在飞行器, 尤其是无人机领域。

应理解, 本申请各实施例的电路、子电路、子单元的划分只是示意性的。本领域普通技术人员可以意识到, 本文中所公开的实施例描述的各示例的电路、子电路和子单元, 能够再行拆分或组合。

在上述实施例中, 可以全部或部分地通过软件、硬件、固件或者其任意组合来实现。当使用软件实现时, 可以全部或部分地以计算机程序产品的形式实现。计算机程序产品包括一个或多个计算机指令。在计算机上加载和执行计算机指令时, 全部或部分地产生按照本申请实施例的流程或功能。计算机可以是通用计算机、专用计算机、计算机网络、或者其他可编程装置。计算机指令可以存储在计算机可读存储介质中, 或者从一个计算机可读存储介质向另一个计算机可读存储介质传输, 例如, 计算机指令可以从一个网站站点、计算机、服务器或数据中心通过有线 (例如同轴电缆、光纤、数字用户线 (Digital Subscriber Line, DSL)) 或无线 (例如红外、无

线、微波等)方式向另一个网站站点、计算机、服务器或数据中心进行传输。计算机可读存储介质可以是计算机能够存取的任何可用介质或者是包含一个或多个可用介质集成的服务器、数据中心等数据存储设备。可用介质可以是磁性介质(例如,软盘、硬盘、磁带)、光介质(例如,高密度数字视频光盘(Digital Video Disc, DVD))、或者半导体介质(例如,固态硬盘(Solid State Disk, SSD))等。

应理解,本申请各实施例均是以总位宽为16位(bit)为例进行说明的,本申请各实施例可以适用于其他的位宽。

应理解,说明书通篇中提到的“一个实施例”或“一实施例”意味着与实施例有关的特定特征、结构或特性包括在本申请的至少一个实施例中。因此,在整个说明书各处出现的“在一个实施例中”或“在一实施例中”未必一定指相同的实施例。此外,这些特定的特征、结构或特性可以任意适合的方式结合在一个或多个实施例中。

应理解,在本申请的各种实施例中,上述各过程的序号的大小并不意味着执行顺序的先后,各过程的执行顺序应以其功能和内在逻辑确定,而不应对本申请实施例的实施过程构成任何限定。

应理解,在本申请实施例中,“与A相应的B”表示B与A相关联,根据A可以确定B。但还应理解,根据A确定B并不意味着仅仅根据A确定B,还可以根据A和/或其它信息确定B。

应理解,本文中术语“和/或”,仅仅是一种描述关联对象的关联关系,表示可以存在三种关系,例如,A和/或B,可以表示:单独存在A,同时存在A和B,单独存在B这三种情况。另外,本文中字符“/”,一般表示前后关联对象是一种“或”的关系。

本领域普通技术人员可以意识到,结合本文中所公开的实施例描述的各示例的单元及算法步骤,能够以电子硬件、或者计算机软件和电子硬件的结合来实现。这些功能究竟以硬件还是软件方式来执行,取决于技术方案的特定应用和设计约束条件。专业技术人员可以对每个特定的应用来使用不同方法来实现所描述的功能,但是这种实现不应认为超出本申请的范围。

所属领域的技术人员可以清楚地了解到,为描述的方便和简洁,上述描述的系统、装置和单元的具体工作过程,可以参考前述方法实施例中的

对应过程，在此不再赘述。

在本申请所提供的几个实施例中，应该理解到，所揭露的系统、装置和方法，可以通过其它的方式实现。例如，以上所描述的装置实施例仅仅是示意性的，例如，所述单元的划分，仅仅为一种逻辑功能划分，实际实现时可以有另外的划分方式，例如多个单元或组件可以结合或者可以集成到另一个系统，或一些特征可以忽略，或不执行。另一点，所显示或讨论的相互之间的耦合或直接耦合或通信连接可以是通过一些接口，装置或单元的间接耦合或通信连接，可以是电性，机械或其它的形式。

所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部单元来实现本实施例方案的目的。

另外，在本申请各个实施例中的各功能单元可以集成在一个处理单元中，也可以是各个单元单独物理存在，也可以两个或两个以上单元集成在一个单元中。

以上所述，仅为本申请的具体实施方式，但本申请的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本申请揭露的技术范围内，可轻易想到变化或替换，都应涵盖在本申请的保护范围之内。因此，本申请的保护范围应以所述权利要求的保护范围为准。

权利要求

1. 一种数据定点化方法，其特征在于，包括：

计算多个输入样本中每个所述输入样本在神经网络的第一目标层的最大输出值；

从多个所述最大输出值中选取至少两个所述最大输出值作为定点参考值；

根据每个所述定点参考值确定参考整数部分位宽；

基于预设的输出值总位宽和每个所述参考整数部分位宽进行准确率测试，将使得准确率最高的参考整数部分确定为所述第一目标层在输出值定点化时使用的整数部分位宽。

2. 根据权利要求 1 所述的数据定点化方法，其特征在于，所述从多个所述最大输出值中选取至少两个所述最大输出值作为定点参考值，包括：

对多个所述最大输出值进行排序，按照预设的选取参数从所述多个最大输出值中选取至少两个所述最大输出值作为所述定点参考值。

3. 根据权利要求 1 或 2 所述的数据定点化方法，其特征在于，所述根据每个所述定点参考值确定参考整数部分位宽，包括：

根据所述定点参考值的大小确定所述参考整数部分位宽；

所述方法还包括：

确定预设的第一符号位宽和所述输出值总位宽；

根据所述第一符号位宽、所述输出值总位宽和所述参考整数部分位宽，确定参考小数部分位宽。

4. 根据权利要求 1 或 2 所述的数据定点化方法，其特征在于，所述数据定点化方法还包括：

确定为权重系数预设的权重定点总位宽和预设的第二符号位宽；

根据所述第一目标层中最大权重系数的大小确定权重整数部分位宽；

根据所述权重定点总位宽、所述第二符号位宽和所述权重整数部分位宽，确定权重小数部分位宽，

其中，将所述第二符号位宽、所述权重整数部分位宽和所述权重小数部分位宽确定为所述第一目标层在权重系数定点化时使用的符号位宽、整数部分位宽和小数部分位宽。

5. 根据权利要求 4 所述的数据定点化方法, 其特征在于, 所述最大权重系数为对所述神经网络的至少两层进行合并预处理后形成的所述第一目标层的权重系数的最大值。

6. 根据权利要求 1 所述的数据定点化方法, 其特征在于, 所述数据定点化方法还包括:

对所述神经网络的至少两层进行合并预处理, 得到合并后形成的所述第一目标层。

7. 根据权利要求 6 所述的数据定点化方法, 其特征在于, 所述最大输出值为所述多个输入样本中每个所述输入样本在合并后形成的所述第一目标层的所述最大输出值。

8. 根据权利要求 6 所述的数据定点化方法, 其特征在于, 所述对所述神经网络的至少两层进行合并预处理, 得到合并后形成的所述第一目标层, 包括:

对所述神经网络的卷积层和归一化层进行合并预处理, 得到所述第一目标层; 或,

对所述神经网络的卷积层和缩放层进行合并预处理, 得到所述第一目标层; 或,

对所述神经网络的卷积层、归一化层和缩放层进行合并预处理, 得到所述第一目标层。

9. 根据权利要求 1 所述的数据定点化方法, 其特征在于, 所述第一目标层包括卷积层、转置卷积层、归一化层、缩放层、池化层、全连接层、拼接层、元素智能加法层和激活层中的一层或至少两层合并后的层。

10. 根据权利要求 1 所述的数据定点化方法, 其特征在于, 所述数据定点化方法还包括:

确定所述神经网络的第二目标层在输出值定点化时使用的整数部分位宽, 使所述第二目标层在输出值定点化时使用的整数部分位宽和所述第一目标层在输出值定点化时使用的整数部分位宽相等。

11. 根据权利要求 10 所述的数据定点化方法, 其特征在于, 所述确定所述神经网络的第二目标层在输出值定点化时使用的整数部分位宽, 包括:

将所述第一目标层和所述第二目标层输出值定点化应使用的整数部分位宽中的最大值, 确定为所述第一目标层和所述第二目标层输出值定点化最

终使用的整数部分位宽。

12. 根据权利要求 10 或 11 所述的数据定点化方法, 其特征在于, 所述第一目标层的输出值与所述第二目标层的输出值在拼接层和/或元素智能加法层进行后续处理。

- 5 13. 一种数据定点化方法, 其特征在于, 包括:
 计算输入样本在神经网络的第一目标层的参考输出值;
 确定为输出值预设的输出值总位宽和预设的第一符号位宽;
 根据所述参考输出值的大小确定输出值整数部分位宽;
 根据所述输出值总位宽、所述第一符号位宽和所述输出值整数部分位
10 宽, 确定输出值小数部分位宽, 其中, 所述第一符号位宽、所述输出值整数
 部分位宽和所述输出值小数部分位宽作为所述第一目标层在输出值定点化
 时使用的符号位宽、整数部分位宽和小数部分位宽。

14. 根据权利要求 13 所述的数据定点化方法, 其特征在于, 所述参考输出值为所述输入样本在所述第一目标层的最大输出值。

- 15 15. 根据权利要求 13 或 14 所述的数据定点化方法, 其特征在于, 所述
 数据定点化方法还包括:
 确定为权重系数预设的权重定点总位宽和预设的第二符号位宽;
 根据所述第一目标层中最大权重系数的大小确定权重整数部分位宽;
 根据所述权重定点总位宽、所述第二符号位宽和所述权重整数部分位
20 宽, 确定权重小数部分位宽,
 其中, 将所述第二符号位宽、所述权重整数部分位宽和所述权重小数部
 分位宽确定为所述第一目标层在权重系数定点化时使用的符号位宽、整数部
 分位宽和小数部分位宽。

- 25 16. 根据权利要求 15 所述的数据定点化方法, 其特征在于, 所述最大
 权重系数为对所述神经网络的至少两层进行合并预处理后形成的所述第一
 目标层的权重系数的最大值。

17. 根据权利要求 13 所述的数据定点化方法, 其特征在于, 所述数据定点化方法还包括:

- 30 对所述神经网络的至少两层进行合并预处理, 得到合并后形成的所述第
 一目标层。

18. 根据权利要求 17 所述的数据定点化方法, 其特征在于, 所述参考

输出值为所述多个输入样本中每个所述输入样本在合并后形成的所述第一目标层的所述参考输出值。

19. 根据权利要求 17 所述的数据定点化方法，其特征在于，所述对所述神经网络的至少两层进行合并预处理，得到合并后形成的所述第一目标层，包括：

对所述神经网络的卷积层和归一化层进行合并预处理，得到所述第一目标层；或，

对所述神经网络的卷积层和缩放层进行合并预处理，得到所述第一目标层；或，

对所述神经网络的卷积层、归一化层和缩放层进行合并预处理，得到所述第一目标层。

20. 根据权利要求 13 所述的数据定点化方法，其特征在于，所述第一目标层包括卷积层、转置卷积层、归一化层、缩放层、池化层、全连接层、拼接层、元素智能加法层和激活层中的一层或至少两层合并后的层。

21. 根据权利要求 13 所述的数据定点化方法，其特征在于，所述数据定点化方法还包括：

确定所述神经网络的第二目标层在输出值定点化时使用的整数部分位宽，使所述第二目标层在输出值定点化时使用的整数部分位宽和所述第一目标层在输出值定点化时使用的整数部分位宽相等。

22. 根据权利要求 21 所述的数据定点化方法，其特征在于，所述第一目标层的输出值与所述第二目标层的输出值在拼接层和/或元素智能加法层进行后续处理。

23. 一种数据处理方法，其特征在于，包括：

对神经网络的至少两层进行合并预处理；

基于合并预处理后的神经网络进行神经网络运算。

24. 根据权利要求 23 所述的数据处理方法，其特征在于，所述对神经网络的至少两层进行合并预处理，包括：

对所述神经网络的卷积层和归一化层进行合并预处理；或，

对所述神经网络的卷积层和缩放层进行合并预处理；或，

对所述神经网络的卷积层、归一化层和缩放层进行合并预处理。

25. 根据权利要求 23 或 24 所述的数据处理方法，其特征在于，所述数

据处理方法还包括:

确定所述至少两层进行合并预处理后形成的第一目标层的权重系数。

26. 根据权利要求 23 或 24 所述的数据处理方法, 其特征在于, 所述基于合并预处理后的神经网络进行神经网络运算, 包括:

5 对所述至少两层进行合并预处理后形成的第一目标层进行定点化计算。

27. 根据权利要求 26 所述的数据处理方法, 其特征在于, 所述对所述至少两层进行合并预处理后形成的第一目标层进行定点化计算, 包括:

根据权利要求 1 至 12 中任一项所述的数据定点化方法确定所述第一目标层定点化使用的整数部分位宽。

10 28. 一种数据对齐方法, 其特征在于, 包括:

从神经网络中确定需数据对齐的多个层;

根据所述多个层中每层输出值定点化应使用的整数部分位宽, 确定所述多个层输出值定点化最终使用的整数部分位宽, 其中, 所述多个层中任意两层输出值定点化最终使用的整数部分位宽相等。

15 29. 根据权利要求 28 所述的数据对齐方法, 其特征在于, 所述数据对齐方法还包括:

根据权利要求 1 至 12 中任一项所述的数据定点化方法确定所述多个层中每层输出值定点化应使用的整数部分位宽。

20 30. 根据权利要求 28 或 29 所述的数据对齐方法, 其特征在于, 所述确定所述多个层输出值定点化最终使用的整数部分位宽, 包括:

将所述多个层中定点化应使用的所有整数部分位宽中的最大值, 确定为所述多个层输出值定点化最终使用的整数部分位宽。

31. 一种数据定点化方法, 其特征在于, 包括:

25 计算多个输入样本中每个所述输入样本在神经网络的第一目标层的最大输出值;

从多个所述最大输出值中选取一个所述最大输出值作为定点参考值;

根据所述定点参考值确定参考整数部分位宽, 作为所述第一目标层在输出值定点化时使用的整数部分位宽。

30 32. 根据权利要求 31 所述的数据定点化方法, 其特征在于, 所述根据所述定点参考值确定参考整数部分位宽, 作为所述第一目标层在输出值定点化时使用的整数部分位宽, 包括:

根据所述定点参考值确定所述参考整数部分位宽;

基于预设的输出值总位宽和所述参考整数部分位宽进行准确率测试,在准确率不小于预设阈值时,将所述参考整数部分位宽作为所述第一目标层在输出值定点化时使用的整数部分位宽。

5 33. 一种数据定点化装置, 其特征在于, 包括:

前向传播计算模块, 用于计算多个输入样本中每个所述输入样本在神经网络的第一目标层的最大输出值;

定点参考选取模块, 用于从所述前向传播计算模块得到的多个所述最大输出值中选取至少两个所述最大输出值作为定点参考值;

10 参考位宽确定模块, 用于根据所述定点参考选取模块选取的每个所述定点参考值确定参考整数部分位宽;

准确率测试模块, 用于基于预设的输出值总位宽和所述参考位宽确定模块确定的每个所述参考整数部分位宽进行准确率测试, 将使得准确率最高的参考整数部分确定为所述第一目标层在输出值定点化时使用的整数部分位宽。

15 34. 根据权利要求 33 所述的数据定点化装置, 其特征在于, 所述定点参考选取模块从多个所述最大输出值中选取至少两个所述最大输出值作为定点参考值, 包括:

所述定点参考选取模块对多个所述最大输出值进行排序, 按照预设的选取参数从所述多个最大输出值中选取至少两个所述最大输出值作为所述定点参考值。

35. 根据权利要求 33 或 34 所述的数据定点化装置, 其特征在于, 所述参考位宽确定模块根据每个所述定点参考值确定参考整数部分位宽, 包括:

25 所述参考位宽确定模块根据所述定点参考值的大小确定所述参考整数部分位宽;

所述参考位宽确定模块还用于:

确定预设的第一符号位宽和所述输出值总位宽;

根据所述第一符号位宽、所述输出值总位宽和所述参考整数部分位宽, 确定参考小数部分位宽。

30 36. 根据权利要求 33 或 34 所述的数据定点化装置, 其特征在于, 所述数据定点化装置还包括权重位宽确定模块, 用于:

确定为权重系数预设的权重定点总位宽和预设的第二符号位宽；

根据所述第一目标层中最大权重系数的大小确定权重整数部分位宽；

根据所述权重定点总位宽、所述第二符号位宽和所述权重整数部分位宽，确定权重小数部分位宽，

- 5 其中，将所述第二符号位宽、所述权重整数部分位宽和所述权重小数部分位宽确定为所述第一目标层在权重系数定点化时使用的符号位宽、整数部分位宽和小数部分位宽。

37. 根据权利要求 36 所述的数据定点化装置，其特征在于，所述最大权重系数为对所述神经网络的至少两层进行合并预处理后形成的所述第一目标层的权重系数的最大值。

38. 根据权利要求 33 所述的数据定点化装置，其特征在于，所述数据定点化装置还包括预处理模块，用于：

对所述神经网络的至少两层进行合并预处理，得到合并后形成的所述第一目标层。

- 15 39. 根据权利要求 38 所述的数据定点化装置，其特征在于，所述最大输出值为所述多个输入样本中每个所述输入样本在合并后形成的所述第一目标层的所述最大输出值。

40. 根据权利要求 38 所述的数据定点化装置，其特征在于，所述预处理模块对所述神经网络的至少两层进行合并预处理，得到合并后形成的所述第一目标层，包括：

所述预处理模块对所述神经网络的卷积层和归一化层进行合并预处理，得到所述第一目标层；或，

所述预处理模块对所述神经网络的卷积层和缩放层进行合并预处理，得到所述第一目标层；或，

- 25 所述预处理模块对所述神经网络的卷积层、归一化层和缩放层进行合并预处理，得到所述第一目标层。

41. 根据权利要求 33 所述的数据定点化装置，其特征在于，所述第一目标层包括卷积层、转置卷积层、归一化层、缩放层、池化层、全连接层、拼接层、元素智能加法层和激活层中的一层或至少两层合并后的层。

- 30 42. 根据权利要求 33 所述的数据定点化装置，其特征在于，所述数据定点化装置还包括对齐模块，用于：

确定所述神经网络的第二目标层在输出值定点化时使用的整数部分位宽，使所述第二目标层在输出值定点化时使用的整数部分位宽和所述第一目标层在输出值定点化时使用的整数部分位宽相等。

43. 根据权利要求 42 所述的数据定点化装置，其特征在于，所述对齐模块确定所述神经网络的第二目标层在输出值定点化时使用的整数部分位宽，包括：

所述对齐模块将所述第一目标层和所述第二目标层输出值定点化应使用的整数部分位宽中的最大值，确定为所述第一目标层和所述第二目标层输出值定点化最终使用的整数部分位宽。

44. 根据权利要求 42 或 43 所述的数据定点化装置，其特征在于，所述第一目标层的输出值与所述第二目标层的输出值在拼接层和/或元素智能加法层进行后续处理。

45. 一种数据定点化装置，其特征在于，包括：

前向传播计算模块，用于计算输入样本在神经网络的第一目标层的参考输出值；

确定模块，用于确定为输出值预设的输出值总位宽和预设的第一符号位宽；

输出值位宽确定模块，用于根据所述前向传播计算模块得到的所述参考输出值的大小确定输出值整数部分位宽；根据所述确定模块确定的所述输出值总位宽、所述第一符号位宽和所述输出值整数部分位宽，确定输出值小数部分位宽，其中，所述第一符号位宽、所述输出值整数部分位宽和所述输出值小数部分位宽作为所述第一目标层在输出值定点化时使用的符号位宽、整数部分位宽和小数部分位宽。

46. 根据权利要求 45 所述的数据定点化装置，其特征在于，所述参考输出值为所述输入样本在所述第一目标层的最大输出值。

47. 根据权利要求 45 或 46 所述的数据定点化装置，其特征在于，所述数据定点化装置还包括权重位宽确定模块，用于：

确定为权重系数预设的权重定点总位宽和预设的第二符号位宽；

根据所述第一目标层中最大权重系数的大小确定权重整数部分位宽；

根据所述权重定点总位宽、所述第二符号位宽和所述权重整数部分位宽，确定权重小数部分位宽，

其中，将所述第二符号位宽、所述权重整数部分位宽和所述权重小数部分位宽确定为所述第一目标层在权重系数定点化时使用的符号位宽、整数部分位宽和小数部分位宽。

48. 根据权利要求 47 所述的数据定点化装置，其特征在于，所述最大权重系数为对所述神经网络的至少两层进行合并预处理后形成的所述第一目标层的权重系数的最大值。

49. 根据权利要求 45 所述的数据定点化装置，其特征在于，所述数据定点化装置还包括预处理模块，用于：

对所述神经网络的至少两层进行合并预处理，得到合并后形成的所述第一目标层。

50. 根据权利要求 49 所述的数据定点化装置，其特征在于，所述参考输出值为所述多个输入样本中每个所述输入样本在合并后形成的所述第一目标层的所述参考输出值。

51. 根据权利要求 49 所述的数据定点化装置，其特征在于，所述预处理模块对所述神经网络的至少两层进行合并预处理，得到合并后形成的所述第一目标层，包括：

所述预处理模块对所述神经网络的卷积层和归一化层进行合并预处理，得到所述第一目标层；或，

所述预处理模块对所述神经网络的卷积层和缩放层进行合并预处理，得到所述第一目标层；或，

所述预处理模块对所述神经网络的卷积层、归一化层和缩放层进行合并预处理，得到所述第一目标层。

52. 根据权利要求 45 所述的数据定点化装置，其特征在于，所述第一目标层包括卷积层、转置卷积层、归一化层、缩放层、池化层、全连接层、拼接层、元素智能加法层和激活层中的一层或至少两层合并后的层。

53. 根据权利要求 45 所述的数据定点化装置，其特征在于，所述数据定点化装置还包括对齐模块，用于：

确定所述神经网络的第二目标层在输出值定点化时使用的整数部分位宽，使所述第二目标层在输出值定点化时使用的整数部分位宽和所述第一目标层在输出值定点化时使用的整数部分位宽相等。

54. 根据权利要求 53 所述的数据定点化装置，其特征在于，所述第一目

标层的输出值与所述第二目标层的输出值在拼接层和/或元素智能加法层进行后续处理。

55. 一种数据处理装置，其特征在于，包括：

预处理模块，用于对神经网络的至少两层进行合并预处理；

5 运算模块，用于基于所述预处理模块合并预处理后的神经网络进行神经网络运算。

56. 根据权利要求 55 所述的数据处理装置，其特征在于，所述预处理模块对神经网络的至少两层进行合并预处理，包括：

所述预处理模块对所述神经网络的卷积层和归一化层进行合并预处理；

10 或，

所述预处理模块对所述神经网络的卷积层和缩放层进行合并预处理；

或，

所述预处理模块对所述神经网络的卷积层、归一化层和缩放层进行合并预处理。

15 57. 根据权利要求 55 或 56 所述的数据处理装置，其特征在于，所述数据处理装置还包括确定模块，用于：

确定所述至少两层进行合并预处理后形成的第一目标层的权重系数。

58. 根据权利要求 55 或 56 所述的数据处理装置，其特征在于，所述运算模块基于合并预处理后的神经网络进行神经网络运算，包括：

20 所述运算模块对所述至少两层进行合并预处理后形成的第一目标层进行定点化计算。

59. 根据权利要求 58 所述的数据处理装置，其特征在于，所述运算模块对所述至少两层进行合并预处理后形成的第一目标层进行定点化计算，包括：

25 所述运算模块根据权利要求 1 至 12 中任一项所述的数据定点化方法确定所述第一目标层定点化使用的整数部分位宽。

60. 一种数据对齐装置，其特征在于，包括：

第一确定模块，用于从神经网络中确定需数据对齐的多个层；

30 第二确定模块，用于根据所述多个层中每层输出值定点化应使用的整数部分位宽，确定所述多个层输出值定点化最终使用的整数部分位宽，其中，所述多个层中任意两层输出值定点化最终使用的整数部分位宽相等。

61. 根据权利要求 60 所述的数据对齐装置, 其特征在于, 所述数据对齐装置还包括第三确定模块, 用于:

根据权利要求 1 至 12 中任一项所述的数据定点化方法确定所述多个层中每层输出值定点化应使用的整数部分位宽。

5 62. 根据权利要求 60 或 61 所述的数据对齐装置, 其特征在于, 所述第二确定模块确定所述多个层输出值定点化最终使用的整数部分位宽, 包括:

所述第二确定模块将所述多个层中定点化应使用的所有整数部分位宽中的最大值, 确定为所述多个层输出值定点化最终使用的整数部分位宽。

63. 一种数据定点化装置, 其特征在于, 包括:

10 前向传播计算模块, 用于计算多个输入样本中每个所述输入样本在神经网络的第一目标层的最大输出值;

定点参考选取模块, 用于从所述前向传播计算模块得到的多个所述最大输出值中选取一个所述最大输出值作为定点参考值;

15 位宽确定模块, 用于根据所述定点参考选取模块选取的所述定点参考值确定参考整数部分位宽, 作为所述第一目标层在输出值定点化时使用的整数部分位宽。

64. 根据权利要求 63 所述的数据定点化装置, 其特征在于, 所述位宽确定模块根据所述定点参考值确定参考整数部分位宽, 作为所述第一目标层在输出值定点化时使用的整数部分位宽, 包括:

20 所述位宽确定模块根据所述定点参考值确定所述参考整数部分位宽;

所述位宽确定模块基于预设的输出值总位宽和所述参考整数部分位宽进行准确率测试, 在准确率不小于预设阈值时, 将所述参考整数部分位宽作为所述第一目标层在输出值定点化时使用的整数部分位宽。

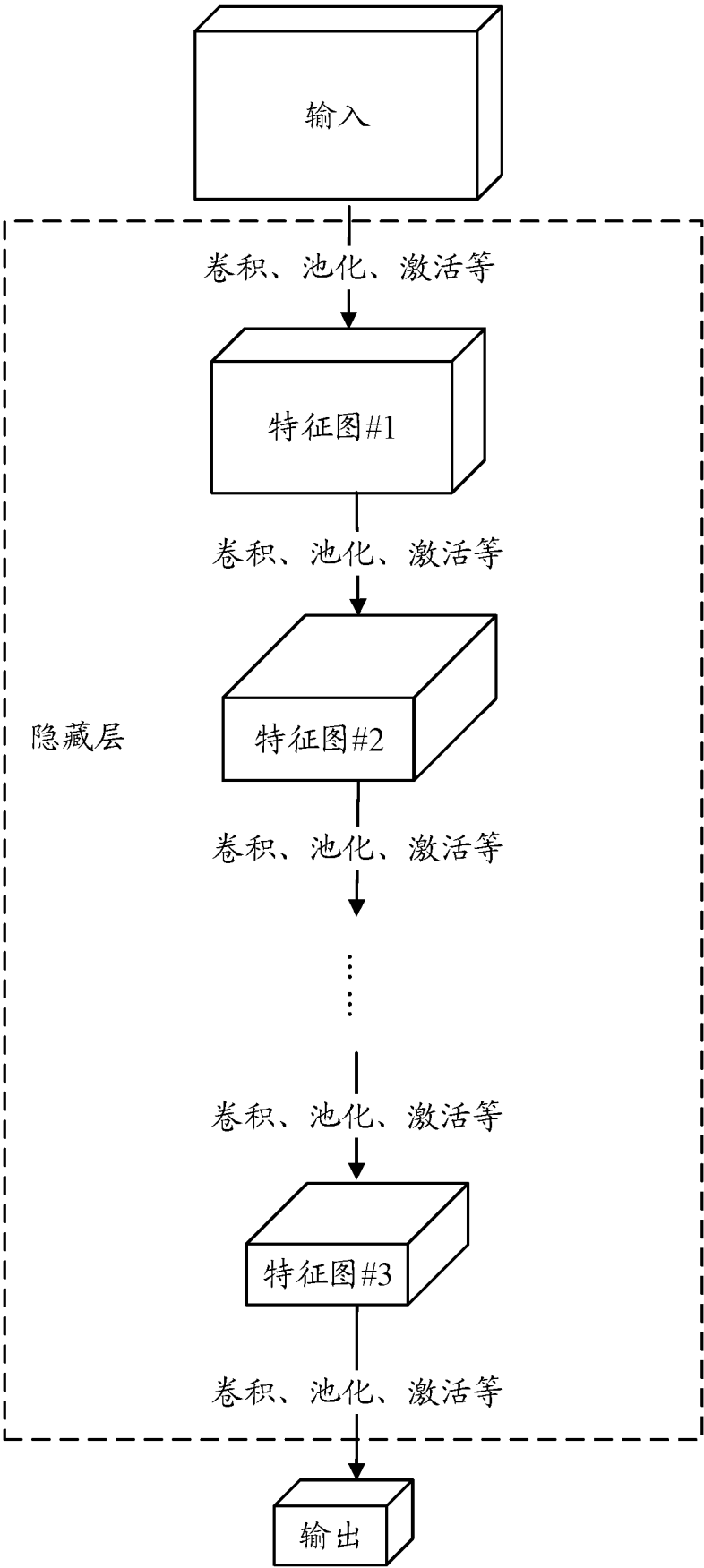


图 1

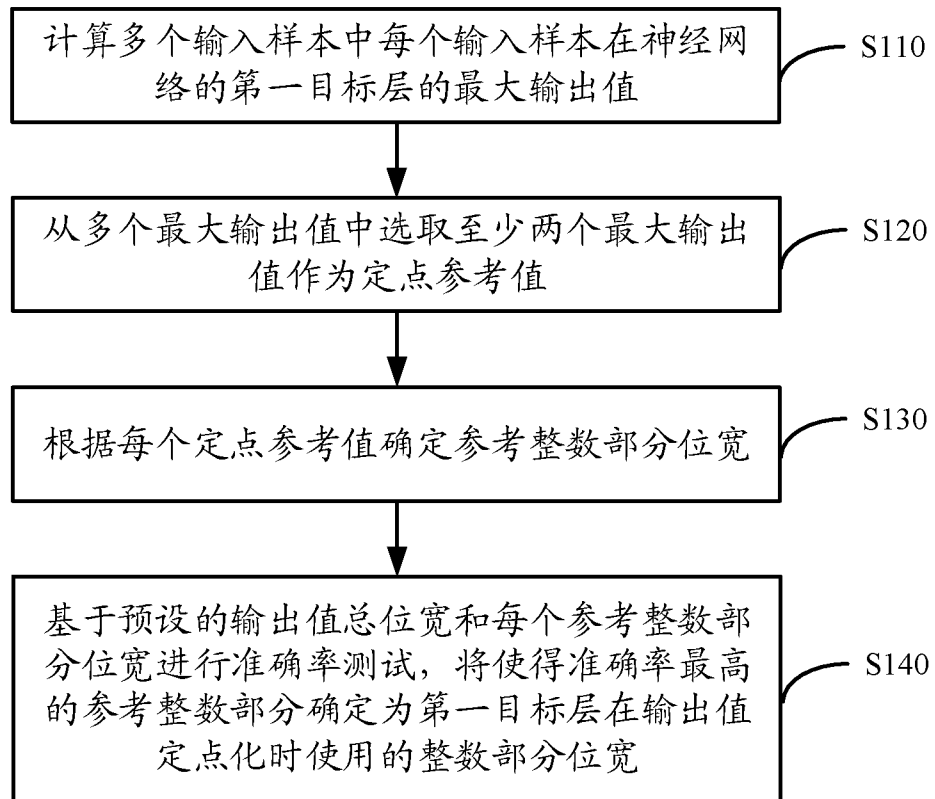


图 2

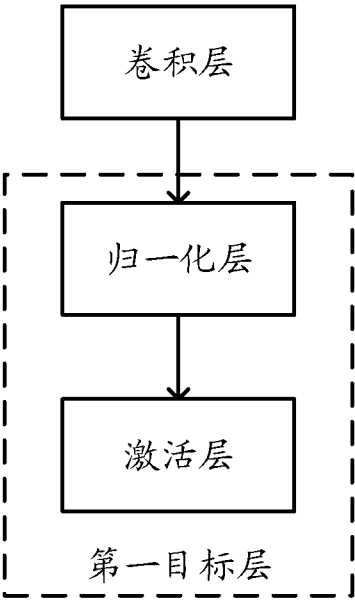


图3A

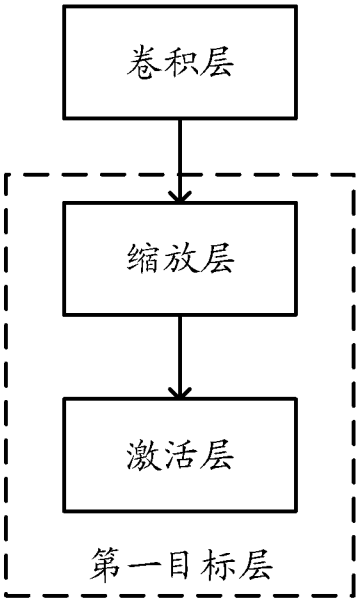


图3B

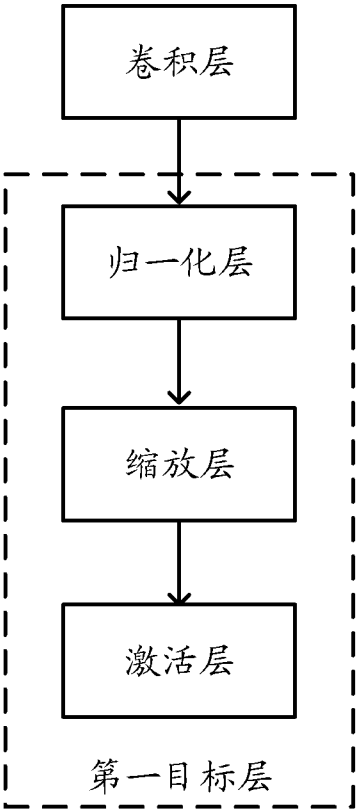


图3C

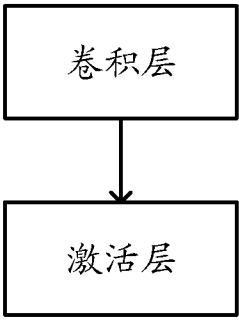


图3D

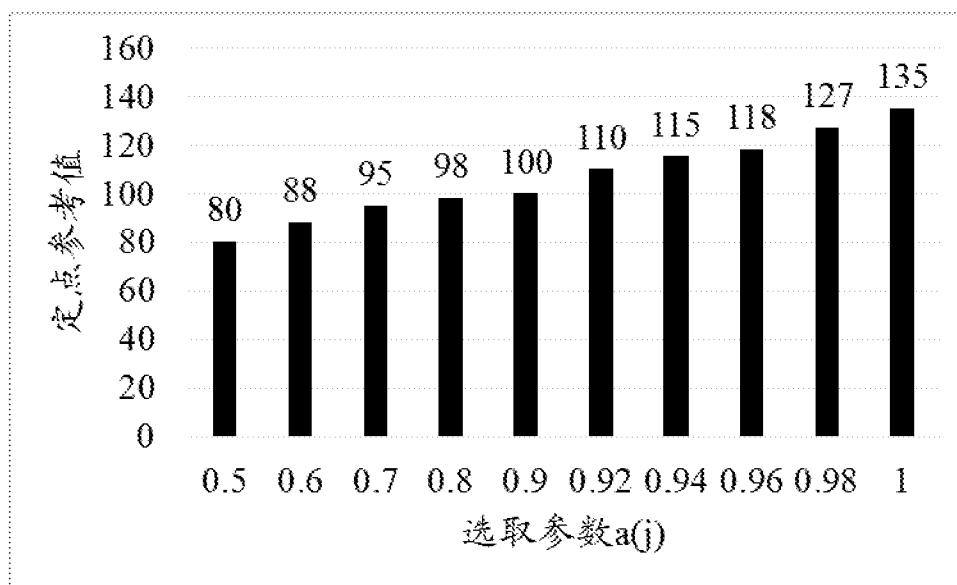


图 4

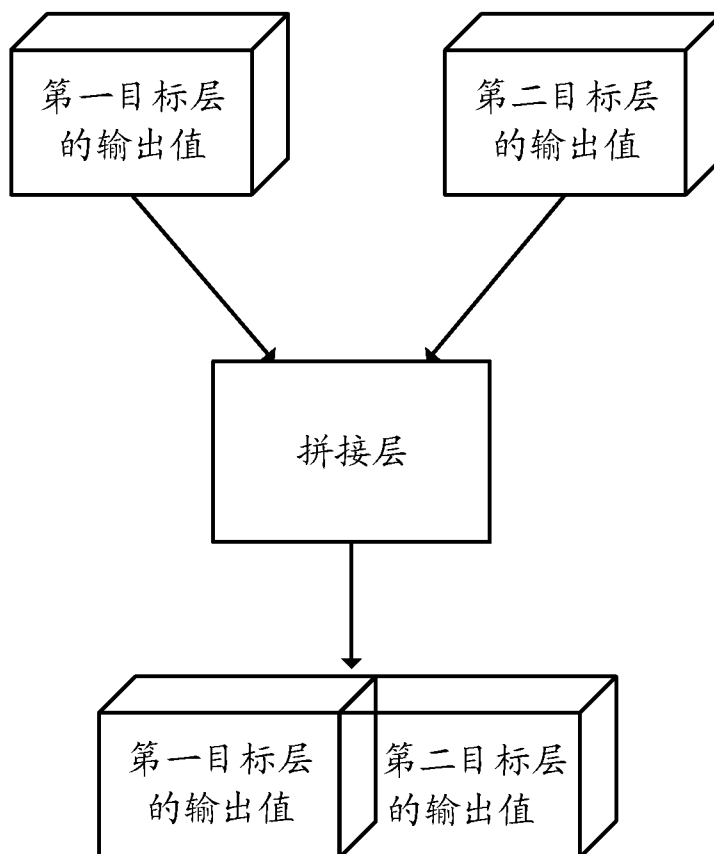


图 5

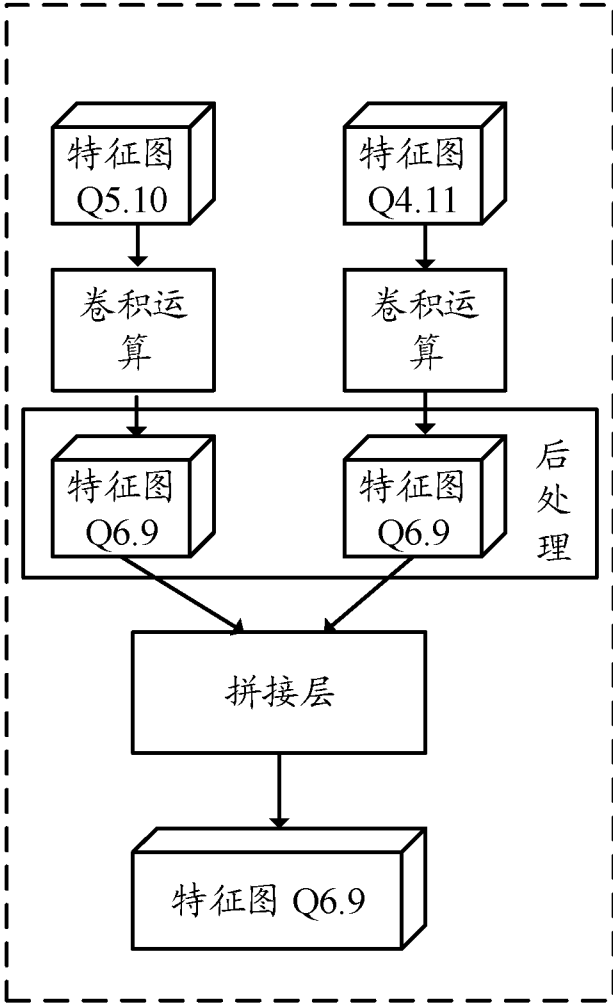


图 6

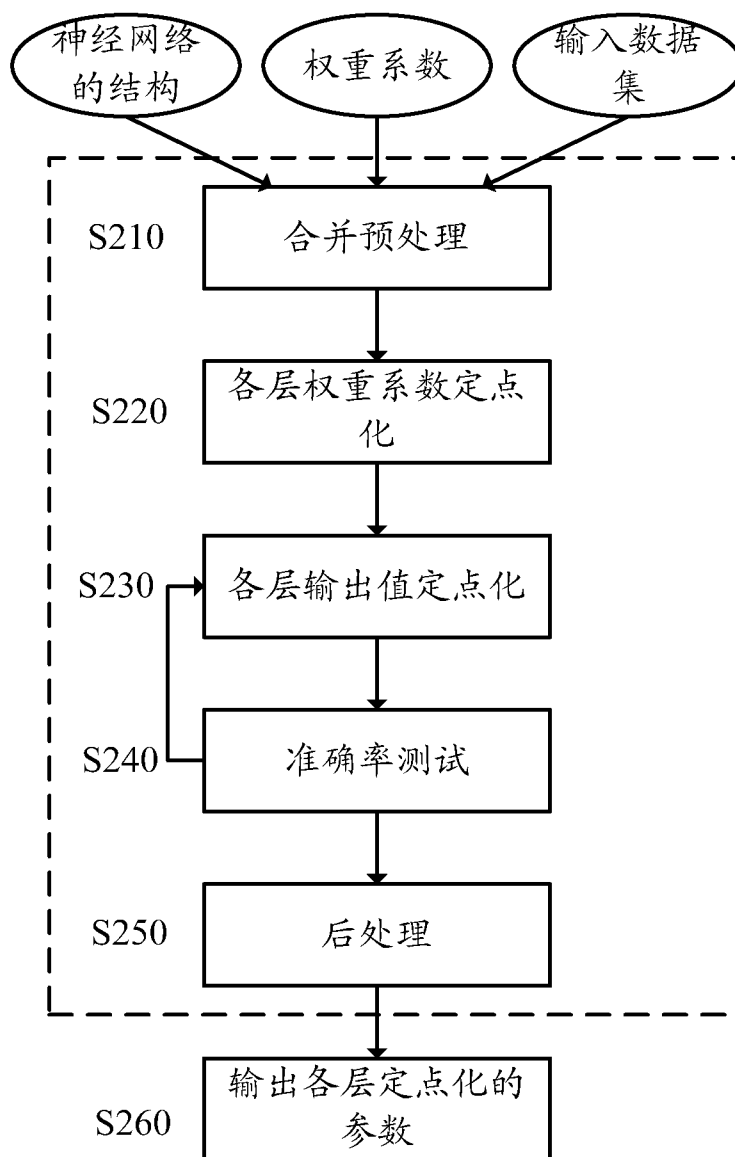


图 7

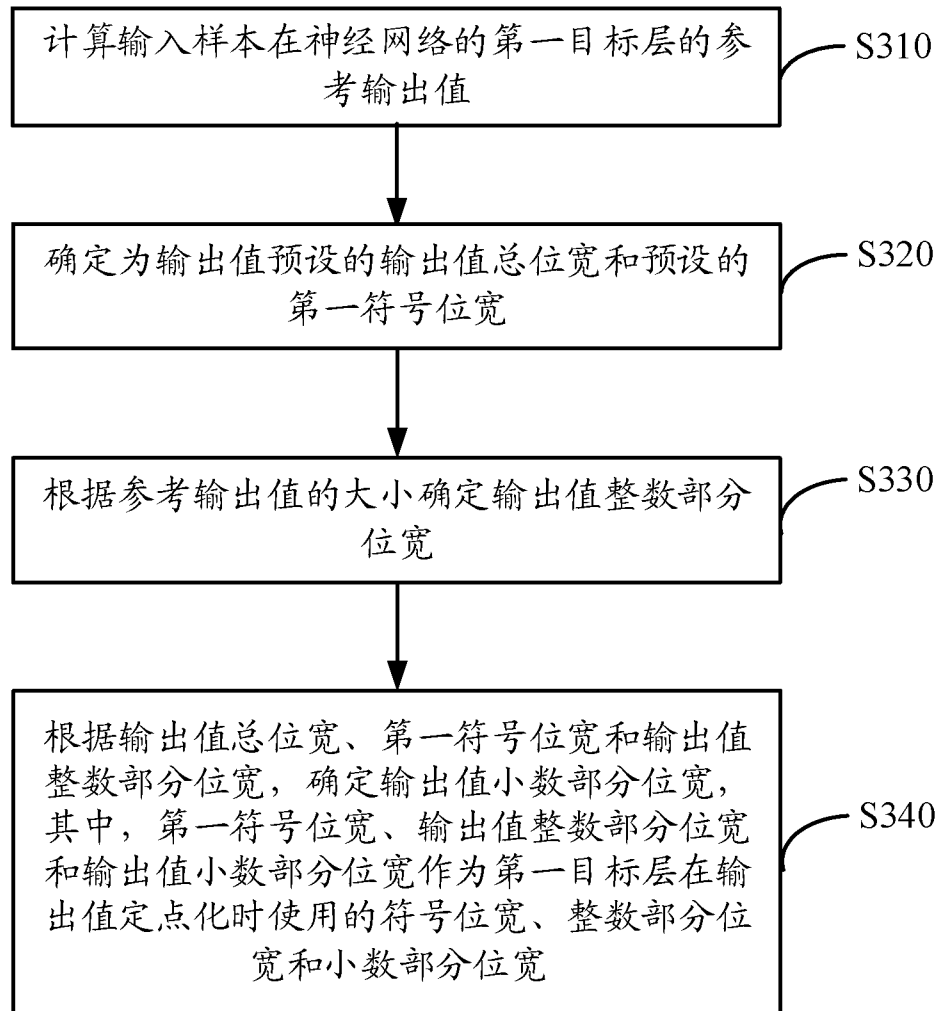


图 8

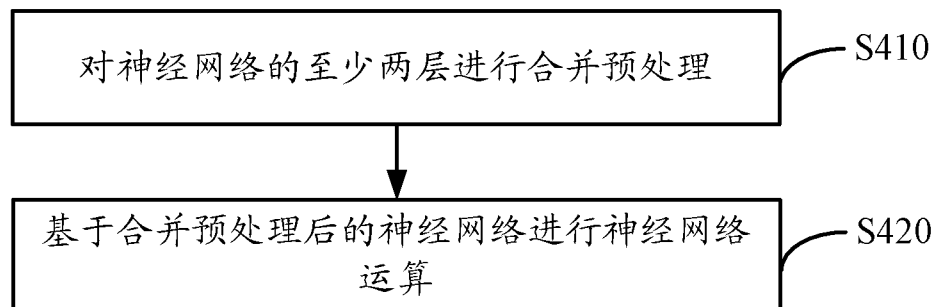


图 9

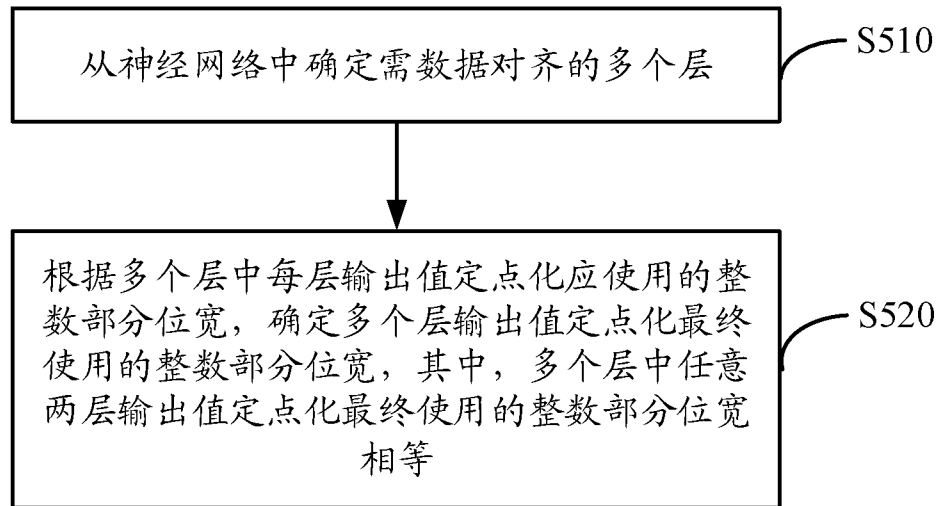


图 10

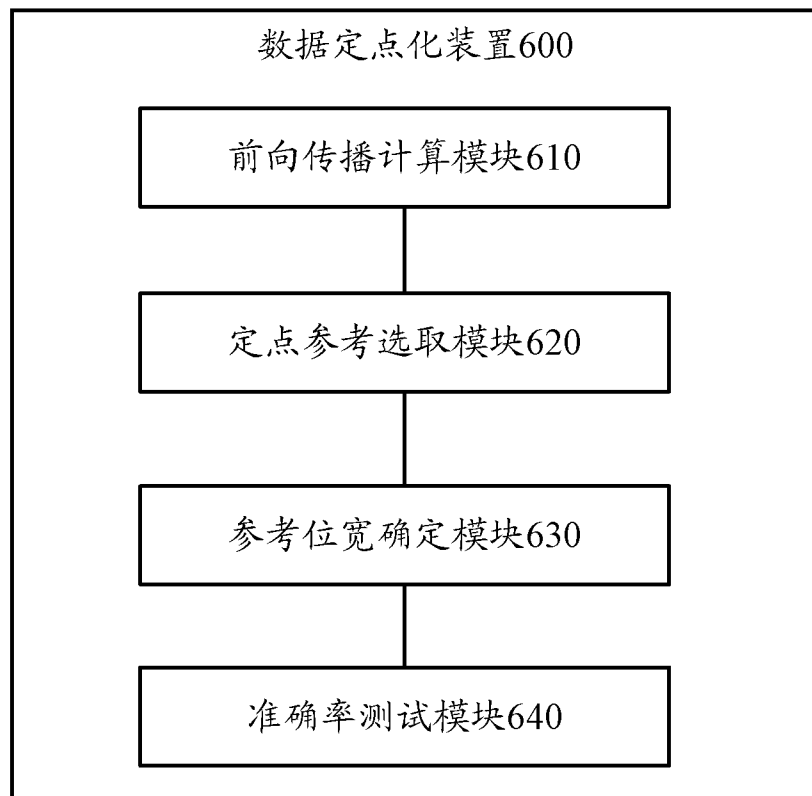


图 11

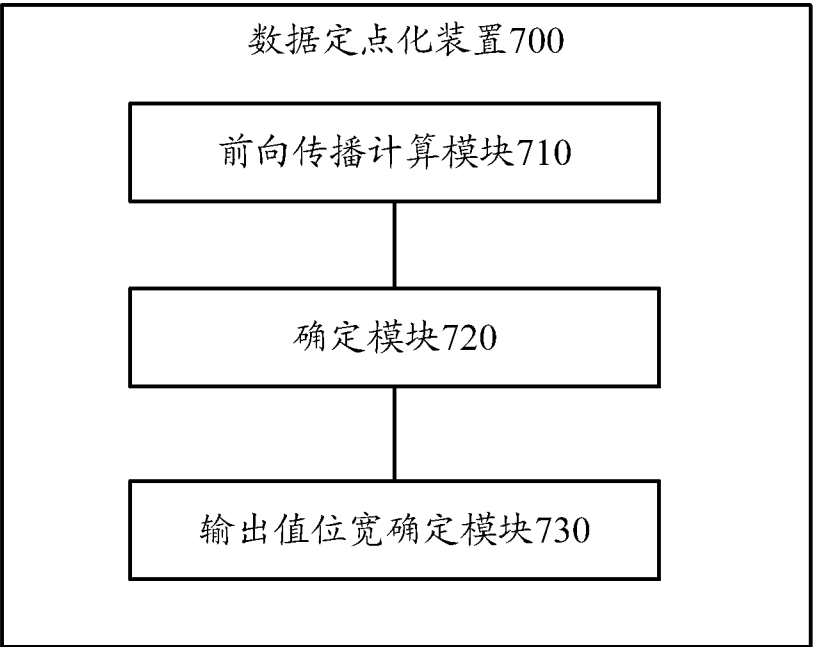


图 12



图 13



图 14

INTERNATIONAL SEARCH REPORT

International application No.

PCT/CN2017/106333

A. CLASSIFICATION OF SUBJECT MATTER

G06N 3/08(2006.01)i; G06F 17/11(2006.01)i; G06F 9/318(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06N; G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

WPI, PODOC, NKI, NPAT, GOOGLE, 数据, 定点, 输出, 层, 神经网络, 整数, 位宽, 准确率, 合并, data, fixed point, bit width, layer, accuracy, learning, model, integer

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2016328647 A1 (QUALCOMM INCORPORATED) 10 November 2016 (2016-11-10) claims 1-14, description, paragraphs 25-101, and figures 1-5	28, 30, 60, 62
A	US 2016328647 A1 (QUALCOMM INCORPORATED) 10 November 2016 (2016-11-10) claims 1-14, description, paragraphs 25-101, and figures 1-5	1-22, 27, 29, 31-54, 59, 61, 63-64
X	CN 104915654 A (ZHEJIANG UNIVERSITY OF TECHNOLOGY) 16 September 2015 (2015-09-16) description, paragraphs 25-53, and figures 1-2	23-26, 55-58
A	CN 102622207 A (ZTE CORPORATION) 01 August 2012 (2012-08-01) entire document	1-64
A	US 2016328646 A1 (QUALCOMM INCORPORATED) 10 November 2016 (2016-11-10) entire document	1-64

☐ Further documents are listed in the continuation of Box C.☒ See patent family annex.

* Special categories of cited documents:

“A” document defining the general state of the art which is not considered to be of particular relevance

“E” earlier application or patent but published on or after the international filing date

“L” document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

“O” document referring to an oral disclosure, use, exhibition or other means

“P” document published prior to the international filing date but later than the priority date claimed

“T” later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

“X” document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

“Y” document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

“&” document member of the same patent family

Date of the actual completion of the international search

03 July 2018

Date of mailing of the international search report

18 July 2018

Name and mailing address of the ISA/CN

State Intellectual Property Office of the P. R. China
No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing
100088
China

Facsimile No. (86-10)62019451

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.

PCT/CN2017/106333

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
US	2016328647	A1	10 November 2016	WO	2016182659	A1	17 November 2016
				CN	107646116	A	30 January 2018
				EP	3295382	A1	21 March 2018
				IN	201747033714	A	29 September 2017
CN	104915654	A	16 September 2015	None			
CN	102622207	A	01 August 2012	None			
US	2016328646	A1	10 November 2016	EP	3295385	A1	21 March 2018
				CN	107636697	A	26 January 2018
				WO	2016182671	A1	17 November 2016
				IN	201747033566	A	29 September 2017

国际检索报告

国际申请号

PCT/CN2017/106333

A. 主题的分类 G06N 3/08(2006.01)i; G06F 17/11(2006.01)i; G06F 9/318(2006.01)i 按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类																				
B. 检索领域 检索的最低限度文献(标明分类系统和分类号) G06N; G06F 包含在检索领域中的除最低限度文献以外的检索文献 在国际检索时查阅的电子数据库(数据库的名称, 和使用的检索词(如使用)) WPI, EPODOC, CNKI, CNPAT, GOOGLE, 数据, 定点, 输出, 层, 神经网络, 整数, 位宽, 准确率, 合并, data, fixed point, bit width, layer, accuracy, learning, model, integer																				
C. 相关文件 <table border="1"> <thead> <tr> <th>类 型*</th> <th>引用文件, 必要时, 指明相关段落</th> <th>相关的权利要求</th> </tr> </thead> <tbody> <tr> <td>X</td> <td>US 2016328647 A1 (QUALCOMM INCORPORATED) 2016年 11月 10日 (2016 - 11 - 10) 权利要求1至14, 说明书25至101段及图1至5</td> <td>28, 30, 60, 62</td> </tr> <tr> <td>A</td> <td>US 2016328647 A1 (QUALCOMM INCORPORATED) 2016年 11月 10日 (2016 - 11 - 10) 权利要求1至14, 说明书25至101段及图1至5</td> <td>1-22, 27, 29, 31-54, 59, 61, 63-64</td> </tr> <tr> <td>X</td> <td>CN 104915654 A (浙江工业大学) 2015年 9月 16日 (2015 - 09 - 16) 说明书第25至53段及图1至2</td> <td>23-26, 55-58</td> </tr> <tr> <td>A</td> <td>CN 102622207 A (中兴通讯股份有限公司) 2012年 8月 1日 (2012 - 08 - 01) 全文</td> <td>1-64</td> </tr> <tr> <td>A</td> <td>US 2016328646 A1 (QUALCOMM INCORPORATED) 2016年 11月 10日 (2016 - 11 - 10) 全文</td> <td>1-64</td> </tr> </tbody> </table>			类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求	X	US 2016328647 A1 (QUALCOMM INCORPORATED) 2016年 11月 10日 (2016 - 11 - 10) 权利要求1至14, 说明书25至101段及图1至5	28, 30, 60, 62	A	US 2016328647 A1 (QUALCOMM INCORPORATED) 2016年 11月 10日 (2016 - 11 - 10) 权利要求1至14, 说明书25至101段及图1至5	1-22, 27, 29, 31-54, 59, 61, 63-64	X	CN 104915654 A (浙江工业大学) 2015年 9月 16日 (2015 - 09 - 16) 说明书第25至53段及图1至2	23-26, 55-58	A	CN 102622207 A (中兴通讯股份有限公司) 2012年 8月 1日 (2012 - 08 - 01) 全文	1-64	A	US 2016328646 A1 (QUALCOMM INCORPORATED) 2016年 11月 10日 (2016 - 11 - 10) 全文	1-64
类 型*	引用文件, 必要时, 指明相关段落	相关的权利要求																		
X	US 2016328647 A1 (QUALCOMM INCORPORATED) 2016年 11月 10日 (2016 - 11 - 10) 权利要求1至14, 说明书25至101段及图1至5	28, 30, 60, 62																		
A	US 2016328647 A1 (QUALCOMM INCORPORATED) 2016年 11月 10日 (2016 - 11 - 10) 权利要求1至14, 说明书25至101段及图1至5	1-22, 27, 29, 31-54, 59, 61, 63-64																		
X	CN 104915654 A (浙江工业大学) 2015年 9月 16日 (2015 - 09 - 16) 说明书第25至53段及图1至2	23-26, 55-58																		
A	CN 102622207 A (中兴通讯股份有限公司) 2012年 8月 1日 (2012 - 08 - 01) 全文	1-64																		
A	US 2016328646 A1 (QUALCOMM INCORPORATED) 2016年 11月 10日 (2016 - 11 - 10) 全文	1-64																		
☐ 其余文件在C栏的续页中列出。 ☒ 见同族专利附件。																				
* 引用文件的具体类型: “A” 认为不特别相关的表示了现有技术一般状态的文件 “E” 在国际申请日的当天或之后公布的在先申请或专利 “L” 可能对优先权要求构成怀疑的文件, 或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件(如具体说明的) “O” 涉及口头公开、使用、展览或其他方式公开的文件 “P” 公布日先于国际申请日但迟于所要求的优先权日的文件 “T” 在申请日或优先权日之后公布, 与申请不相抵触, 但为了理解发明之理论或原理的在后文件 “X” 特别相关的文件, 单独考虑该文件, 认定要求保护的发明不是新颖的或不具有创造性 “Y” 特别相关的文件, 当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时, 要求保护的发明不具有创造性 “&” 同族专利的文件																				
国际检索实际完成的日期 2018年 7月 3日		国际检索报告邮寄日期 2018年 7月 18日																		
ISA/CN的名称和邮寄地址 中华人民共和国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088 传真号 (86-10)62019451		受权官员 沈敏洁 电话号码 86-(10)-53961293																		

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2017/106333

检索报告引用的专利文件			公布日 (年/月/日)	同族专利			公布日 (年/月/日)
US	2016328647	A1	2016年 11月 10日	WO	2016182659	A1	2016年 11月 17日
				CN	107646116	A	2018年 1月 30日
				EP	3295382	A1	2018年 3月 21日
				IN	201747033714	A	2017年 9月 29日
CN	104915654	A	2015年 9月 16日	无			
CN	102622207	A	2012年 8月 1日	无			
US	2016328646	A1	2016年 11月 10日	EP	3295385	A1	2018年 3月 21日
				CN	107636697	A	2018年 1月 26日
				WO	2016182671	A1	2016年 11月 17日
				IN	201747033566	A	2017年 9月 29日

表 PCT/ISA/210 (同族专利附件) (2015年1月)