(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 112200310 B
(45)授权公告日 2023.11.24

(21)申请号 202010883908.7

(22)申请日 2020.08.28

(65)同一申请的已公布的文献号
申请公布号 CN 112200310 A

(43)申请公布日 2021.01.08

(73)专利权人 星宸科技股份有限公司
地址 361005 福建省厦门市火炬高新区软件园创新大厦A区1501

(72)发明人 邓亚明

(74)专利代理机构 深圳紫藤知识产权代理有限公司 44570
专利代理师 远明

(51)Int.Cl.
*G06N 3/063*(2023.01)
*G06N 3/0464*(2023.01)
*G06N 3/048*(2023.01)
*G06N 20/00*(2019.01)

(56)对比文件
CN 108710530 A,2018.10.26
JP 2010231645 A,2010.10.14
潘曦等.《数字系统与微处理器》.北京理工大学出版社,2018,第137-140页.
翁钧朋.“三维CT图像中肺结节自动检测预处理算法的并行化研究”.《中国优秀硕士学位论文全文数据库 医药卫生科技辑》.2020,正文第3.1.2节.

审查员 陈佳怡

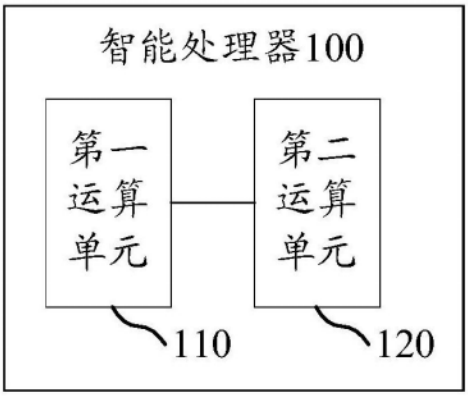
权利要求书2页 说明书8页 附图3页

(54)发明名称

智能处理器、数据处理方法及存储介质

(57)摘要

本申请实施例公开了一种智能处理器、数据处理方法及存储介质,其中,智能处理器包括一第一运算单元和一第二运算单元,第一运算单元用于获取对应一第一算子的一第一输入数据,并将第一输入数据划分为多个第一子输入数据,以及运行第一算子对每一第一子输入数据进行运算,得到对应每一第一子输入数据的多个第一子输出数据;第二运算单元用于将每一子输出数据作为第二算子的一第二输入数据,以及运行第二算子对每一第二输入数据进行运算,得到对应每一第二输入数据的一第二输出数据。这样,第二算子无需等待第一算子完成对第一输入数据的全部运算,可以使得相邻的两个算子在一定程度上实现部分并行计算,达到提高运算效率的目的。




CN 112200310 B

1.一种智能处理器,其特征在于,应用于卷积网络模型,包括:

一第一运算单元,用于获取对应一第一算子的一第一输入数据,并将所述第一输入数据划分为多个第一子输入数据,及运行所述第一算子对每一所述第一子输入数据进行运算,得到对应每一所述第一子输入数据的一第一子输出数据;以及

一第二运算单元,用于将每一所述第一子输出数据作为一第二算子的一第二输入数据,及运行所述第二算子对每一所述第二输入数据进行运算,得到对应每一所述第二输入数据的一第二输出数据;

其中,所述第二算子与所述第一算子属于同一网络模型且相邻,所述第一输入数据为具有高、宽及通道维度的数据,而将所述第一输入数据划分为所述多个第一子输入数据的操作为在所述第一输入数据的高、宽维度上进行划分,使所述第一输入数据与所述多个第一子输入数据具有相同的通道维度。

2.根据权利要求1所述的智能处理器,其特征在于,所述第二运算单元用于在所述第一运算单元每次得到所述第一子输出数据时,将所述第一子输出数据作为所述第二输入数据。

3.根据权利要求1所述的智能处理器,其特征在于,所述第二运算单元用于在得到的第一子输出数据的数据量达到一预设数据量时,将已得到的第一子输出数据作为所述第二输入数据。

4.根据权利要求1所述的智能处理器,其特征在于,所述智能处理器还包括一缓存单元,所述第一运算单元还用于将所述第一子输出数据写入所述缓存单元,而所述第二运算单元还用于从所述缓存单元读取所述第一子输出数据。

5.根据权利要求4所述的智能处理器,其特征在于,所述第一运算单元还用于:

根据所述第一算子的类型信息确定对应所述第一输入数据的一目标划分策略;以及

按照所述目标划分策略将所述第一输入数据划分为所述多个第一子输入数据。

6.根据权利要求5所述的智能处理器,其特征在于:

当所述目标划分策略为一第一划分策略时,根据所述第一运算单元的处理能力以及所述缓存单元的缓存空间,确定划分的多个第一子输入数据的一第一目标数据大小。

7.根据权利要求5所述的智能处理器,其特征在于:

当所述目标划分策略为一第二划分策略时,根据所述第一算子的运算逻辑,确定划分的多个第一子输入数据的一第二目标数据大小。

8.根据权利要求1所述的智能处理器,其特征在于,所述智能处理器包含于一芯片中。

9.一种数据处理方法,应用于一智能处理器,所述智能处理器包括一第一运算单元和一第二运算单元,其特征在于,所述数据处理方法包括:

所述第一运算单元获取对应一第一算子的一第一输入数据,并将所述第一输入数据划分为多个第一子输入数据;

所述第一运算单元运行所述第一算子对每一所述第一子输入数据进行运算,得到对应每一所述第一子输入数据的一第一子输出数据;

所述第二运算单元将每一所述第一子输出数据作为一第二算子的一第二输入数据;以及

所述第二运算单元运行所述第二算子对每一所述第二输入数据进行运算,得到对应每

一所述第二输入数据的一第二输出数据；

其中，所述第二算子与所述第一算子属于同一网络模型且相邻，所述第一输入数据为具有高、宽及通道维度的数据，而将所述第一输入数据划分为所述多个第一子输入数据的操作为在所述第一输入数据的高、宽维度上进行划分，使所述第一输入数据与所述多个第一子输入数据具有相同的通道维度。

10.根据权利要求9所述的数据处理方法，其特征在于，所述第二运算单元在所述第一运算单元每次得到所述第一子输出数据时，将所述第一子输出数据作为所述第二输入数据。

11.根据权利要求9所述的数据处理方法，其特征在于，所述第二运算单元在得到的第一子输出数据的数据量达到一预设数据量时，将已得到的第一子输出数据作为所述第二输入数据。

12.一种存储介质，其上存储有一计算机程序，所述计算机程序用以执行一数据处理方法，所述数据处理方法应用于一智能处理器，所述智能处理器包括一第一运算单元和一第二运算单元，其特征在于，所述数据处理方法包括：

所述第一运算单元获取对应一第一算子的一第一输入数据，并将所述第一输入数据划分为多个第一子输入数据；

所述第一运算单元运行所述第一算子对每一所述第一子输入数据进行运算，得到对应每一所述第一子输入数据的一第一子输出数据；

所述第二运算单元将每一所述第一子输出数据作为一第二算子的一第二输入数据；以及

所述第二运算单元运行所述第二算子对每一所述第二输入数据进行运算，得到对应每一所述第二输入数据的一第二输出数据；

其中，所述第二算子与所述第一算子属于同一网络模型且相邻，所述第一输入数据为具有高、宽及通道维度的数据，而将所述第一输入数据划分为所述多个第一子输入数据的操作为在所述第一输入数据的高、宽维度上进行划分，使所述第一输入数据与所述多个第一子输入数据具有相同的通道维度。

# 智能处理器、数据处理方法及存储介质

## 技术领域

[0001] 本申请涉及人工智能技术领域，具体涉及一种智能处理器、数据处理方法及存储介质。

## 背景技术

[0002] 人工智能(Artificial Intelligence,AI)是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能，感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。换句话说，人工智能是计算机科学的一个综合技术，它企图了解智能的实质，并生产出一种新的能以人类智能相似的方式做出反应的智能机器。人工智能也就是研究各种智能机器的设计原理与实现方法，使机器具有感知、运算与决策的功能。

[0003] 人工智能技术是一门综合学科，涉及领域广泛，既有硬件层面的技术也有软件层面的技术。人工智能基础技术一般包括如传感器、专用人工智能芯片、云计算、分布式存储、大数据处理技术、操作/交互系统、机电一体化等技术。人工智能软件技术主要包括计算机视觉技术、语音处理技术、自然语言处理技术以及机器学习/深度学习等几大方向。

[0004] 其中，机器学习(Machine Learning,ML)是一门多领域交叉学科，涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。专门研究计算机怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。机器学习是人工智能的核心，是使计算机具有智能的根本途径，其应用遍及人工智能的各个领域。机器学习和深度学习通常包括人工神经网络、置信网络、强化学习、迁移学习、归纳学习等技术。利用机器学习技术，以及对应的训练数据集，能够训练得到实现不同功能的网络模型，比如，基于一训练数据集能够训练得到用于性别分类的网络模型，基于另一训练数据集能够训练得到图像优化的网络模型。

[0005] 目前，随着人工智能技术的不断发展，网络模型被部署在如智能手机、平板电脑等电子设备上，用于增强电子设备的处理能力。比如，电子设备通过其部署的图像优化模型，可以对其拍摄的图像进行优化，提升图像质量。

[0006] 相关技术中，网络模型中相邻的两个算子串行运算，在网络模型的运算逻辑上，前一个算子运算完成的输出结果作为后一个算子的输入数据，后一个算子依赖于前一个算子的输出。比如，请参照图1，示出了某卷积网络模型中相邻的卷积算子和加法算子，在卷积网络模型的运算逻辑上，卷积算子的卷积结果作为加法算子的输入数据，加法算子依赖卷积算子的输出结果。实际运算中，加法算子需要等待卷积算子完全运算完毕，才能够根据卷积算子的卷积结果进行加法运算。如图2所示，加法算子需要等待卷积算子完成对高为H、宽为W、C通道的卷积输入数据的卷积运算，得到高为H、宽为W、C’通道的卷积结果后，再将高为H、宽为W、C’通道的卷积结果作为加法输入数据进行加法运算，得到高为H、宽为W、C’通道的加法结果。可以看出，这种相邻算子间的强行等待，将严重拖慢网络模型的运算效率。

[0007] 从拓扑结构上看，网络模型由多种不同类型的算子组成，相邻的算子串行运算，即后一个算子运算依赖于前一个算子的运算结果，这种串行结构会造成强行等待，不利于网

络模型运算效率的提高。基于此，本申请提供一种智能处理器、芯片、电子设备、数据处理方法及数据处理装置，以提高网络模型的运算效率。

## 发明内容

[0008] 本申请提供了一种智能处理器、数据处理方法及存储介质，能够提高网络模型的运算效率。

[0009] 本申请提供一种智能处理器，包括一第一运算单元及一第二运算单元。所述第一运算单元，用于获取对应一第一算子的一第一输入数据，并将所述第一输入数据划分为多个第一子输入数据，及运行所述第一算子对每一所述第一子输入数据进行运算，得到对应每一所述第一子输入数据的一第一子输出数据。所述第二运算单元，用于将每一所述第一子输出数据作为一第二算子的一第二输入数据，及运行所述第二算子对每一所述第二输入数据进行运算，得到对应每一所述第二输入数据的一第二输出数据。

[0010] 本申请提供一种数据处理方法，应用于一智能处理器，所述智能处理器包括一第一运算单元和一第二运算单元，所述数据处理方法包括：所述第一运算单元获取对应一第一算子的一第一输入数据，并将所述第一输入数据划分为多个第一子输入数据；所述第一运算单元运行所述第一算子对每一所述第一子输入数据进行运算，得到对应每一所述第一子输入数据的一第一子输出数据；所述第二运算单元将每一所述第一子输出数据作为一第二算子的一第二输入数据；以及，所述第二运算单元运行所述第二算子对每一所述第二输入数据进行运算，得到对应每一所述第二输入数据的一第二输出数据。

[0011] 本申请提供一种存储介质，其上存储有一计算机程序，所述计算机程序用以执行一数据处理方法，所述数据处理方法应用于一智能处理器，所述智能处理器包括一第一运算单元和一第二运算单元，其特征在于，所述数据处理方法包括：所述第一运算单元获取对应一第一算子的一第一输入数据，并将所述第一输入数据划分为多个第一子输入数据；所述第一运算单元运行所述第一算子对每一所述第一子输入数据进行运算，得到对应每一所述第一子输入数据的一第一子输出数据；所述第二运算单元将每一所述第一子输出数据作为一第二算子的一第二输入数据；以及，所述第二运算单元运行所述第二算子对每一所述第二输入数据进行运算，得到对应每一所述第二输入数据的一第二输出数据。

[0012] 前述的智能处理器、数据处理方法及存储介质中，第二算子无需等待第一算子完成对第一输入数据的全部运算，即可在不同的小块输入数据上实现与第一算子的并行计算。由此，利用不同的运算单元，可以使得相邻的两个算子在一定程度上实现部分并行计算，达到提高网络模型运算效率的目的。

## 附图说明

[0013] 为了更清楚地说明本申请实施例中的技术方案，下面将对实施例描述中所需要使用的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本申请的一些实施例，对于本领域技术人员来讲，在不付出创造性劳动的前提下，还可以根据这些附图获得其他的附图。

[0014] 图1为相关技术中相邻两个算子示意图。

[0015] 图2为图1中在卷积网络模型的运算逻辑上相邻两个算子进行串行运算的示意图。

[0016] 图3是本申请实施例中提供的智能处理器100的一结构示意图。

[0017] 图4是本申请实施例中将第一输入数据拆分为N个子输入数据的示意图。

[0018] 图5是本申请实施例中提供的智能处理器100的另一结构示意图。

[0019] 图6是本申请实施例提供的数据处理方法的流程示意图。

[0020] 图7是本申请实施例提供的数据处理装置300的结构示意图。

## 具体实施方式

[0021] 应当说明的是,本申请的原理是以实施在一适当的运算环境中来举例说明。以下的说明是通过所例示的本申请具体实施例,其不应被视为限制本申请未在此详述的其它具体实施例。

[0022] 本申请提供一种智能处理器、芯片、电子设备、数据处理方法及数据处理装置,以提高网络模型的运算效率。

[0023] 本申请实施例提供的方案涉及人工智能的机器学习技术,具体涉及网络模型的运算阶段,通过如下实施例进行说明:

[0024] 请参照图3,图3为本申请实施例提供的智能处理器100的第一种结构示意图。该智能处理器100可包括相互连接的一第一运算单元110和一第二运算单元120。

[0025] 应当说明的是,第一运算单元110和第二运算单元120为不同类型的硬件运算单元(或称硬件引擎,运算加速单元等),比如,在第一运算单元110为卷积运算单元时,第二运算单元120为向量运算单元;又比如,在第一运算单元110为向量运算单元时,第二运算单元120为卷积运算单元。

[0026] 其中,卷积运算单元也称卷积计算引擎,其内部可包括多个乘加单元(Multiplication Add Cell,MAC),该乘加单元的个数可以为几千个,比如卷积运算单元内部可包括4096个乘加单元,这4096个乘加单元可以进一步分成16个cell,每个cell可以进行最大元素数为256向量内积操作。

[0027] 向量运算单元也称单指令多数据(Single Instruction Multiple Data,SIMD)处理单元。向量运算单元是一个元素级向量计算引擎,可以处理常规的向量间的加减乘除等算术运算,同时也可以处理比特级的与、或、非、异或等逻辑运算。此外,向量运算单元支持线性整流函数(Rectified Linear Unit,ReLU)、PReLU等常见的激活函数操作。

[0028] 本申请实施例中,第一运算单元110用于获取对应一第一算子的一输入数据,记为第一输入数据,即第一算子进行运算所需的数据。

[0029] 其中,第一算子可以为第一运算单元110所能够运算的任意类型算子,比如,在第一运算单元110为卷积运算单元时,第一算子可以为卷积算子,又比如,在第一运算单元110为向量运算单元时,第一算子可以为加法算子、减法算子、乘法算子、除法算子或者激活算子等。

[0030] 在获取到对应第一算子的第一输入数据之后,第一运算单元100进一步按照配置数据划分策略,将第一输入数据划分为多个第一子输入数据。

[0031] 应当说明的是,本申请实施例中对于数据划分策略的配置不做具体限制,可由本领域普通技术人员根据实际需要进行配置。

[0032] 比如,请参照图4,假设第一运算单元110为卷积运算单元,第一算子为卷积算子,

如图4所示，获取到高为H、宽为W、C通道的第一输入数据，第一运算单元110对该第一输入数据进行拆分，得到N个高为H'、宽为W'、C通道的第一子输入数据。

[0033] 在将对应第一算子的第一输入数据划分为多个第一子输入数据之后，第一运算单元110即运行第一算子对划分得到的每一第一子输入数据进行运算，相应得到对应每一第一子输入数据的一输出数据，记为第一子输出数据。

[0034] 应当说明的是，本申请实施例中对于第一运算单元110运行第一算子对各第一子输入数据的运算先后顺序不做具体限制，可由本领域普通技术人员根据实际需要进行配置。

[0035] 本申请实施例中，第二算子可以为第二运算单元120所能够运算的任意类型算子，比如，在第二运算单元120为卷积运算单元时，第二算子可以为卷积算子，又比如，在第二运算单元120为向量运算单元时，第二算子可以为加法算子、减法算子、乘法算子、除法算子或者激活算子等。应当说明的是，本申请实施例中的第一算子和第二算子属于同一网络模型且相邻。比如，第一运算单元110为卷积运算单元，第二运算单元120为向量运算单元，则第一算子可以为一卷积网络模型中的卷积算子，第二算子可以为同一卷积网络模型中与前述卷积算子相邻的加法算子。

[0036] 在第一运算单元110运行第一算子对第一子输入数据进行运算，得到对应第一子输入数据的第一子输出数据之后，第二运算单元120将第一运算单元110得到的第一子输出数据作为第二算子的一输入数据，记为第二输入数据。然后，第二运算单元120即运行第二算子对第二输入数据进行运算，相应得到对应第二输入数据的一第二输出数据。

[0037] 由上可知，本申请提供的智能处理器包括不同的运算单元，分别为一第一运算单元和一第二运算单元，其中，第一运算单元用于获取对应一第一算子的一第一输入数据，并将第一输入数据划分为多个第一子输入数据，以及运行第一算子对每一第一子输入数据进行运算，得到对应每一第一子输入数据的一第一子输出数据；第二运算单元用于将每一第一子输出数据作为第二算子的一第二输入数据，以及运行第二算子对每一第二输入数据进行运算，得到对应每一第二输入数据的一第二输出数据。这样，第二算子无需等待第一算子完成对第一输入数据运算，即可在不同的小块输入数据上实现与第一算子的并行计算。由此，利用不同的运算单元，可以使得相邻的两个算子在一定程度上实现部分并行计算，达到提高网络模型运算效率的目的。

[0038] 可选地，在一实施例中，第二运算单元120用于在第一运算单元110每次得到第一子输出数据时，将第一运算单元110得到的第一子输出数据作为第二算子的第二输入数据。

[0039] 其中，在第一运算单元110每次得到第一子输出数据时，第二运算单元120即将第一运算单元110得到的第一子输出数据作为第二算子的第二输入数据，并运行第二算子对第二输入数据进行运算，相应得到对应第二输入数据的第二输出数据。

[0040] 比如，假设第一运算单元110将第一输入数据拆分为5个第一子输入数据，分别为第一子输入数据A、第一子输入数据B、第一子输入数据C、第一子输入数据D以及第一子输入数据E。第一运算单元110运行第一算子先对第一子输入数据A进行运算，得到对应第一子输入数据A的第一子输出数据A'，此时，第二运算单元120即将第一子输出数据A'作为第二算子的第二输入数据，并运行第二算子对第一子输出数据A'进行运算，得到对应第一子输出数据A'(即第二输入数据)的第二输出数据。其中，在第二运算单元120运行第二算子对第一

子输出数据A'进行运算的同时，第一运算单元110可以继续运行第一算子对第一子输入数据B进行运算。这样，当第一运算单元110运行第一算子完成对第一子输入数据B的运算，并得到对应第一子输入数据B的第一子输出数据B'时，第二运算单元120即将第一子输出数据B'作为第二算子的第二输入数据，运行第二算子对第一子输出数据B'进行运算，得到对应第一子输出数据B'(即第二输入数据)的第二输出数据。

[0041] 如上，在第一运算单元110运行第一算子完成对第一子输入数据E的运算，并得到对应第一子输入数据E的第一子输出数据E'时，第二运算单元120即将第一子输出数据E'作为第二算子的第二输入数据，运行第二算子对第一子输出数据E'进行运算，得到对应第一子输出数据E'(即第二输入数据)的第二输出数据。

[0042] 可以看出，以上在第一运算单元110运行第一算子对第一输入数据进行分块运算的过程中，第二运算单元120运行第二算子对第一算子的分块运算结果进行运算，使得第一算子和第二算子在不同数据块上实现了并行计算，当第一算子完成对第一输入数据的运算时，第二算子随即完成后续运算，由此，使得网络模型整体的运算效率得以提高。

[0043] 可选地，在一实施例中，第二运算单元120用于在得自第一运算单元110的第一子输出数据的数据量达到预设数据量时，将已得到的第一子输出数据作为第二算子的第二输入数据。

[0044] 应当说明的是，算子在进行运算时，存在最小数据量的要求，即算子正常运算的前提是提供足够其运算所需数据量的数据。因此，本申请实施例设置预设数据量来约束第二算子进行运算的数据量。其中，本申请实施例中对预设数据量的取值不做具体限定，以预设数据量大于第二算子进行运算的最小数据量为约束，可由本领域普通技术人员根据第二算子的特性取经验值。

[0045] 本申请实施例中，第二运算单元120在第一运算单元110每次运算得到第一子输出数据时，并不直接将第一运算单元110运算得到的第一子输出数据作为第二算子的第二输入数据，而是先识别第一运算单元110已经得到的第一子输出数据的数据量是否达到预设数据量，若已达到，第二运算单元120将第一运算单元110已得到的第一子输出数据作为第二算子的第二输入数据，并进一步运行第二算子对第二输入数据进行运算，得到对应第二输入数据的第二输出数据；若未达到，则等待第一运算单元110下次运算得到第一子输出数据，再次进行识别，直至第一运算单元110已得到的第一子输出数据的数据量达到预设数据量。

[0046] 可选地，在一实施例中，请参照图5，图5为本申请实施例提供的智能处理器110的第二种结构示意图。该智能处理器110包括一第一运算单元110、一第二运算单元120以及一缓存单元130，其中，第一运算单元110和第二运算单元120可以相应参照图3中的第一运算单元110和第二运算单元120，此处不再赘述。

[0047] 本申请实施例中，第一运算单元110还用于将第一子输出数据写入缓存单元130，而第二运算单元120还用于从缓存单元130读取第一子输出数据。第一运算单元110和第二运算单元120并不直接进行数据交换，而是利用第三方器件缓存单元130实现数据交换。

[0048] 详细来说，第一运算单元110在每次运行第一算子对一第一子输入数据进行运算时，将得到的对应该第一子输入数据的第一子输出数据写入缓存单元130中进行缓存。而第二运算单元120在第一运算单元110每次运算得到第一子输出数据并写入缓存单元130时，

从缓存单元130中读取第一子输出数据，将其作为第二算子的第二输入数据，并运行第二算子对第二输入数据进行运算，得到对应第二输入数据的第二输出数据；或者，第二运算单元120在第一运算单元110每次运算得到第一子输出数据并写入缓存单元130时，识别缓存单元130中所缓存的第一运算单元110已得到的第一子输出数据的数据量是否达到预设数据量，若已达到，则说明缓存单元130中缓存的第一子输出数据足够第二运算单元120进行运算，此时，第二运算单元120从缓存单元130中读取其中缓存的第一子输出数据，作为第二算子的第二输入数据，并进一步运行第二算子对第二输入数据进行运算，得到对应第二输入数据的第二输出数据；若未达到，则等待第一运算单元110下次运算得到第一子输出数据，再次进行识别，直至第一运算单元110已得到的第一子输出数据的数据量达到预设数据量。

[0049] 可选地，在一实施例中，第一运算单元110还用于：根据第一算子的类型信息确定对应第一输入数据的一目标划分策略，并按照目标划分策略将第一输入数据划分为多个第一子输入数据，再写入缓存单元中130。

[0050] 应当说明的是，本申请实施例中预先针对不同类型的算子，设置有与之对应的划分策略。

[0051] 相应的，在将对应第一算子的第一输入数据划分为多个第一子输入数据时，第一运算单元首先获取到第一算子的类型信息，也即是识别第一算子为何种类型的算子；然后，根据第一算子的类型信息确定与之对应的划分策略，记为目标划分策略；然后，按照目标划分策略将第一输入数据划分为多个第一子输入数据，并写入缓存单元130中。由此，第一运算单元110可以进一步从缓存单元130中读取第一子输入数据，并运行第一算子对第一子输入数据进行运算，得到对应第一子输入数据的第一子输出数据。

[0052] 可选地，在一实施例中，第一运算单元110可用于：在确定的目标划分策略为一第一划分策略时，根据第一运算单元110的处理能力以及缓存单元的缓存空间，确定划分的多个第一子输入数据的一第一目标数据大小；并按照第一目标数据大小将第一输入数据划分为多个第一子输入数据，并写入缓存单元130中。

[0053] 举例来说，本申请实施例中将算子类型分为两类，其中，第一类型算子为运算只发生在向量分量之间的算子，其输入数据的结构与第二输出数据的结构相同，第一类型算子比如加法算子、减法算子、乘法算子、除法算子以及激活算子等。第二类型算子即非第一类型算子，其输入数据的结构与第二输出数据的结构不同，比如卷积算子、池化算子等。

[0054] 本申请实施例中，针对第一类型算子，根据其特性可知，对其输入数据的拆分，无需考虑算子本身在运算逻辑上的特征，只需考虑硬件资源的限制。

[0055] 相应的，针对第一类型算子，设置有一第一划分策略，其根据第一运算单元110的处理能力以及缓存单元的缓存空间，确定划分的多个第一子输入数据的一第一目标数据大小。

[0056] 详细来说，第一运算单元110根据第一运算单元110的处理能力确定出其能够运算的输入数据的最大数据量，并根据缓存单元130的缓存空间确定缓存单元130能存放的数据的最大数据量，然后，以缓存单元130能存放的最大数据量，第一运算单元110能够运算的最大数据量为约束，使得划分的第一子输入数据在不超过缓存单元130能存放的最大数据量的前提下，尽可能的达到第一运算单元110能够运算的最大数据量。

[0057] 比如，假设第一运算单元110能够运算的最大数据量为5，缓存单元能够存放的最

大数据量为3，则可以确定划分的第一子输入数据的第一目标数据大小为3；又比如，假设第一运算单元110能够运算的最大数据量为4，缓存单元能够存放的最大数据量为5，则可以确定划分的第一子输入数据的第一目标数据大小为4。

[0058] 可选地，在一实施例中，第一运算单元110还用于：当目标划分策略为一第二划分策略时，根据第一算子的运算逻辑，确定划分的多个第一子输入数据的一第二目标数据大小，及按照第二目标数据大小将第一输入数据划分为多个第一子输入数据，并写入缓存单元130中。

[0059] 本申请实施例中，针对第二类型算子，根据其特性可知，只需考虑算子本身在运算逻辑上的特征。

[0060] 比如，当第一算子为卷积算子时，根据卷积算子的运算逻辑，以卷积可运算第一子输出数据的数据大小来确定划分的第一子输入数据的数据大小。其中，假定第二输出数据的通道为C，以数据数据的宽度和高度能够被存入缓存单元130为约束，确定划分的第一子输入数据的第二目标数据大小。

[0061] 又比如，当第一算子为池化算子时，根据池化算子的运算逻辑，可以根据池化算子中滤波器(Filter)的尺寸和步长来确定划分的第一子输入数据的数据大小。其中，可以将滤波器看做是一个矩形窗口，其在输入数据按照步长进行滑动，被滤波器框住的数据执行池化运算，相应的，将滤波器框住的数据的数据大小确定为第二目标数据大小。

[0062] 本申请还提供一种芯片，其包括：一中央处理器及一智能处理器。其中此智能处理器可为本申请任一实施例所提供的智能处理器，用于从中央处理器获取运算所需的输入数据，并将运算得到的输出数据返回至中央处理器。

[0063] 本申请还提供一种数据处理方法，应用于本申请提供的智能处理器，请参照图6，该数据处理方法的流程如下：

[0064] 在210中，第一运算单元获取对应一第一算子的一第一输入数据，并将第一输入数据划分为多个第一子输入数据；

[0065] 在220中，第一运算单元运行第一算子对每一第一子输入数据进行运算，得到对应每一第一子输入数据的一第一子输出数据；

[0066] 在230中，第二运算单元将每一第一子输出数据作为一第二算子的一第二输入数据；

[0067] 在240中，第二运算单元运行第二算子对每一第二输入数据进行运算，得到对应每一第二输入数据的一第二输出数据。

[0068] 请参照图3，以本申请提供的数据处理方法适用于图3所示的智能处理器100为例，相关说明请参照前述关于图3的说明，在此不再赘述。

[0069] 可选地，在一实施例中，请参照图5，智能处理器100还包括一缓存单元130，本申请提供的数据处理方法还包括：第一运算单元110将第一子输出数据写入缓存单元130；以及，第二运算单元120从缓存单元读取130第一子输出数据。相关说明请参照前述关于图5的说明，在此不再赘述。

[0070] 可选地，本申请还提供一种数据处理装置，应用于本申请提供的智能处理器，此智能处理器包括一第一运算单元和一第二运算单元，请参照图7，数据处理装置300包括一数据获取模块310、一第一运算模块320、一输入设定模块330及一第二运算模块340。数据获取

模块310，用于通过第一运算单元获取对应一第一算子的一第一输入数据，并将第一输入数据划分为多个第一子输入数据。第一运算模块320，用于通过第一运算单元运行第一算子对每一第一子输入数据进行运算，得到对应每一第一子输入数据的一第一子输出数据。输入设定模块330，用于通过第二运算单元将每一第一子输出数据作为一第二算子的一第二输入数据。第二运算模块340，用于通过第二运算单元运行第二算子对每一第二输入数据进行运算，得到对应每一第二输入数据的一第二输出数据。

[0071] 可选地，在一实施例中，在将每一第一子输出数据作为一第二算子的一第二输入数据时，输入设定模块330用于通过第二运算单元在第一运算单元每次得到第一子输出数据时，将该第一子输出数据作为第二输入数据。

[0072] 可选地，在一实施例中，在将每一第一子输出数据作为一第二算子的一第二输入数据时，输入设定模块330通过第二运算单元在得自第一运算单元的第一子输出数据的数据量达到预设数据量时，将已得到的第一子输出数据作为第二输入数据。

[0073] 可选地，在一实施例中，智能处理器还包括一缓存单元，第一运算模块320还用于通过第一运算单元将第一子输出数据写入缓存单元，而输入设定模块330还用于通过第二运算单元从缓存单元读取第一子输出数据。

[0074] 需要说明的是，对本申请实施例的数据处理方法而言，本领域普通技术人员可以理解实现本申请实施例的数据处理方法的全部或部分流程，是可以通过计算机程序来控制相关的硬件来完成，该计算机程序可存储于一计算机可读取存储介质中，其在被包括一第一运算单元和一第二运算单元的一智能处理器加载时执行可包括如数据处理方法的实施例的流程。其中，存储介质可为磁碟、光盘、只读存储器、随机存取记忆体等。

[0075] 以上对本申请实施例提供的智能处理器、芯片、数据处理方法、数据处理装置及存储介质进行了详细介绍。本文中应用了具体个例对本申请的原理及实施方式进行了阐述，以上实施例的说明只是用于帮助理解本申请。同时，对于本领域的技术人员，依据本申请的思想，在具体实施方式及应用范围上均会有改变之处，综上所述，本说明书内容不应理解为对本申请的限制。


卷积算子
加法算子

图1

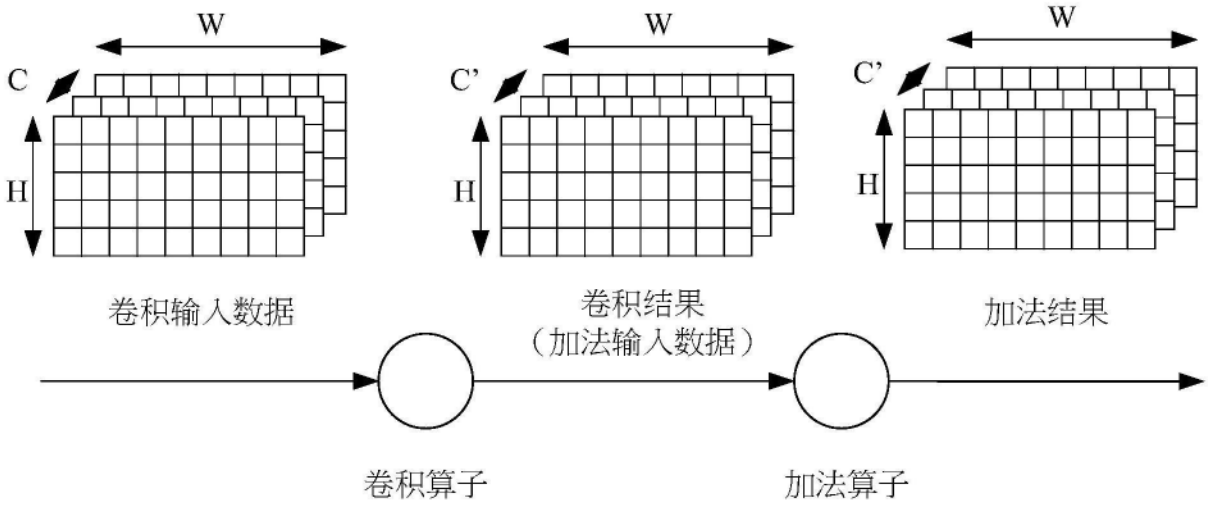

W
C
H
卷积输入数据
W
C'
H
卷积结果
（加法输入数据）
W
C'
H
加法结果
卷积算子
加法算子

图2

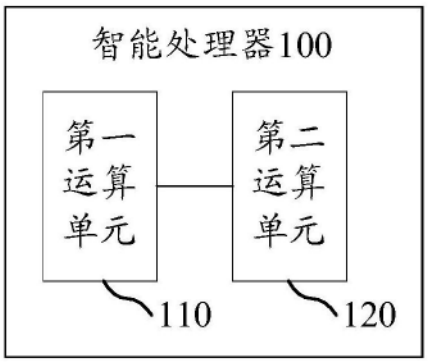

智能处理器100
第一运算单元
第二运算单元
110
120

图3

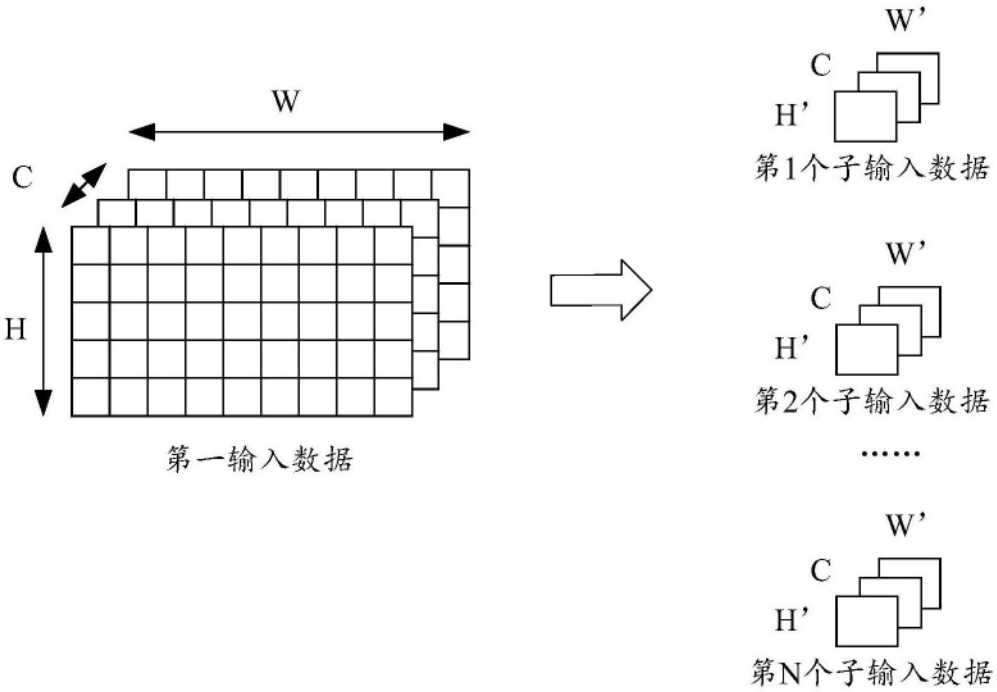

W
C
H
第一输入数据
W'
C
H'
第1个子输入数据
W'
C
H'
第2个子输入数据
.....
W'
C
H'
第N个子输入数据

图4

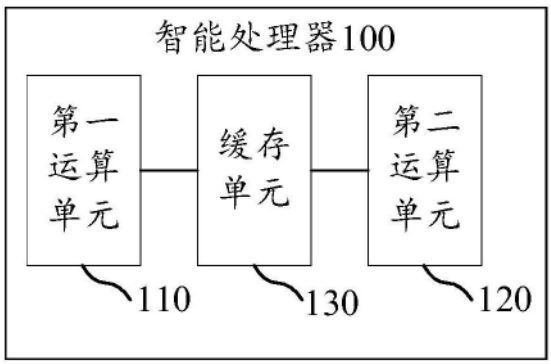

智能处理器100
第一运算单元
缓存单元
第二运算单元
110
130
120

图5

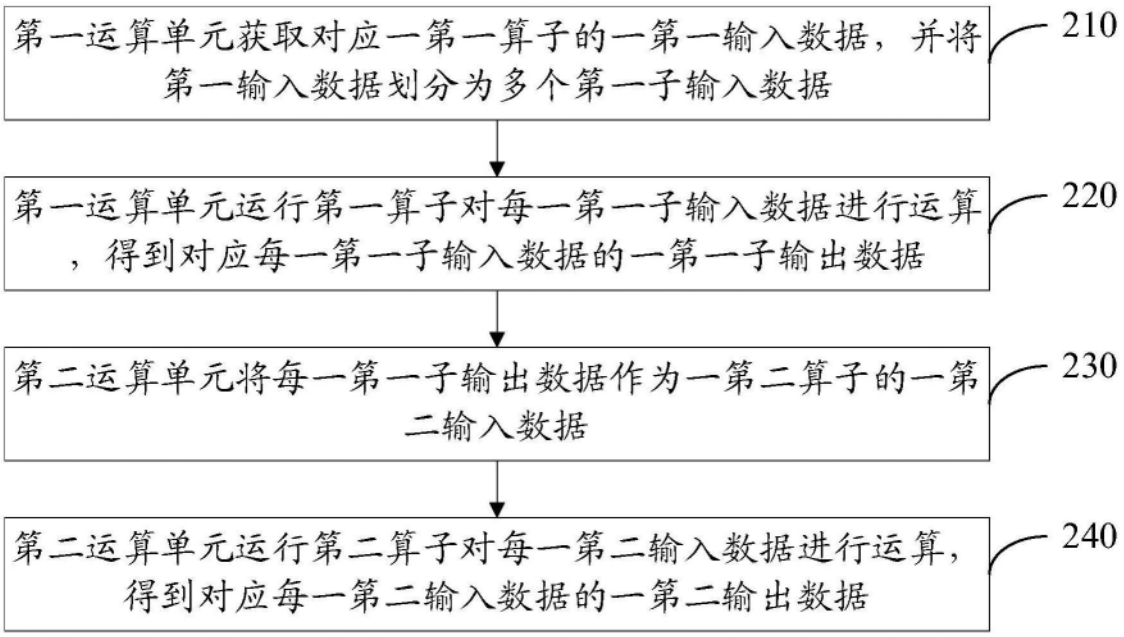

第一运算单元获取对应一第一算子的一第一输入数据，并将第一输入数据划分为多个第一子输入数据
210
第一运算单元运行第一算子对每一第一子输入数据进行运算，得到对应每一第一子输入数据的一第一子输出数据
220
第二运算单元将每一第一子输出数据作为一第二算子的一第二输入数据
230
第二运算单元运行第二算子对每一第二输入数据进行运算，得到对应每一第二输入数据的一第二输出数据
240

图6

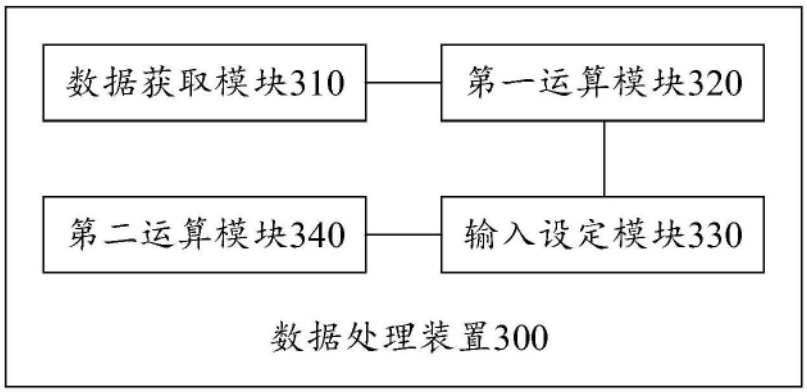

数据获取模块310
第一运算模块320
第二运算模块340
输入设定模块330
数据处理装置300

图7