(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 105893159 B
(45)授权公告日 2018.06.19

(21)申请号 201610454483.1

(22)申请日 2016.06.21

(65)同一申请的已公布的文献号
申请公布号 CN 105893159 A

(43)申请公布日 2016.08.24

(73)专利权人 北京百度网讯科技有限公司
地址 100085 北京市海淀区上地十街10号百度大厦2层

(72)发明人 欧阳剑 漆维 王勇

(74)专利代理机构 北京英赛嘉华知识产权代理有限责任公司 11204
代理人 王达佐 马晓亚

(51)Int.Cl.
G06F 9/50(2006.01)

(56)对比文件
CN 104484703 A,2015.04.01,全文.
CN 103970719 A,2014.08.06,全文.
CN 105512723 A,2016.04.20,全文.
CN 104572504 A,2015.04.29,全文.
US 5226092 A,1993.07.06,全文.
US 5479574 A,1995.12.26,全文.

审查员 詹芊芊

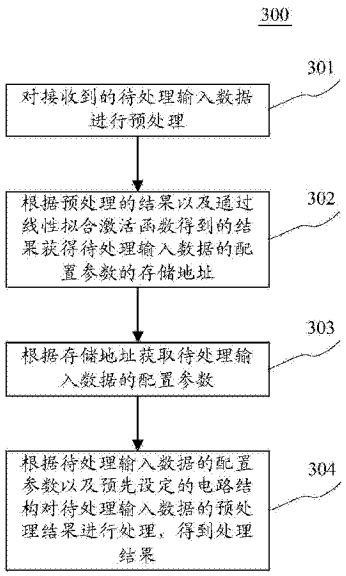
权利要求书3页 说明书9页 附图4页

(54)发明名称

数据处理方法和装置

(57)摘要

本申请公开了数据处理方法和装置。所述方法的一具体实施方式包括:对接收到的待处理输入数据进行预处理;根据预处理的结果以及通过线性拟合激活函数得到的结果获得所述待处理输入数据的配置参数的存储地址,其中,配置参数是根据激活函数的曲线特性预先设置的;根据所述存储地址获取所述待处理输入数据的配置参数;根据所述待处理输入数据的配置参数以及预先设定的电路结构对所述待处理输入数据的预处理结果进行处理,得到处理结果。该实施方式实现了使用配置参数和预先设定的电路结构实现对待处理输入数据的处理,不需要使用用于实现激活函数的专用电路,从而简化了电路结构,且同时可以支持多种激活函数,提高了灵活性。



CN 105893159 B

1.一种数据处理方法,其特征在于,所述方法包括:

对接收到的待处理输入数据进行预处理;

根据预处理的结果以及通过线性拟合激活函数得到的结果获得所述待处理输入数据的配置参数的存储地址,其中,配置参数是根据激活函数的曲线特性预先设置的;

根据所述存储地址获取所述待处理输入数据的配置参数;

根据所述待处理输入数据的配置参数以及预先设定的电路结构对所述待处理输入数据的预处理结果进行处理,得到处理结果。

2.根据权利要求1所述的方法,其特征在于,所述待处理输入数据为浮点数;以及

所述对接收到的待处理输入数据进行预处理,包括:

将所述待处理输入数据转化为定点数。

3.根据权利要求1所述的方法,其特征在于,所述方法还包括:

根据激活函数的曲线特征预先设置配置参数的步骤,包括:

根据激活函数曲线的斜率变化得到激活函数输入数据的第一阈值和第二阈值,其中,小于第一阈值或者大于第二阈值的输入数据对应的激活函数曲线的斜率变化小于预先设定的斜率变化阈值;

分别计算小于第一阈值、大于第二阈值、以及第一阈值与第二阈值之间的输入数据对应的配置参数;

将输入数据与配置参数关联存储。

4.根据权利要求3所述的方法,其特征在于,所述分别计算小于第一阈值、大于第二阈值、以及第一阈值与第二阈值之间的输入数据对应的配置参数,包括:

根据至少两个小于第一阈值的输入数据对应的激活函数的输出值计算小于第一阈值的输入数据对应的配置参数;

根据至少两个大于第二阈值的输入数据对应的激活函数的输出值计算大于第二阈值的输入数据对应的配置参数;

将第一阈值与第二阈值之间的输入数据进行非均匀分段,计算各个分段内的输入数据对应的配置参数。

5.根据权利要求4所述的方法,其特征在于,所述将第一阈值与第二阈值之间的输入数据进行非均匀分段,计算各个分段内的输入数据对应的配置参数,包括:

采用逐步逼近的方法将第一阈值与第二阈值之间的输入数据划分为至少一个非均匀分段;

根据各个非均匀分段内至少两点的输入数据对应的激活函数的输出值计算各个非均匀分段内的输入数据对应的配置参数。

6.根据权利要求5所述的方法,其特征在于,所述采用逐步逼近的方法将第一阈值与第二阈值之间的输入数据划分为至少一个非均匀分段,包括:

将第一阈值与第二阈值之间的输入数据均匀划分为设定个区间,并将每个区间均匀划分为设定个子区间;

连接各区间中相邻的子区间的1/2抽样值,得到至少一条直线,判断各条直线与相应位置的激活函数的输出值曲线的误差是否超过预先设定的误差阈值;

如果没有超过,则将得到的各条直线中的两两相邻的直线进行合并,得到一条直线,判

断该直线与相应位置激活函数的输出值曲线的误差是否超过所述误差阈值；

依次类推，得到各个区间内的至少一条直线，每条直线对应的取值范围为输入数据的一个分段。

7.根据权利要求6所述的方法，所述根据预处理的结果以及通过线性拟合激活函数得到的结果获得所述待处理输入数据的配置参数的存储地址，包括：

将所述待处理输入数据预处理后的结果与第一阈值和第二阈值进行比较；

如果所述待处理输入数据预处理后的结果小于第一阈值，则得到小于第一阈值的输入数据对应的配置参数的存储地址；

如果所述待处理输入数据预处理后的结果大于第二阈值，则得到大于第二阈值的输入数据对应的配置参数的存储地址；

如果所述待处理输入数据预处理后的结果在所述第一阈值和所述第二阈值之间，则查找所述待处理输入数据预处理后的结果所在区间的分段范围，并得到该分段内的输入数据对应的配置参数的存储地址。

8.一种数据处理装置，其特征在于，所述装置包括：

预处理单元，用于对接收到的待处理输入数据进行预处理；

存储地址获取单元，用于根据预处理的结果以及通过线性拟合激活函数得到的结果获得所述待处理输入数据的配置参数的存储地址，其中，配置参数是根据激活函数的曲线特性预先设置的；

配置参数获取单元，用于根据所述存储地址获取所述待处理输入数据的配置参数；

处理单元，用于根据所述待处理输入数据的配置参数以及预先设定的电路结构对所述待处理输入数据的预处理结果进行处理，得到处理结果。

9.根据权利要求8所述的装置，其特征在于，所述待处理输入数据为浮点数；以及所述预处理单元进一步用于：

将所述待处理输入数据转化为定点数。

10.根据权利要求8所述的装置，其特征在于，所述装置还包括参数配置单元，所述参数配置单元包括阈值获取单元、计算单元和存储单元；以及

所述阈值获取单元，用于根据激活函数曲线的斜率变化得到激活函数输入数据的第一阈值和第二阈值，其中，小于第一阈值或者大于第二阈值的输入数据对应的激活函数曲线的斜率变化小于预先设定的斜率变化阈值；

所述计算单元，用于分别计算小于第一阈值、大于第二阈值、以及第一阈值与第二阈值之间的输入数据对应的配置参数；

所述存储单元，用于将输入数据与配置参数关联存储。

11.根据权利要求10所述的装置，其特征在于，所述计算单元包括第一计算子单元、第二计算子单元和第三计算子单元；以及

所述第一计算子单元，用于根据至少两个小于第一阈值的输入数据对应的激活函数的输出值计算小于第一阈值的输入数据对应的配置参数；

所述第二计算子单元，用于根据至少两个大于第二阈值的输入数据对应的激活函数的输出值计算大于第二阈值的输入数据对应的配置参数；

所述第三计算子单元，用于将第一阈值与第二阈值之间的输入数据进行非均匀分段，

计算各个分段内的输入数据对应的配置参数。

12.根据权利要求11所述的装置，其特征在于，所述第三计算子单元包括分段单元和段内配置参数计算单元；以及

所述分段单元，用于采用逐步逼近的方法将第一阈值与第二阈值之间的输入数据划分为至少一个非均匀分段；

所述段内配置参数计算单元，包括根据各个非均匀分段内至少两点的输入数据对应的激活函数的输出值计算各个非均匀分段内的输入数据对应的配置参数。

13.根据权利要求12所述的装置，其特征在于，所述分段单元进一步用于：

将第一阈值与第二阈值之间的输入数据均匀划分为设定个区间，并将每个区间均匀划分为设定个子区间；

连接各区间中相邻的子区间的1/2抽样值，得到至少一条直线，判断各条直线与相应位置的激活函数的输出值曲线的误差是否超过预先设定的误差阈值；

如果没有超过，则将得到的各条直线中的两两相邻的直线进行合并，得到一条直线，判断该直线与相应位置激活函数的输出值曲线的误差是否超过所述误差阈值；

依次类推，得到各个区间内的至少一条直线，每条直线对应的取值范围为输入数据的一个分段。

14.根据权利要求13所述的装置，其特征在于，所述存储地址获取单元进一步用于：

将所述待处理输入数据预处理后的结果与第一阈值和第二阈值进行比较；

如果所述待处理输入数据预处理后的结果小于第一阈值，则得到小于第一阈值的输入数据对应的配置参数的存储地址；

如果所述待处理输入数据预处理后的结果大于第二阈值，则得到大于第二阈值的输入数据对应的配置参数的存储地址；

如果所述待处理输入数据预处理后的结果在所述第一阈值和所述第二阈值之间，则查找所述待处理输入数据预处理后的结果所在区间的分段范围，并得到该分段内的输入数据对应的配置参数的存储地址。

15.一种人工智能处理器，包括如权利要求8-14中任一项所述的数据处理装置，所述人工智能处理器包括两组或两组以上的寄存器堆和静态随机存取存储器，每组寄存器堆和静态随机存取存储器用于存储一种激活函数的配置参数，且每组寄存器堆和静态随机存取存储器中存储的激活函数的配置参数是动态更新的。

数据处理方法和装置

技术领域

[0001] 本申请涉及通信技术领域，具体涉及数据处理领域，尤其涉及数据处理方法和装置。

背景技术

[0002] 深度学习算法是人工智能的核心，对很多领域(例如语音识别、图片识别、自然语言处理等)的发展起到了极大的推动作用。深度学习算法是典型计算密集型算法，一般计算复杂度是O(N^3)(立方阶)，往往比传统的机器学习算法有一到两个数量级的提高。另一方面，深度学习算法往往与大数据紧密联系在一起，一般需要TB到PB级别的训练数据，数百万到千亿级别的训练参数，才能得到精度较好的模型。综合两点，在实际的应用中，深度学习算法对计算量要求极大，传统的CPU(Central Processing Unit，中央处理器)根本无法满足其计算需求，为了解决深度学习算法的计算瓶颈，很多公司设计了深度学习算法的专有芯片，如百度的人工智能计算机，Google(谷歌)的TPU(Tensor Processing Unit，张量处理单元)等。

[0003] 如图1所示，深度学习算法的网络结构一般有N层(layer)网络，其中，N可以是几层到数十层，每一个layer可以是DNN(Deep Neural Network，深度神经网络)、RNN(Recurrent Neural Networks，循环神经网络)、或者CNN(Convolutional Neural Network，卷积神经网络)结构等，layer之间有激活函数，常用激活函数有十几种，每一层之间的激活函数可能会重复，也可能不一样。现有的技术一般采用以下两种方式计算各种激活函数：一、采用通用处理器通过软件编程的方式，这种方式效率比较低，因为通用处理器处理激活函数这种复杂计算比较慢；二、采用专有硬件电路实现，专有电路实现激活函数的代价很高，一方面激活函数复杂，每个函数都需要消耗很多电路资源，另一方面，同时支持多种激活函数，总的电路资源消耗大，此外，专有的电路结构不灵活，如果有新的激活函数不能灵活支持。

发明内容

[0004] 本申请的目的在于提出一种改进的数据处理方法和装置，来解决以上背景技术部分提到的技术问题。

[0005] 第一方面，本申请提供了一种数据处理方法，所述方法包括：对接收到的待处理输入数据进行预处理；根据预处理的结果以及通过线性拟合激活函数得到的结果获得所述待处理输入数据的配置参数的存储地址，其中，配置参数是根据激活函数的曲线特性预先设置的；根据所述存储地址获取所述待处理输入数据的配置参数；根据所述待处理输入数据的配置参数以及预先设定的电路结构对所述待处理输入数据的预处理结果进行处理，得到处理结果。

[0006] 在一些实施例中，所述待处理输入数据为浮点数；以及所述对接收到的待处理输入数据进行预处理，包括：将所述待处理输入数据转化为定点数。

[0007] 在一些实施例中，所述方法还包括：根据激活函数的曲线特征预先设置配置参数

的步骤，包括：根据激活函数曲线的斜率变化得到激活函数输入数据的第一阈值和第二阈值，其中，小于第一阈值或者大于第二阈值的输入数据对应的激活函数曲线的斜率变化小于预先设定的斜率变化阈值；分别计算小于第一阈值、大于第二阈值、以及第一阈值与第二阈值之间的输入数据对应的配置参数；将输入数据与配置参数关联存储。

[0008] 在一些实施例中，所述分别计算小于第一阈值、大于第二阈值、以及第一阈值与第二阈值之间的输入数据对应的配置参数，包括：根据至少两个小于第一阈值的输入数据对应的激活函数的输出值计算小于第一阈值的输入数据对应的配置参数；根据至少两个大于第二阈值的输入数据对应的激活函数的输出值计算大于第二阈值的输入数据对应的配置参数；将第一阈值与第二阈值之间的输入数据进行非均匀分段，计算各个分段内的输入数据对应的配置参数。

[0009] 在一些实施例中，所述将第一阈值与第二阈值之间的输入数据进行非均匀分段，计算各个分段内的输入数据对应的配置参数，包括：采用逐步逼近的方法将第一阈值与第二阈值之间的输入数据划分为至少一个非均匀分段；根据各个非均匀分段内至少两点的输入数据对应的激活函数的输出值计算各个非均匀分段内的输入数据对应的配置参数。

[0010] 在一些实施例中，所述采用逐步逼近的方法将第一阈值与第二阈值之间的输入数据划分为至少一个非均匀分段，包括：将第一阈值与第二阈值之间的输入数据均匀划分为设定个区间，并将每个区间均匀划分为设定个子区间；连接各区间中相邻的子区间的1/2抽样值，得到至少一条直线，判断各条直线与相应位置的激活函数的输出值曲线的误差是否超过预先设定的误差阈值；如果没有超过，则将得到的各条直线中的两两相邻的直线进行合并，得到一条直线，判断该直线与相应位置激活函数的输出值曲线的误差是否超过所述误差阈值；依次类推，得到各个区间内的至少一条直线，每条直线对应的取值范围为输入数据的一个分段。

[0011] 在一些实施例中，所述根据预处理的结果以及通过线性拟合激活函数得到的结果获得所述待处理输入数据的配置参数的存储地址，包括：将所述待处理输入数据预处理后的结果与第一阈值和第二阈值进行比较；如果所述待处理输入数据预处理后的结果小于第一阈值，则得到小于第一阈值的输入数据对应的配置参数的存储地址；如果所述待处理输入数据预处理后的结果大于第二阈值，则得到大于第二阈值的输入数据对应的配置参数的存储地址；如果所述待处理输入数据预处理后的结果在所述第一阈值和所述第二阈值之间，则查找所述待处理输入数据预处理后的结果所在区间的分段范围，并得到该分段内的输入数据对应的配置参数的存储地址。

[0012] 第二方面，本申请提供了一种数据处理装置，所述装置包括：预处理单元，用于对接收到的待处理输入数据进行预处理；存储地址获取单元，用于根据预处理的结果以及通过线性拟合激活函数得到的结果获得所述待处理输入数据的配置参数的存储地址，其中，配置参数是根据激活函数的曲线特性预先设置的；配置参数获取单元，用于根据所述存储地址获取所述待处理输入数据的配置参数；处理单元，用于根据所述待处理输入数据的配置参数以及预先设定的电路结构对所述待处理输入数据的预处理结果进行处理，得到处理结果。

[0013] 在一些实施例中，所述待处理输入数据为浮点数；以及所述预处理单元进一步用于：将所述待处理输入数据转化为定点数。

[0014] 在一些实施例中，所述装置还包括参数配置单元，所述参数配置单元包括阈值获取单元、计算单元和存储单元；以及所述阈值获取单元，用于根据激活函数曲线的斜率变化得到激活函数输入数据的第一阈值和第二阈值，其中，小于第一阈值或者大于第二阈值的输入数据对应的激活函数曲线的斜率变化小于预先设定的斜率变化阈值；所述计算单元，用于分别计算小于第一阈值、大于第二阈值、以及第一阈值与第二阈值之间的输入数据对应的配置参数；所述存储单元，用于将输入数据与配置参数关联存储。

[0015] 在一些实施例中，所述计算单元包括第一计算子单元、第二计算子单元和第三计算子单元；以及所述第一计算子单元，用于根据至少两个小于第一阈值的输入数据对应的激活函数的输出值计算小于第一阈值的输入数据对应的配置参数；所述第二计算子单元，用于根据至少两个大于第二阈值的输入数据对应的激活函数的输出值计算大于第二阈值的输入数据对应的配置参数；所述第三计算子单元，用于将第一阈值与第二阈值之间的输入数据进行非均匀分段，计算各个分段内的输入数据对应的配置参数。

[0016] 在一些实施例中，所述第三计算子单元包括分段单元和段内配置参数计算单元；以及所述分段单元，用于采用逐步逼近的方法将第一阈值与第二阈值之间的输入数据划分为至少一个非均匀分段；所述段内配置参数计算单元，包括根据各个非均匀分段内至少两点的输入数据对应的激活函数的输出值计算各个非均匀分段内的输入数据对应的配置参数。

[0017] 在一些实施例中，所述分段单元进一步用于：将第一阈值与第二阈值之间的输入数据均匀划分为设定个区间，并将每个区间均匀划分为设定个子区间；连接各区间中相邻的子区间的1/2抽样值，得到至少一条直线，判断各条直线与相应位置的激活函数的输出值曲线的误差是否超过预先设定的误差阈值；如果没有超过，则将得到的各条直线中的两两相邻的直线进行合并，得到一条直线，判断该直线与相应位置激活函数的输出值曲线的误差是否超过所述误差阈值；依次类推，得到各个区间内的至少一条直线，每条直线对应的取值范围为输入数据的一个分段。

[0018] 在一些实施例中，所述存储地址获取单元进一步用于：将所述待处理输入数据预处理后的结果与第一阈值和第二阈值进行比较；如果所述待处理输入数据预处理后的结果小于第一阈值，则得到小于第一阈值的输入数据对应的配置参数的存储地址；如果所述待处理输入数据预处理后的结果大于第二阈值，则得到大于第二阈值的输入数据对应的配置参数的存储地址；如果所述待处理输入数据预处理后的结果在所述第一阈值和所述第二阈值之间，则查找所述待处理输入数据预处理后的结果所在区间的分段范围，并得到该分段内的输入数据对应的配置参数的存储地址。

[0019] 第三方面，本申请提供了人工智能处理器，包括如第二方面中任一项所述的数据处理装置，所述人工智能处理器包括两组或两组以上的寄存器堆和静态随机存取存储器，每组寄存器堆和静态随机存取存储器用于存储一种激活函数的配置参数，且每组寄存器堆和静态随机存取存储器中存储的激活函数的配置参数是动态更新的。

[0020] 本申请提供的数据处理方法和装置，将待处理输入数据预处理之后，根据预处理的结果以及通过线性拟合激活函数得到的结果获得该待处理输入数据的配置参数的存储地址，并从该存储地址中获取待处理输入数据的配置参数，最后根据获取的配置参数以及预先设定的电路结构对待处理输入数据的预处理结果进行处理，因此使用配置参数和预先

设定的电路结构可以实现对待处理输入数据的处理，不需要使用用于实现激活函数的专用电路，从而简化了电路结构，且同时可以支持多种激活函数，提高了灵活性。

附图说明

[0021] 通过阅读参照以下附图所作的对非限制性实施例所作的详细描述，本申请的其它特征、目的和优点将会变得更明显：

[0022] 图1是现有的深度学习算法的网络结构的示例图；

[0023] 图2是本申请可以应用于其中的示例性系统架构图；

[0024] 图3是根据本申请的数据处理方法的一个实施例的流程图；

[0025] 图4是本申请对待处理输入数据进行处理的示例性说明；

[0026] 图5是激活函数(1)的曲线图；

[0027] 图6是根据比较器获取配置参数的存储地址的示例性说明；

[0028] 图7是根据本申请的数据处理装置的一个实施例的结构示意图；

具体实施方式

[0029] 下面结合附图和实施例对本申请作进一步的详细说明。可以理解的是，此处所描述的具体实施例仅仅用于解释相关发明，而非对该发明的限定。另外还需要说明的是，为了便于描述，附图中仅示出了与有关发明相关的部分。

[0030] 需要说明的是，在不冲突的情况下，本申请中的实施例及实施例中的特征可以相互组合。下面将参考附图并结合实施例来详细说明本申请。

[0031] 图2示出了可以应用本申请的数据处理方法或数据处理装置的实施例的示例性系统架构200。

[0032] 如图2所示，系统架构200可以包括CPU(Central Processing Unit,中央处理器)201和人工智能处理器202，CPU201和人工智能处理器202之间可以通过pcie等总线连接。

[0033] CPU201可以用于根据激活函数的曲线特征对激活函数进行分段线性拟合，并根据拟合结果生成配置参数。CPU201将生成的配置参数通过总线传输给人工智能处理器202，人工智能处理器202可以将接收到的配置参数存储到寄存器堆和/或SRAM(Static Random Access Memory,静态随机存取存储器)中，例如使用多条直线分段拟合激活函数曲线，可以将对激活函数进行分段线性拟合得到的与分段相关的数据(例如各个区间的起始坐标，区间内段的数量等)存储在寄存器堆中，将分段线性拟合得到的各个分段对应的直线的相关数据(例如斜率、截距等)存储在SRAM中。应该理解，人工智能处理器202中可以根据实现需要设置任意数目的寄存器堆和SRAM，例如可以设置有两组或两组以上的寄存器堆和SRAM，在使用过程中，每组寄存器堆和SRAM可以存储有一种激活函数的相关配置参数，且每组寄存器堆和SRAM中存储的数据是可以根据处理需要动态重配置的，两组或两组以上的寄存器堆和SRAM可以提高人工智能处理器202的数据处理并行度。例如当前layer在计算的同时，可以并行配置下一个layer的激活函数参数表，配置和layer计算并行，能隐藏配置的时间，以保证在切换激活函数的时候，不需要额外的时间等待参数的配置。

[0034] 需要说明的是，本申请实施例所提供的数据处理方法一般由人工智能处理器202执行，相应地，数据处理装置一般设置于人工智能处理器202中。

[0035] 继续参考图3，示出了根据本申请的数据处理方法的一个实施例的流程300。所述的数据处理方法，包括以下步骤：

[0036] 步骤301，对接收到的待处理输入数据进行预处理。

[0037] 在本实施例中，数据处理方法运行于其上的电子设备(例如图2所示的人工智能处理器202)可以对深度学习算法两层之间的激活函数接收到的待处理输入数据进行各种预处理。在这里，上述激活函数可以是各种激活函数。

[0038] 在本实施例的一些可选的实现方式中，上述待处理输入数据可以为浮点数；以及上述电子设备可以将上述待处理输入数据通过各种方式转化为定点数，例如，可以通过预先设置的浮点转定点电路将上述待处理输入数据转化为定点数。

[0039] 步骤302，根据预处理的结果以及通过线性拟合激活函数得到的结果获得待处理输入数据的配置参数的存储地址。

[0040] 在本实施例中，上述电子设备可以根据步骤301得到的预处理结果以及通过线性拟合激活函数得到的结果获得待处理输入数据的配置参数的存储地址，其中，配置参数是根据激活函数的曲线特性预先设置的。例如，上述电子设备可以根据线性拟合激活函数得到的各条直线所在的区间获取待处理输入数据所在区间的配置参数的存储地址。

[0041] 步骤303，根据存储地址获取待处理输入数据的配置参数。

[0042] 在本实施例中，上述电子设备可以从步骤302中得到存储地址中获取待处理输入数据的配置参数。

[0043] 步骤304，根据待处理输入数据的配置参数以及预先设定的电路结构对待处理输入数据的预处理结果进行处理，得到处理结果。

[0044] 在本实施例中，上述电子设备可以根据步骤303获取的配置参数以及预先设定的电路结构对待处理输入数据的预处理结果进行计算等处理。上述预先设定的电路结构可以是各种形式的电路结构，例如当使用多条直线分段拟合激活函数时，上述电路结构可以为实现直线公式Y＝A*X+B的电路结构，其中，X表示输入，Y表示输出，A、B表示配置参数。例如，如图4所示，其示出了本实施例对待处理输入数据进行处理的示例性说明，首先待处理输入数据通过预先设定的浮点转定点电路转化为定点数，之后根据得到的定点数获取配置参数的存储地址，然后根据获取的存储地址在分别用于存储配置参数A、B的参数表A和参数表B中获取配置参数A和B，最后根据获取的配置参数A和B计算Y＝A*X+B，从而得到处理结果，其中，参数表A和参数表B中存储的参数不是固定不变的，可以根据处理过程中所需激活函数的类型动态更新。

[0045] 在本实施例的一些可选的实现方式中，上述数据处理方法还可以包括根据激活函数的曲线特征预先设置配置参数的步骤，该步骤可以由上述电子设备执行，也可以由其他电子设备(例如图2所示的CPU201)执行，该步骤可以包括：

[0046] 首先，根据激活函数曲线的斜率变化得到激活函数输入数据的第一阈值和第二阈值，其中，小于第一阈值或者大于第二阈值的输入数据对应的激活函数曲线的斜率变化小于预先设定的斜率变化阈值，上述斜率变化阈值可以根据实际需要进行设定，在这里，上述激活函数可以为各种激活函数，理论上激活函数的输入数据x的取值范围可以为(-∞，+∞)，但根据激活函数的曲线特性可以发现当输入数据x大于或者小于某个值之后，激活函数曲线的斜率变化就很小，此时非常接近于一条直线，因此可以各用一条直线拟和，例如图

5所示的曲线为如下激活函数(1)的曲线图，通过该曲线图可以看出当x小于-3的时候，曲线就非常接近f(x)＝-1这条直线，当x大于3的时候，曲线就非常接近f(x)＝1这条直线，因此，对于激活函数(1)可以将-3作为第一阈值，将3作为第二阈值；

[0047] $$f(x)=\tanh(x)=\frac{2}{1+e^{-2x}}-1 \qquad (1)$$

[0048] 之后，分别计算小于第一阈值、大于第二阈值、以及第一阈值与第二阈值之间的输入数据对应的配置参数；在这里，上述配置参数可以是线性拟合激活函数所得到的参数；例如当激活函数中x小于第一阈值的区间拟合为直线f(x)＝A_0x+B_0，则A_0和B_0可以为配置参数；当激活函数中x大于第二阈值的区间拟合为直线f(x)＝A_1x+B_1，则A_1和B_1可以为配置参数；

[0049] 最后，将输入数据与配置参数关联存储，当执行该步骤的设备为上述电子设备时，可以将输入数据与配置参数关联存储到本地的存储单元；当执行该步骤的设备为其他电子设备时，该其他电子设备可以将关联存储的输入数据与配置参数通过总线传输给上述电子设备，由上述电子设备进行存储。

[0050] 在一些可选的实现方式中，上述分别计算小于第一阈值、大于第二阈值、以及第一阈值与第二阈值之间的输入数据对应的配置参数，可以通过以下步骤实现：

[0051] 首先，根据至少两个小于第一阈值的输入数据对应的激活函数的输出值计算小于第一阈值的输入数据对应的配置参数；例如，当第一阈值为-3，且激活函数小于-3的区间可以拟合为直线f(x)＝A_0x+B_0，可以选取x＝-300和x＝-600两点处的激活函数的值带入直线公式f(x)＝A_0x+B_0，从而得到A_0和B_0的值；

[0052] 之后，根据至少两个大于第二阈值的输入数据对应的激活函数的输出值计算大于第二阈值的输入数据对应的配置参数，具体方法可以参考以上步骤，此处不再赘述；

[0053] 最后，将第一阈值与第二阈值之间的输入数据进行非均匀分段，计算各个分段内的输入数据对应的配置参数，例如可以将第一阈值与第二阈值之间输入数据根据曲线斜率的变化分段，然后将各个段包含的曲线拟合为直线，在根据以上步骤计算各段的配置参数。通过非均匀分段可以将激活函数曲线中曲线斜率变化较多的部分划分较多的分段，将曲线斜率变化较少的部分划分较少的分段，这样可以实现用较少的配置参数实现高精度的曲线拟合。

[0054] 可选的，上述将第一阈值与第二阈值之间的输入数据进行非均匀分段，计算各个分段内的输入数据对应的配置参数，可以通过以下步骤实现：首先，采用逐步逼近的方法将第一阈值与第二阈值之间的输入数据划分为至少一个非均匀分段；然后，根据各个非均匀分段内至少两点的输入数据对应的激活函数的输出值计算各个非均匀分段内的输入数据对应的配置参数。

[0055] 可选的，上述采用逐步逼近的方法将第一阈值与第二阈值之间的输入数据划分为至少一个非均匀分段，可以通过以下步骤实现：首先，将第一阈值与第二阈值之间的输入数据均匀划分为设定个区间，并将每个区间均匀划分为设定个子区间，其中，划分区间和子区间的多少可以根据计算精度的需要以及硬件设备的存储能力进行设定；然后，连接各区间中相邻的子区间的1/2抽样值，得到至少一条直线，判断各条直线与相应位置的激活函数的输出值曲线的误差是否超过预先设定的误差阈值；如果没有超过，则将得到的各条直线中

的两两相邻的直线进行合并，得到一条直线，判断该直线与相应位置激活函数的输出值曲线的误差是否超过所述误差阈值；依次类推，得到各个区间内的至少一条直线，每条直线对应的取值范围为输入数据的一个分段。

[0056] 可选的，步骤302根据预处理的结果以及通过线性拟合激活函数得到的结果获得待处理输入数据的配置参数的存储地址，可以通过以下方式实现：将待处理输入数据预处理后的结果与第一阈值和第二阈值进行比较；如果待处理输入数据预处理后的结果小于第一阈值，则得到小于第一阈值的输入数据对应的配置参数的存储地址；如果待处理输入数据预处理后的结果大于第二阈值，则得到大于第二阈值的输入数据对应的配置参数的存储地址；如果待处理输入数据预处理后的结果在第一阈值和所述第二阈值之间，则查找待处理输入数据预处理后的结果所在区间的分段范围，并得到该分段内的输入数据对应的配置参数的存储地址。例如，当激活函数的输入数据x在第一阈值和第一阈值之间，配置参数的存储地址address可以通过以下计算方法得到：address＝n*(x-i_{base})/(i_{max}-i_{base})+m，m表示i区间之前所有区间的总段数，可以通过累加的方式计算出来，由于片内SRAM存储空间的限制m应当控制在合理的范围内(例如小于4096)，i_{base}和i_{max}分别表示第一阈值和第二阈值之间均匀划分的设定个区间中的第i个区间的x坐标最小值和最大值，i_{base}和i_{max}可以用x和寄存器堆里面的区间信息逐一比较可以得到，n表示第i个区间划分的段的数量(寄存器堆里面有该信息)，其中，m可以用于表示第i个区间在片内SRAM中的基地址，n*(x-i_{base})/(i_{max}-i_{base})可以用于计算x在第i个区间内的偏移地址。在这里，上述比较的过程可以通过软件的方式实现，也可以通过硬件的方式(例如使用比较器)实现。例如，如图6所示，其示出了一种通过使用比较器获取配置参数的存储地址的一种实现方式，通过线性拟合激活函数可以得到分段相关的数据(例如各个分段的起始坐标)和配置参数，其中与分段相关的数据可以存储到寄存器中，输入数据x通过浮点转定点电路601转化为定点数之后，比较器602通过将输入数据x与寄存器中存储的数据进行比较从而得出输入数据x所在的段，如果小于第一阈值则可以直接得到相对应的配置参数A_0和B_0，如果大于第二阈值则可以直接得到相对应的配置参数A_1和B_1，如果在第一阈值和所述第二阈值之间则可以根据上述用于计算存储地址address的公式address＝n*(x-i_{base})/(i_{max}-i_{base})+m计算配置参数的存储地址。本申请的上述实施例提供的方法根据激活函数的曲线特性分段拟合激活函数得到各段的配置参数，并根据待处理输入数据所在的分段获取待处理输入数据的配置参数，根据获取的该配置参数以及预先设定的电路结构得到待处理输入数据的处理结果，整个处理过程不需要使用用于实现激活函数的专用电路，从而简化了电路结构，同时上述实施例提供的方法可以适用于多种激活函数，因此提高了数据处理的灵活性。

[0057] 进一步参考图7，作为对上述各图所示方法的实现，本申请提供了一种数据处理装置的一个实施例，该装置实施例与图3所示的方法实施例相对应，该装置具体可以应用于各种电子设备中。

[0058] 如图7所示，本实施例所述的数据处理装置700包括：预处理单元701、存储地址获取单元702、配置参数获取单元703和处理单元704。其中，预处理单元701用于对接收到的待处理输入数据进行预处理；存储地址获取单元702用于根据预处理的结果以及通过线性拟合激活函数得到的结果获得上述待处理输入数据的配置参数的存储地址，其中，配置参数是根据激活函数的曲线特性预先设置的；配置参数获取单元703用于根据上述存储地址获

取上述待处理输入数据的配置参数；处理单元704用于根据上述待处理输入数据的配置参数以及预先设定的电路结构对上述待处理输入数据的预处理结果进行处理，得到处理结果。

[0059] 在本实施例中，预处理单元701、存储地址获取单元702、配置参数获取单元703和处理单元704的具体处理可以参考图3对应实施例步骤301、步骤302、步骤303和步骤304的详细描述，在此不再赘述。

[0060] 在本实施例的一些可选的实现方式中，上述待处理输入数据为浮点数；以及上述预处理单元701进一步用于：将上述待处理输入数据转化为定点数。该实现方式可参考上述图3对应实施例中相应实现方式的详细描述，在此不再赘述。

[0061] 在本实施例的一些可选的实现方式中，上述装置还包括参数配置单元(未示出)，上述参数配置单元包括阈值获取单元(未示出)、计算单元(未示出)和存储单元(未示出)；以及上述阈值获取单元，用于根据激活函数曲线的斜率变化得到激活函数输入数据的第一阈值和第二阈值，其中，小于第一阈值或者大于第二阈值的输入数据对应的激活函数曲线的斜率变化小于预先设定的斜率变化阈值；上述计算单元，用于分别计算小于第一阈值、大于第二阈值、以及第一阈值与第二阈值之间的输入数据对应的配置参数；上述存储单元，用于将输入数据与配置参数关联存储。该实现方式可参考上述图3对应实施例中相应实现方式的详细描述，在此不再赘述。

[0062] 在本实施例的一些可选的实现方式中，上述计算单元包括第一计算子单元(未示出)、第二计算子单元(未示出)和第三计算子单元(未示出)；以及上述第一计算子单元，用于根据至少两个小于第一阈值的输入数据对应的激活函数的输出值计算小于第一阈值的输入数据对应的配置参数；上述第二计算子单元，用于根据至少两个大于第二阈值的输入数据对应的激活函数的输出值计算大于第二阈值的输入数据对应的配置参数；上述第三计算子单元，用于将第一阈值与第二阈值之间的输入数据进行非均匀分段，计算各个分段内的输入数据对应的配置参数。该实现方式可参考上述图3对应实施例中相应实现方式的详细描述，在此不再赘述。

[0063] 在本实施例的一些可选的实现方式中，上述第三计算子单元包括分段单元(未示出)和段内配置参数计算单元(未示出)；以及上述分段单元，用于采用逐步逼近的方法将第一阈值与第二阈值之间的输入数据划分为至少一个非均匀分段；上述段内配置参数计算单元，包括根据各个非均匀分段内至少两点的输入数据对应的激活函数的输出值计算各个非均匀分段内的输入数据对应的配置参数。该实现方式可参考上述图3对应实施例中相应实现方式的详细描述，在此不再赘述。

[0064] 在本实施例的一些可选的实现方式中，上述分段单元进一步用于：将第一阈值与第二阈值之间的输入数据均匀划分为设定个区间，并将每个区间均匀划分为设定个子区间；连接各区间中相邻的子区间的1/2抽样值，得到至少一条直线，判断各条直线与相应位置的激活函数的输出值曲线的误差是否超过预先设定的误差阈值；如果没有超过，则将得到的各条直线中的两两相邻的直线进行合并，得到一条直线，判断该直线与相应位置激活函数的输出值曲线的误差是否超过上述误差阈值；依次类推，得到各个区间内的至少一条直线，每条直线对应的取值范围为输入数据的一个分段。该实现方式可参考上述图3对应实施例中相应实现方式的详细描述，在此不再赘述。

[0065] 在本实施例的一些可选的实现方式中，上述存储地址获取单元702进一步用于：将上述待处理输入数据预处理后的结果与第一阈值和第二阈值进行比较；如果上述待处理输入数据预处理后的结果小于第一阈值，则得到小于第一阈值的输入数据对应的配置参数的存储地址；如果上述待处理输入数据预处理后的结果大于第二阈值，则得到大于第二阈值的输入数据对应的配置参数的存储地址；如果上述待处理输入数据预处理后的结果在上述第一阈值和上述第二阈值之间，则查找上述待处理输入数据预处理后的结果所在区间的分段范围，并得到该分段内的输入数据对应的配置参数的存储地址。该实现方式可参考上述图3对应实施例中相应实现方式的详细描述，在此不再赘述。

[0066] 本发明实施例还提供了一种人工智能处理器，该人工智能处理器可以包括图7对应实施例中所描述的数据处理装置，该人工智能处理器包括两组或两组以上的寄存器堆和静态随机存取存储器，每组寄存器堆和静态随机存取存储器用于存储一种激活函数的配置参数，且每组寄存器堆和静态随机存取存储器中存储的激活函数的配置参数是动态更新的。

[0067] 以上描述仅为本申请的较佳实施例以及对所运用技术原理的说明。本领域技术人员应当理解，本申请中所涉及的发明范围，并不限于上述技术特征的特定组合而成的技术方案，同时也应涵盖在不脱离所述发明构思的情况下，由上述技术特征或其等同特征进行任意组合而形成的其它技术方案。例如上述特征与本申请中公开的(但不限于)具有类似功能的技术特征进行互相替换而形成的技术方案。

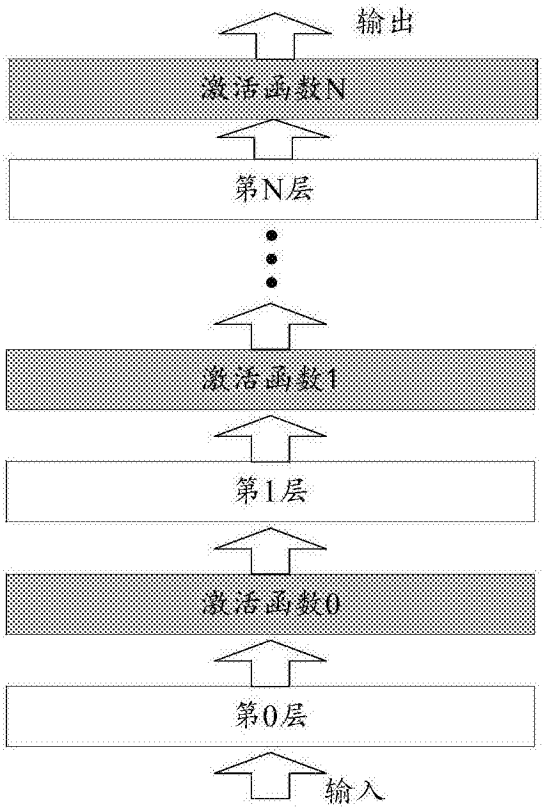

输出
激活函数N
第N层
激活函数1
第1层
激活函数0
第0层
输入

图1

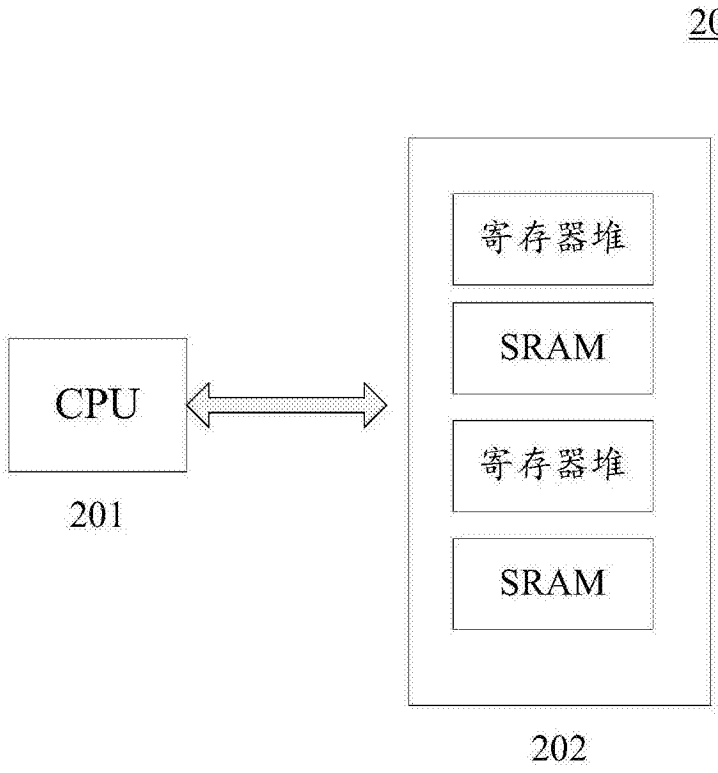

200
寄存器堆
SRAM
寄存器堆
SRAM
CPU
201
202

图2

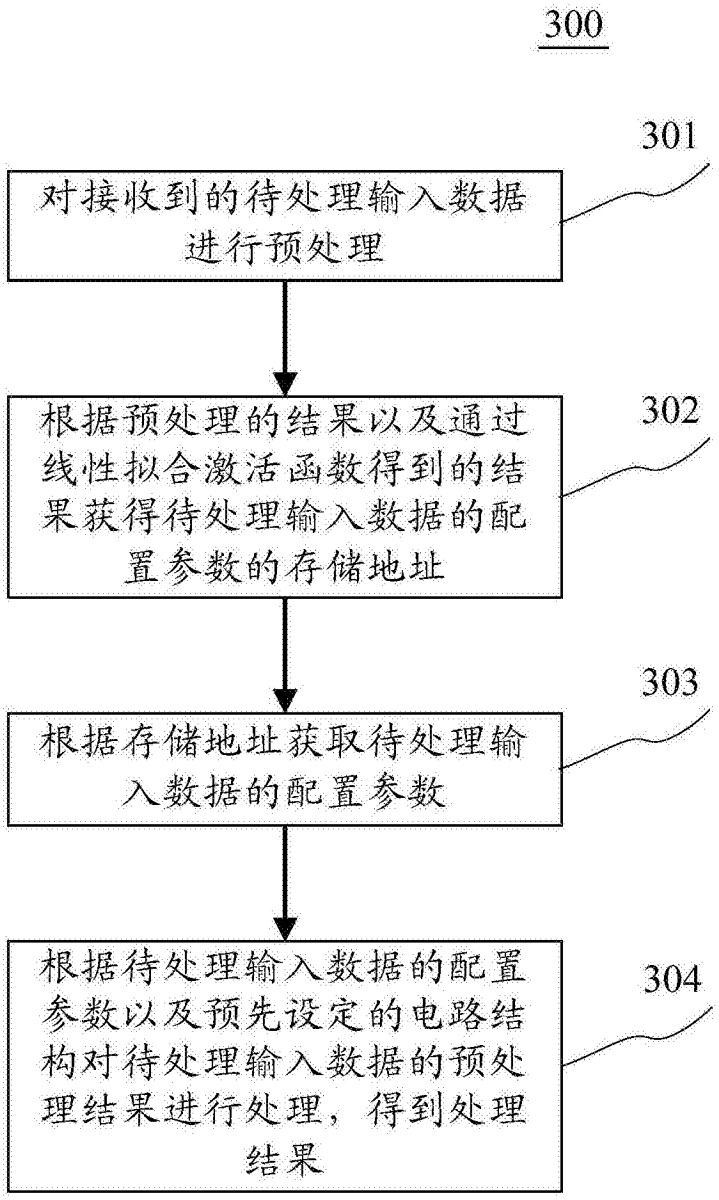

300
301
对接收到的待处理输入数据进行预处理
302
根据预处理的结果以及通过线性拟合激活函数得到的结果获得待处理输入数据的配置参数的存储地址
303
根据存储地址获取待处理输入数据的配置参数
304
根据待处理输入数据的配置参数以及预先设定的电路结构对待处理输入数据的预处理结果进行处理，得到处理结果

图3

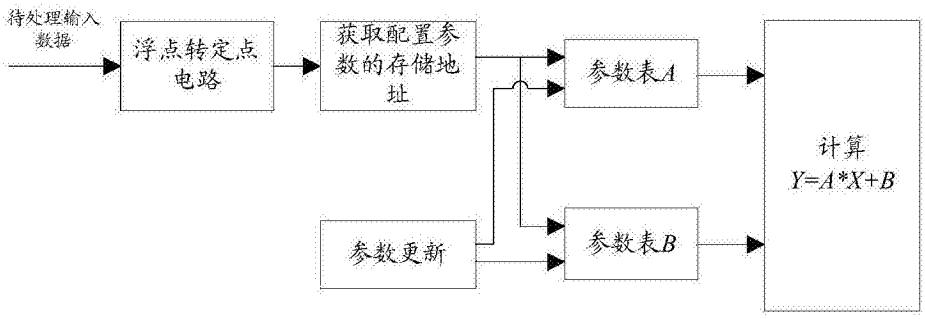

待处理输入数据
浮点转定点电路
获取配置参数的存储地址
参数表A
参数表B
参数更新
计算
Y=A*X+B

图4

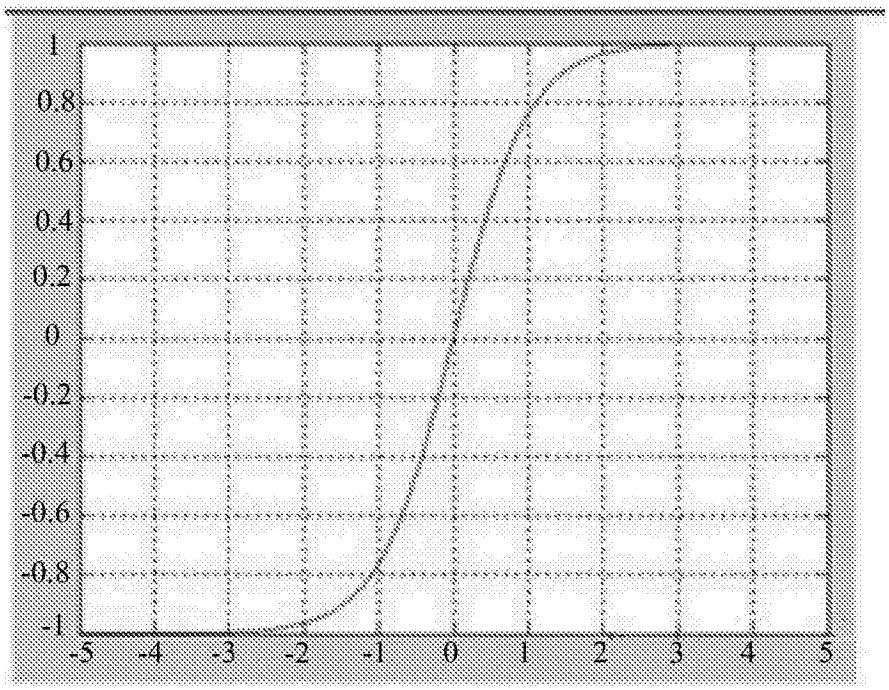

1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-5
-4
-3
-2
-1
0
1
2
3
4
5

图5

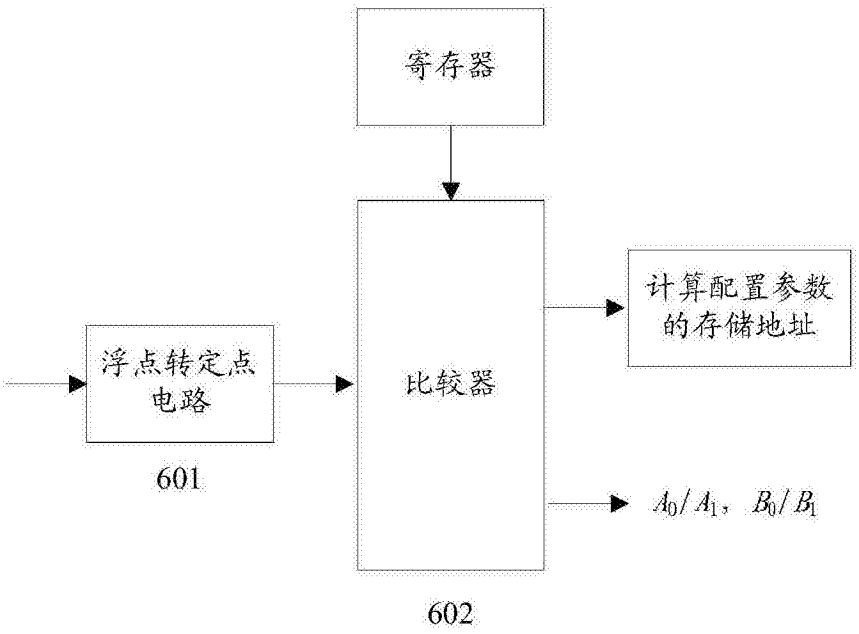

寄存器
浮点转定点电路
601
比较器
602
计算配置参数的存储地址
A_0/A_1，B_0/B_1

图6

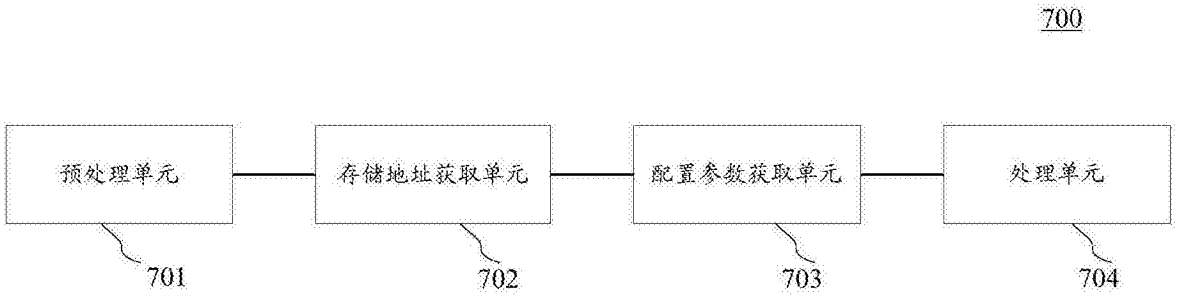


图7