(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 109543139 B
(45)授权公告日 2021.09.17

(21)申请号 201710866060.5
(22)申请日 2017.09.22
(65)同一申请的已公布的文献号
申请公布号 CN 109543139 A
(43)申请公布日 2019.03.29
(73)专利权人 杭州海康威视数字技术股份有限公司
地址 310051 浙江省杭州市滨江区阡陌路555号
(72)发明人 张渊
(74)专利代理机构 北京柏杉松知识产权代理事务所(普通合伙) 11413
代理人 马敬 项京
(51)Int.Cl.
G06F 17/15(2006.01)
G06N 3/063(2006.01)
G06N 3/04(2006.01)

审查员 吴姝泓

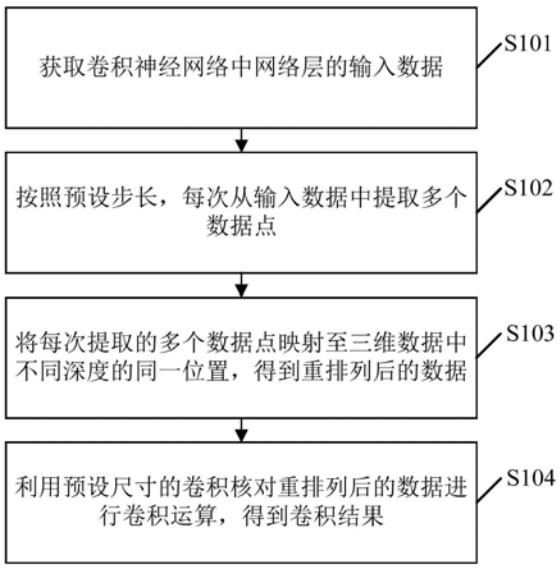
权利要求书1页 说明书10页 附图4页

(54)发明名称

卷积运算方法、装置、计算机设备及计算机可读存储介质

(57)摘要

本发明实施例提供了一种卷积运算方法、装置、计算机设备及计算机可读存储介质，其中，卷积运算方法包括：获取卷积神经网络中网络层的输入数据；按照预设步长，每次从输入数据中提取多个数据点；将每次提取的多个数据点映射至三维数据中不同深度的同一位置，得到重排列后的数据；利用预设尺寸的卷积核对重排列后的数据进行卷积运算，得到卷积结果。通过本发明可以提高卷积神经网络的运算效率。



CN 109543139 B

1.一种卷积运算方法,其特征在于,涉及深度学习技术领域,应用于具有图像处理功能的摄像机,所述方法包括:

获取卷积神经网络中网络层的输入图像数据;

将所述输入图像数据沿深度方向进行划分,得到多个切片;

针对各切片,每次按照预设步长,分别提取该切片中各深度的数据点,得到多个数据点;

将每次从各切片中提取的多个数据点映射至三维数据中不同深度的同一位置,分别得到各切片对应的待合并数据;

沿深度方向,将多个待合并数据进行排列,得到重排列后的数据;

利用预设尺寸的卷积核对所述重排列后的数据进行卷积运算,得到卷积结果。

2.一种卷积运算装置,其特征在于,涉及深度学习技术领域,应用于具有图像处理功能的摄像机,所述装置包括:

获取模块,用于获取卷积神经网络中网络层的输入图像数据;

划分模块,用于将所述输入图像数据沿深度方向进行划分,得到多个切片;

提取模块,用于针对各切片,每次按照预设步长,分别提取该切片中各深度的数据点,得到多个数据点;

映射模块,用于将每次从各切片中提取的多个数据点映射至三维数据中不同深度的同一位置,分别得到各切片对应的待合并数据;沿深度方向,将多个待合并数据进行排列,得到重排列后的数据;

运算模块,用于利用预设尺寸的卷积核对所述重排列后的数据进行卷积运算,得到卷积结果。

3.一种计算机设备,其特征在于,包括处理器、通信接口、存储器和通信总线,其中,所述处理器,所述通信接口,所述存储器通过所述通信总线完成相互间的通信;

所述存储器,用于存放计算机程序;

所述处理器,用于执行所述存储器上所存放的程序时,实现权利要求1所述的方法步骤。

4.一种计算机可读存储介质,其特征在于,所述计算机可读存储介质内存储有计算机程序,所述计算机程序被处理器执行时实现权利要求1所述的方法步骤。

卷积运算方法、装置、计算机设备及计算机可读存储介质

技术领域

[0001] 本发明涉及深度学习技术领域，特别是涉及一种卷积运算方法、装置、计算机设备及计算机可读存储介质。

背景技术

[0002] 在CNN(Convolutional Neural Network，卷积神经网络)中，对于每个网络层而言，由于输入数据的大小往往是不同的，因此，每个网络层进行卷积运算的卷积核也相应的设置为不同尺寸。然而，卷积核的尺寸大小直接影响到CNN对应硬件平台的设计，如果CNN中具有多种尺寸的卷积核，则需要设计复杂的硬件平台以支持CNN的运行，导致硬件资源的开销较大。

[0003] 针对上述问题，相应的卷积运算方法中，对于采用较大尺寸的卷积核的网络层，利用两个小尺寸的卷积核代替该较大尺寸的卷积核，对输入数据进行卷积运算，例如，利用两个3×3的卷积核代替一个5×5的卷积核，对输入数据进行卷积运算。但是，对于一个网络层而言，原本通过一个卷积核完成卷积运算，而该方法需要通过两个卷积核才可以完成卷积运算，该方法增加了卷积运算的运算量，影响卷积运算的运算效率。

发明内容

[0004] 本发明实施例的目的在于提供一种卷积运算方法、装置、计算机设备及计算机可读存储介质，以提高卷积神经网络的运算效率。具体技术方案如下：

[0005] 第一方面，本发明实施例提供了一种卷积运算方法，所述方法包括：

[0006] 获取卷积神经网络中网络层的输入数据；

[0007] 按照预设步长，每次从所述输入数据中提取多个数据点；

[0008] 将每次提取的多个数据点映射至三维数据中不同深度的同一位置，得到重排列后的数据；

[0009] 利用预设尺寸的卷积核对所述重排列后的数据进行卷积运算，得到卷积结果。

[0010] 第二方面，本发明实施例提供了一种卷积运算装置，所述装置包括：

[0011] 获取模块，用于获取卷积神经网络中网络层的输入数据；

[0012] 提取模块，用于按照预设步长，每次从所述输入数据中提取多个数据点；

[0013] 映射模块，用于将每次提取的多个数据点映射至三维数据中不同深度的同一位置，得到重排列后的数据；

[0014] 运算模块，用于利用预设尺寸的卷积核对所述重排列后的数据进行卷积运算，得到卷积结果。

[0015] 第三方面，本发明实施例提供了一种计算机设备，包括处理器、通信接口、存储器和通信总线，其中，所述处理器，所述通信接口，所述存储器通过所述通信总线完成相互间的通信；

[0016] 所述存储器，用于存放计算机程序；

[0017] 所述处理器，用于执行所述存储器上所存放的程序时，实现如第一方面所述的方法步骤。

[0018] 第四方面，本发明实施例提供了一种计算机可读存储介质，所述计算机可读存储介质内存储有计算机程序，所述计算机程序被处理器执行时实现如第一方面所述的方法步骤。

[0019] 本发明实施例提供的一种卷积运算方法、装置、计算机设备及计算机可读存储介质，通过按照预设步长，每次从获取的卷积神经网络中网络层的输入数据中提取多个数据点，并将每次提取的多个数据点映射至三维数据中不同深度的同一位置，得到重排列后的数据，最后利用预设尺寸的卷积核对重排列后的数据进行卷积运算，得到卷积结果。由于对网络层的输入数据进行多个数据点的提取及映射操作，将输入数据在深度方向进行扩展，并且减小了每个深度的尺寸，由于输入数据的尺寸变小，则可以利用更小的卷积核对该输入数据进行卷积运算，通过该方法，将各网络层的输入数据进行处理，得到的重排列后的数据均可以利用相同的预设尺寸的卷积核进行卷积运算，从而可以减小硬件资源的开销，并且，针对每个网络层，利用相同的更小尺寸的卷积核进行卷积运算，可以提高卷积神经网络的运算效率。

附图说明

[0020] 为了更清楚地说明本发明实施例或现有技术中的技术方案，下面将对实施例或现有技术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发明的一些实施例，对于本领域普通技术人员来讲，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。

[0021] 图1为本发明实施例的卷积运算方法的一种流程示意图；

[0022] 图2为本发明实施例的卷积运算方法的另一种流程示意图；

[0023] 图3为本发明实施例的输入数据重排列的示意图；

[0024] 图4为本发明实施例的卷积运算装置的一种结构示意图；

[0025] 图5为本发明实施例的卷积运算装置的另一种结构示意图；

[0026] 图6为本发明实施例的计算机设备的结构示意图。

具体实施方式

[0027] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0028] 为了提高卷积神经网络的运算效率，本发明实施例提供了一种卷积运算方法、装置、计算机设备及计算机可读存储介质。

[0029] 下面首先对本发明实施例所提供的一种卷积运算方法进行介绍。

[0030] 本发明实施例所提供的一种卷积运算方法的执行主体可以为一种执行卷积运算的计算机设备，例如，图像处理器、具有图像处理功能的摄像机等等。执行主体中至少包括具有数据处理能力的核心处理芯片，其中，核心处理芯片可以为DSP(Digital Signal

Processor,数字信号处理器)、ARM(Advanced Reduced Instruction Set Computer Machines,精简指令集计算机微处理器)、FPGA(Field-Programmable Gate Array,现场可编程门阵列)等核心处理芯片中的任一种。实现本发明实施例所提供的一种卷积运算方法的方式,可以为设置于执行主体中的软件、硬件电路和逻辑电路中的至少一种方式。

[0031] 如图1所示,本发明实施例所提供的一种卷积运算方法,可以包括如下步骤:

[0032] S101,获取卷积神经网络中网络层的输入数据。

[0033] 卷积神经网络中每个网络层的输入数据为一个三维数据,输入数据的大小可以表示为W×H×I,其中,I为输入数据的深度,W×H为每个深度的数据尺寸,即每个深度的数据的宽和高,由于卷积神经网络中每个网络层的输入数据的大小不同,尤其是每个深度的数据尺寸不同,为了提高运算速率,针对大尺寸的输入数据,可以选择大尺寸的卷积核进行卷积运算,针对小尺寸的输入数据,可以选择小尺寸的卷积核进行卷积运算。但是,这样就要求复杂的硬件平台支持多种卷积核对不同网络层的输入数据分别进行卷积运算,影响了卷积神经网络的运算效率。基于上述输入数据对卷积核选择的影响,可以考虑将输入数据每个深度的数据尺寸减少,这样的话,就可以选择尺寸较小的卷积核对输入数据进行卷积运算,这样,针对不同的网络层,可以使用相同的较小尺寸的卷积核,既保证了卷积运算的运算速率,又提高了卷积神经网络的运算效率。因此,本发明实施例中,通过对输入数据进行处理,达到提高卷积神经网络的运算效率的目的。

[0034] S102,按照预设步长,每次从输入数据中提取多个数据点。

[0035] 为了能够减少网络层的输入数据中每个深度的数据尺寸,同时又不影响输入数据原本的数据量,可以考虑将输入数据的深度增加,即将输入数据中多个数据点映射至不同深度的同一位置,这样就可以减少输入数据每个深度的数据尺寸,而又不影响输入数据原始的数据量。在映射前,需要确定映射至不同深度的同一位置的数据点,为了不影响卷积运算的结果,可以对相邻的数据进行映射,也就是说,可以按照预设步长,每次从输入数据中提取多个数据点,该预设步长可以为预先设定的提取多个数据点的规则,例如,预设步长为2×2,则按照2×2的规则,每次提取各深度上的四个数据点。在提取多个数据的过程中,可以一次提取所有深度上满足预设步长的多个数据点,例如,输入数据的深度为256,预设步长为2×2,则一次提取的数据点为2×2×256个;也可以一次提取一个深度中满足预设步长的多个数据点,例如,预设步长为2×2,则一次提取的数据点为2×2个;还可以一次提取多个深度上满足预设步长的多个数据点,例如,预设步长为2×2,一次提取10个深度的数据点,则一次提取的数据点为2×2×10个。

[0036] S103,将每次提取的多个数据点映射至三维数据中不同深度的同一位置,得到重排列后的数据。

[0037] 在提取了多个数据点后,可以将多个数据点进行映射,即排列至三维数据中不同深度的同一位置,例如,通过上述步骤提取到a、b、c、d四个数据点,则可以按照a、b、c、d四个数据点的任一排列顺序,将该四个数据点排列至连续四个深度的同一位置处,排列顺序可以为[a→b→c→d]、[a→b→d→c]、[a→c→b→d]、[a→c→d→b]、[a→d→b→c]、[a→d→c→b]、[b→a→c→d]、[b→a→d→c]、[b→d→a→c]、[b→d→c→a]、[b→c→a→d]、[b→c→d→a]、[c→a→b→d]、[c→a→d→b]、[c→b→a→d]、[c→b→d→a]、[c→d→a→b]、[c→d→b→a]、[d→a→b→c]、[d→a→c→b]、[d→b→a→c]、[d→b→c→a]、[d→c→a→b]、[d→

c→b→a]中的任一种，其中箭头表示了数据排列的顺序。

[0038] 如果一次提取输入数据的一个深度中满足预设步长的多个数据点，则可以直接根据上述映射方式，将提取的多个数据点排列至新建三维数据中不同深度的同一位置；如果一次提取输入数据的多个深度或者所有深度上满足预设步长的多个数据点，则可以将提取的每个深度的数据点先进行排列，再按深度的顺序将多个数据点排列至新建三维数据中不同深度的同一位置。举例说明，如果输入数据的大小为26×26×10，按照2×2的预设步长进行数据的提取，则得到重排列后的数据的大小为13×13×40，或者每间隔一行/一列提取，则得到重排列后的数据的大小为25×25×40。

[0039] 或者，还可以是多次提取输入数据的一个深度中满足预设步长的多个数据点，然后将每次从各深度中提取的多个数据点映射至新建三维数据中不同深度的同一位置，得到多个待合并数据，最后沿深度方向，将多个待合并数据进行排列，得到重排列后的数据。例如，如果输入数据的大小为26×26×10，按照2×2的预设步长对每个深度进行数据提取，得到每个深度分别对应的13×13×4的待合并数据，再对10个深度的待合并数据进行合并，得到13×13×40的重排列后的数据。具体的合并方式可以为对各深度对应的待合并数据进行任意排列，通过排列得到的重排列后的数据。针对上述深度为10的输入数据，可以有10!＝3628800种排列方式，可以从这些排列方式中任选一种作为各深度对应的待合并数据的合并方式。

[0040] S104，利用预设尺寸的卷积核对重排列后的数据进行卷积运算，得到卷积结果。

[0041] 由于利用上述步骤对输入数据进行处理，输入数据每个深度的数据尺寸减小，则可以利用预设尺寸的卷积核对重排列后的数据进行卷积运算，预设尺寸的卷积核可以为较小尺寸的卷积核，例如3×3的卷积核，或者更小尺寸的卷积核。并且，通过对每个网络层的输入数据进行上述步骤的处理，可以利用相同尺寸的卷积核分别进行卷积运算，则对于卷积神经网络而言，可以利用小尺寸的卷积核对各网络层的输入数据进行卷积运算，因此，可以实现利用简单的硬件平台实现卷积运算，从而提高卷积神经网络的运算效率。

[0042] 应用本实施例，通过按照预设步长，每次从获取的卷积神经网络中网络层的输入数据中提取多个数据点，并将每次提取的多个数据点映射至三维数据中不同深度的同一位置，得到重排列后的数据，最后利用预设尺寸的卷积核对重排列后的数据进行卷积运算，得到卷积结果。由于对网络层的输入数据进行多个数据点的提取及映射操作，将输入数据在深度方向进行扩展，并且减小了每个深度的尺寸，由于输入数据的尺寸变小，则可以利用更小的卷积核对该输入数据进行卷积运算，通过该方法，将各网络层的输入数据进行处理，得到的重排列后的数据均可以利用相同的预设尺寸的卷积核进行卷积运算，从而可以减小硬件资源的开销，并且，针对每个网络层，利用相同的更小尺寸的卷积核进行卷积运算，可以提高卷积神经网络的运算效率。

[0043] 基于图1所示实施例，本发明实施例还提供了一种卷积运算方法，如图2所示，该卷积运算方法包括如下步骤：

[0044] S201，获取卷积神经网络中网络层的输入数据。

[0045] S202，将输入数据沿深度方向进行划分，得到多个切片。

[0046] S203，针对各切片，每次按照预设步长，分别提取该切片中各深度的数据点，得到多个数据点。

[0047] 在提取多个数据点的过程中，如果一次提取所有深度上满足预设步长的多个数据点，由于一次提取和映射的运算量过大，各深度的运算无法并行运行，容易影响运算的速率。因此，可以将输入数据沿深度方向进行划分，得到多个切片。在对输入数据进行划分的过程中，可以将每个深度划分为一个切片，也可以将多个深度划分为一个切片，划分之后，对各切片中数据点的提取可以并行执行，从而可以提高运算的速率。

[0048] S204，将每次从各切片中提取的多个数据点映射至三维数据中不同深度的同一位置，分别得到各切片对应的待合并数据。

[0049] 针对各切片，在每次提取了多个数据点后，可以将多个数据点进行映射，即排列至三维数据中不同深度的同一位置，映射的过程如图1所示实施例相同，这里不再赘述。

[0050] 如果一次提取一个深度中满足预设步长的多个数据点，则可以直接将提取的多个数据点排列至不同深度的同一位置；如果一次提取多个深度上满足预设步长的多个数据点，则可以将提取的每个深度的数据点先进行排列，再按深度的顺序将多个数据点排列至不同深度的同一位置。

[0051] S205，沿深度方向，将多个待合并数据进行排列，得到重排列后的数据。

[0052] 将每次从各切片中提取的多个数据点映射至三维数据中不同深度的同一位置，得到多个待合并数据，然后沿深度方向，将多个待合并数据进行排列，得到重排列后的数据。例如，如果输入数据的大小为26×26×10，将该输入数据沿深度方向进行划分，得到26×26×1、26×26×3、26×26×6三个切片，分别按照2×2预设步长对各切片进行数据提取，得到13×13×4、13×13×12、13×13×24的待合并数据，再将这三个待合并数据进行合并，得到13×13×40的重排列后的数据。合并方式可以为对各切片对应的待合并数据进行任意排列，通过排列得到的重排列后的数据。针对上述输入数据，可以有3!＝6种排列方式，可以从这些排列方式中任选一种作为各切片对应的待合并数据的合并方式。

[0053] S206，利用预设尺寸的卷积核对重排列后的数据进行卷积运算，得到卷积结果。

[0054] 应用本实施例，通过将输入数据沿深度方向进行划分，得到多个切片，然后按照预设步长，每次从各切片中提取多个数据点，并将每次提取的多个数据点映射至三维数据中不同深度的同一位置，并通过合并得到重排列后的数据，最后利用预设尺寸的卷积核对重排列后的数据进行卷积运算，得到卷积结果。对输入数据进行划分之后，各切片中多个数据点的提取可以并行执行，从而可以提高运算的速率；并且，由于对各切片进行多个数据点的提取及映射操作，将各切片在深度方向进行扩展，减小了每个深度的尺寸，则可以利用更小的卷积核对该输入数据进行卷积运算，通过该方法，将各网络层的输入数据进行处理，得到的重排列后的数据均可以利用相同的预设尺寸的卷积核进行卷积运算，从而可以减小硬件资源的开销，并且，针对每个网络层，利用相同的更小尺寸的卷积核进行卷积运算，可以提高卷积神经网络的运算效率。

[0055] 为了便于理解，下面结合具体的应用实例，对发明实施例所提供的卷积运算方法进行介绍。

[0056] 第一步，对于需要将卷积核替换为更小尺寸的卷积核的网络层，记该网络层的输入数据为A，其中，A为大小为W×H×I的一个三维数据。

[0057] 第二步，将A在深度方向上划分为I个切片，分别记录各切片为A_i，如图3所示，其中，i∈[1,I]。

[0058] 第三步，在A_i内按照2×2的步长，每次提取4个数据点a_j，如图3的A_i中虚线框所示，其中，j∈[1,4]。

[0059] 第四步，将提取的数据点映射到三维数据A_i^*对应的同一位置，其中，A_i^*的大小为$W^*×H^*×4$，数据点的排列顺序可以为4!＝24种排列顺序中的任一种。

[0060] 第五步，将A_i^*在深度方向上进行合并，得到重排列后的数据A^*，其中，A^*的大小为$W^*×H^*×4I$，A_i^*的合并方式可以为对各A_i^*进行排列合并，排列顺序可以为I!种排列顺序中的任一种。

[0061] 第六步，基于重排列后的数据A^*，可以利用预设尺寸的卷积核$K_r×K_r×I_r×O$进行卷积运算，得到卷积结果。

[0062] 本方案中，通过将输入数据沿深度方向进行划分，得到多个切片，然后按照2×2的步长，每次从各切片中提取多个数据点，并将每次提取的多个数据点映射至三维数据中不同深度的同一位置，并通过合并得到重排列后的数据，最后利用预设尺寸的卷积核对重排列后的数据进行卷积运算，得到卷积结果。在对输入数据进行划分的过程中，将每个深度划分为一个切片，划分之后，各切片中多个数据点的提取可以并行执行，从而可以提高运算的速率；并且，由于对各切片进行多个数据点的提取及映射操作，将各切片在深度方向进行扩展，减小了每个深度的尺寸，则可以利用更小的卷积核对该输入数据进行卷积运算，通过该方法，将各网络层的输入数据进行处理，得到的重排列后的数据均可以利用相同的预设尺寸的卷积核进行卷积运算，从而可以减小硬件资源的开销，并且，针对每个网络层，利用相同的更小尺寸的卷积核进行卷积运算，可以提高卷积神经网络的运算效率。

[0063] 相应于上述卷积运算方法实施例，如图4所示，本发明实施例还提供了一种卷积运算装置，该卷积运算装置可以包括：

[0064] 获取模块410，用于获取卷积神经网络中网络层的输入数据；

[0065] 提取模块420，用于按照预设步长，每次从所述输入数据中提取多个数据点；

[0066] 映射模块430，用于将每次提取的多个数据点映射至三维数据中不同深度的同一位置，得到重排列后的数据；

[0067] 运算模块440，用于利用预设尺寸的卷积核对所述重排列后的数据进行卷积运算，得到卷积结果。

[0068] 应用本实施例，通过按照预设步长，每次从获取的卷积神经网络中网络层的输入数据中提取多个数据点，并将每次提取的多个数据点映射至三维数据中不同深度的同一位置，得到重排列后的数据，最后利用预设尺寸的卷积核对重排列后的数据进行卷积运算，得到卷积结果。由于对网络层的输入数据进行多个数据点的提取及映射操作，将输入数据在深度方向进行扩展，并且减小了每个深度的尺寸，由于输入数据的尺寸变小，则可以利用更小的卷积核对该输入数据进行卷积运算，通过该方法，将各网络层的输入数据进行处理，得到的重排列后的数据均可以利用相同的预设尺寸的卷积核进行卷积运算，从而可以减小硬件资源的开销，并且，针对每个网络层，利用相同的更小尺寸的卷积核进行卷积运算，可以提高卷积神经网络的运算效率。

[0069] 可选的，所述提取模块420，具体可以用于：

[0070] 针对所述输入数据的每个深度，每次按照预设步长，分别提取多个数据点；

[0071] 所述映射模块430，具体可以用于：

[0072] 将每次从所述输入数据中各深度提取的多个数据点映射至三维数据中不同深度的同一位置，得到多个待合并数据；

[0073] 沿深度方向，将多个待合并数据进行排列，得到重排列后的数据。

[0074] 可选的，所述提取模块420，具体还可以用于：

[0075] 对每次提取的多个数据点进行排列；

[0076] 按照排列的顺序，将每次提取的多个数据点存储至三维数据中不同深度的同一位置，得到重排列后的数据。

[0077] 本实施例所提供的卷积运算装置为应用如图1所示实施例的卷积运算方法的装置，因此，上述卷积运算方法的所有实施例均适用于本卷积运算装置，且具有相同或相似的有益效果，这里不再赘述。

[0078] 基于图4所示实施例，本发明实施例还提供了另一种卷积运算装置，如图5所示，该卷积运算装置可以包括：

[0079] 获取模块510，用于获取卷积神经网络中网络层的输入数据；

[0080] 划分模块520，用于将所述输入数据沿深度方向进行划分，得到多个切片；

[0081] 提取模块530，用于针对各切片，每次按照预设步长，分别提取该切片中各深度的数据点，得到多个数据点；

[0082] 映射模块540，用于将每次从各切片中提取的多个数据点映射至三维数据中不同深度的同一位置，分别得到各切片对应的待合并数据；沿深度方向，将多个待合并数据进行排列，得到重排列后的数据；

[0083] 运算模块550，用于利用预设尺寸的卷积核对所述重排列后的数据进行卷积运算，得到卷积结果。

[0084] 应用本实施例，通过将输入数据沿深度方向进行划分，得到多个切片，然后按照预设步长，每次从各切片中提取多个数据点，并将每次提取的多个数据点映射至三维数据中不同深度的同一位置，并通过合并得到重排列后的数据，最后利用预设尺寸的卷积核对重排列后的数据进行卷积运算，得到卷积结果。对输入数据进行划分之后，各切片中多个数据点的提取可以并行执行，从而可以提高运算的速率；并且，由于对各切片进行多个数据点的提取及映射操作，将各切片在深度方向进行扩展，减小了每个深度的尺寸，则可以利用更小的卷积核对该输入数据进行卷积运算，通过该方法，将各网络层的输入数据进行处理，得到的重排列后的数据均可以利用相同的预设尺寸的卷积核进行卷积运算，从而可以减小硬件资源的开销，并且，针对每个网络层，利用相同的更小尺寸的卷积核进行卷积运算，可以提高卷积神经网络的运算效率。

[0085] 本发明实施例还提供了一种计算机设备，如图6所示，包括处理器601、通信接口602、存储器603和通信总线604，其中，处理器601，通信接口602，存储器603通过通信总线604完成相互间的通信，

[0086] 存储器603，用于存放计算机程序；

[0087] 处理器601，用于执行存储器603上所存放的程序时，实现如下步骤：

[0088] 获取卷积神经网络中网络层的输入数据；

[0089] 按照预设步长，每次从所述输入数据中提取多个数据点；

[0090] 将每次提取的多个数据点映射至三维数据中不同深度的同一位置，得到重排列后

的数据；

[0091] 利用预设尺寸的卷积核对所述重排列后的数据进行卷积运算，得到卷积结果。

[0092] 可选的，所述处理器601还可以实现：

[0093] 将所述输入数据沿深度方向进行划分，得到多个切片；

[0094] 所述处理器601在实现所述按照预设步长，每次从所述输入数据中提取多个数据点的步骤中，具体可以实现：

[0095] 针对各切片，每次按照预设步长，分别提取该切片中各深度的数据点，得到多个数据点；

[0096] 所述处理器601在实现所述将每次提取的多个数据点映射至三维数据中不同深度的同一位置，得到重排列后的数据的步骤中，具体可以实现：

[0097] 将每次从各切片中提取的多个数据点映射至三维数据中不同深度的同一位置，分别得到各切片对应的待合并数据；

[0098] 沿深度方向，将多个待合并数据进行排列，得到重排列后的数据。

[0099] 可选的，所述处理器601在实现所述按照预设步长，每次从所述输入数据中提取多个数据点的步骤中，具体可以实现：

[0100] 针对所述输入数据的每个深度，每次按照预设步长，分别提取多个数据点；

[0101] 所述处理器601在实现所述将每次提取的多个数据点映射至三维数据中不同深度的同一位置，得到重排列后的数据的步骤中，具体可以实现：

[0102] 将每次从所述输入数据中各深度提取的多个数据点映射至三维数据中不同深度的同一位置，得到多个待合并数据；

[0103] 沿深度方向，将多个待合并数据进行排列，得到重排列后的数据。

[0104] 可选的，所述处理器601在实现所述将每次提取的多个数据点映射至三维数据中不同深度的同一位置，得到重排列后的数据的步骤中，具体可以实现：

[0105] 对每次提取的多个数据点进行排列；

[0106] 按照排列的顺序，将每次提取的多个数据点存储至三维数据中不同深度的同一位置，得到重排列后的数据。

[0107] 上述计算机设备提到的通信总线可以是PCI(Peripheral Component Interconnect，外设部件互连标准)总线或EISA(Extended Industry Standard Architecture，扩展工业标准结构)总线等。该通信总线可以分为地址总线、数据总线、控制总线等。为便于表示，图中仅用一条粗线表示，但并不表示仅有一根总线或一种类型的总线。

[0108] 通信接口用于上述计算机设备与其他设备之间的通信。

[0109] 存储器可以包括RAM(Random Access Memory，随机存取存储器)，也可以包括NVM(Non-Volatile Memory，非易失性存储器)，例如至少一个磁盘存储器。可选的，存储器还可以是至少一个位于远离前述处理器的存储装置。

[0110] 上述的处理器可以是通用处理器，包括CPU(Central Processing Unit，中央处理器)、NP(Network Processor，网络处理器)等；还可以是DSP(Digital Signal Processing，数字信号处理器)、ASIC(Application Specific Integrated Circuit，专用集成电路)、FPGA(Field-Programmable Gate Array，现场可编程门阵列)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。

[0111] 本实施例中，该计算机设备的处理器通过读取存储器中存储的计算机程序，并通过运行该计算机程序，能够实现：通过将输入数据沿深度方向进行划分，得到多个切片，然后按照预设步长，每次从各切片中提取多个数据点，并将每次提取的多个数据点映射至三维数据中不同深度的同一位置，并通过合并得到重排列后的数据，最后利用预设尺寸的卷积核对重排列后的数据进行卷积运算，得到卷积结果。对输入数据进行划分之后，各切片中多个数据点的提取可以并行执行，从而可以提高运算的速率；并且，由于对各切片进行多个数据点的提取及映射操作，将各切片在深度方向进行扩展，减小了每个深度的尺寸，则可以利用更小的卷积核对该输入数据进行卷积运算，通过该方法，将各网络层的输入数据进行处理，得到的重排列后的数据均可以利用相同的预设尺寸的卷积核进行卷积运算，从而可以减小硬件资源的开销，并且，针对每个网络层，利用相同的更小尺寸的卷积核进行卷积运算，可以提高卷积神经网络的运算效率。

[0112] 另外，相应于上述实施例所提供的卷积运算方法，本发明实施例提供了一种计算机可读存储介质，用于存储计算机程序，所述计算机程序被处理器执行时，实现如下步骤：

[0113] 获取卷积神经网络中网络层的输入数据；

[0114] 按照预设步长，每次从所述输入数据中提取多个数据点；

[0115] 将每次提取的多个数据点映射至三维数据中不同深度的同一位置，得到重排列后的数据；

[0116] 利用预设尺寸的卷积核对所述重排列后的数据进行卷积运算，得到卷积结果。

[0117] 可选的，所述处理器还可以实现：

[0118] 将所述输入数据沿深度方向进行划分，得到多个切片；

[0119] 所述处理器具体可以实现：

[0120] 针对各切片，每次按照预设步长，分别提取该切片中各深度的数据点，得到多个数据点；

[0121] 所述处理器具体可以实现：

[0122] 将每次从各切片中提取的多个数据点映射至三维数据中不同深度的同一位置，分别得到各切片对应的待合并数据；

[0123] 沿深度方向，将多个待合并数据进行排列，得到重排列后的数据。

[0124] 可选的，所述处理器具体可以实现：

[0125] 针对所述输入数据的每个深度，每次按照预设步长，分别提取多个数据点；

[0126] 所述处理器具体可以实现：

[0127] 将每次从所述输入数据中各深度提取的多个数据点映射至三维数据中不同深度的同一位置，得到多个待合并数据；

[0128] 沿深度方向，将多个待合并数据进行排列，得到重排列后的数据。

[0129] 可选的，所述处理器具体可以实现：

[0130] 对每次提取的多个数据点进行排列；

[0131] 按照排列的顺序，将每次提取的多个数据点存储至三维数据中不同深度的同一位置，得到重排列后的数据。

[0132] 本实施例中，计算机可读存储介质存储有在运行时执行本申请实施例所提供的卷

积运算方法的应用程序，因此能够实现：通过将输入数据沿深度方向进行划分，得到多个切片，然后按照预设步长，每次从各切片中提取多个数据点，并将每次提取的多个数据点映射至三维数据中不同深度的同一位置，并通过合并得到重排列后的数据，最后利用预设尺寸的卷积核对重排列后的数据进行卷积运算，得到卷积结果。对输入数据进行划分之后，各切片中多个数据点的提取可以并行执行，从而可以提高运算的速率；并且，由于对各切片进行多个数据点的提取及映射操作，将各切片在深度方向进行扩展，减小了每个深度的尺寸，则可以利用更小的卷积核对该输入数据进行卷积运算，通过该方法，将各网络层的输入数据进行处理，得到的重排列后的数据均可以利用相同的预设尺寸的卷积核进行卷积运算，从而可以减小硬件资源的开销，并且，针对每个网络层，利用相同的更小尺寸的卷积核进行卷积运算，可以提高卷积神经网络的运算效率。

[0133] 对于计算机设备以及计算机可读存储介质实施例而言，由于其所涉及的方法内容基本相似于前述的方法实施例，所以描述的比较简单，相关之处参见方法实施例的部分说明即可。

[0134] 需要说明的是，在本文中，诸如第一和第二等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来，而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且，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。

[0135] 本说明书中的各个实施例均采用相关的方式描述，各个实施例之间相同相似的部分互相参见即可，每个实施例重点说明的都是与其他实施例的不同之处。尤其，对于系统实施例而言，由于其基本相似于方法实施例，所以描述的比较简单，相关之处参见方法实施例的部分说明即可。

[0136] 以上所述仅为本发明的较佳实施例而已，并非用于限定本发明的保护范围。凡在本发明的精神和原则之内所作的任何修改、等同替换、改进等，均包含在本发明的保护范围内。

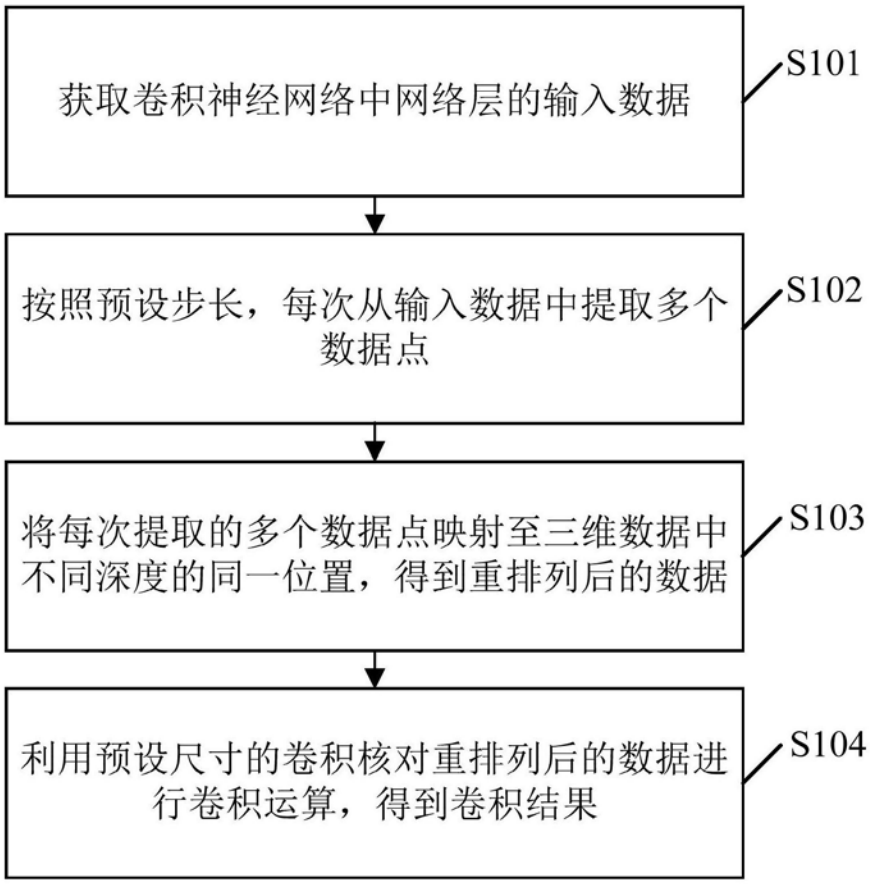

获取卷积神经网络中网络层的输入数据
S101
按照预设步长，每次从输入数据中提取多个数据点
S102
将每次提取的多个数据点映射至三维数据中不同深度的同一位置，得到重排列后的数据
S103
利用预设尺寸的卷积核对重排列后的数据进行卷积运算，得到卷积结果
S104

图1

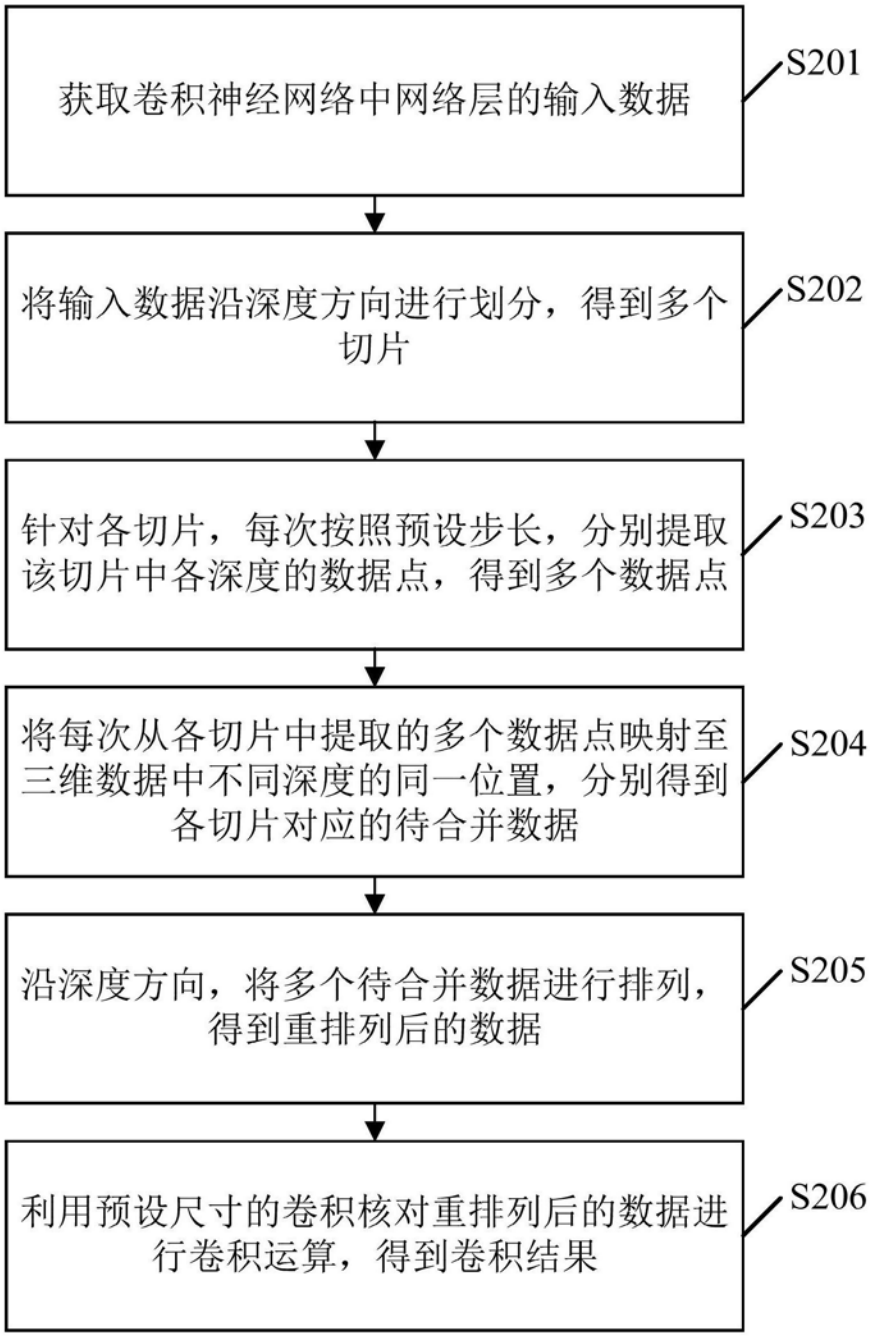

获取卷积神经网络中网络层的输入数据
S201
将输入数据沿深度方向进行划分，得到多个切片
S202
针对各切片，每次按照预设步长，分别提取该切片中各深度的数据点，得到多个数据点
S203
将每次从各切片中提取的多个数据点映射至三维数据中不同深度的同一位置，分别得到各切片对应的待合并数据
S204
沿深度方向，将多个待合并数据进行排列，得到重排列后的数据
S205
利用预设尺寸的卷积核对重排列后的数据进行卷积运算，得到卷积结果
S206

图2

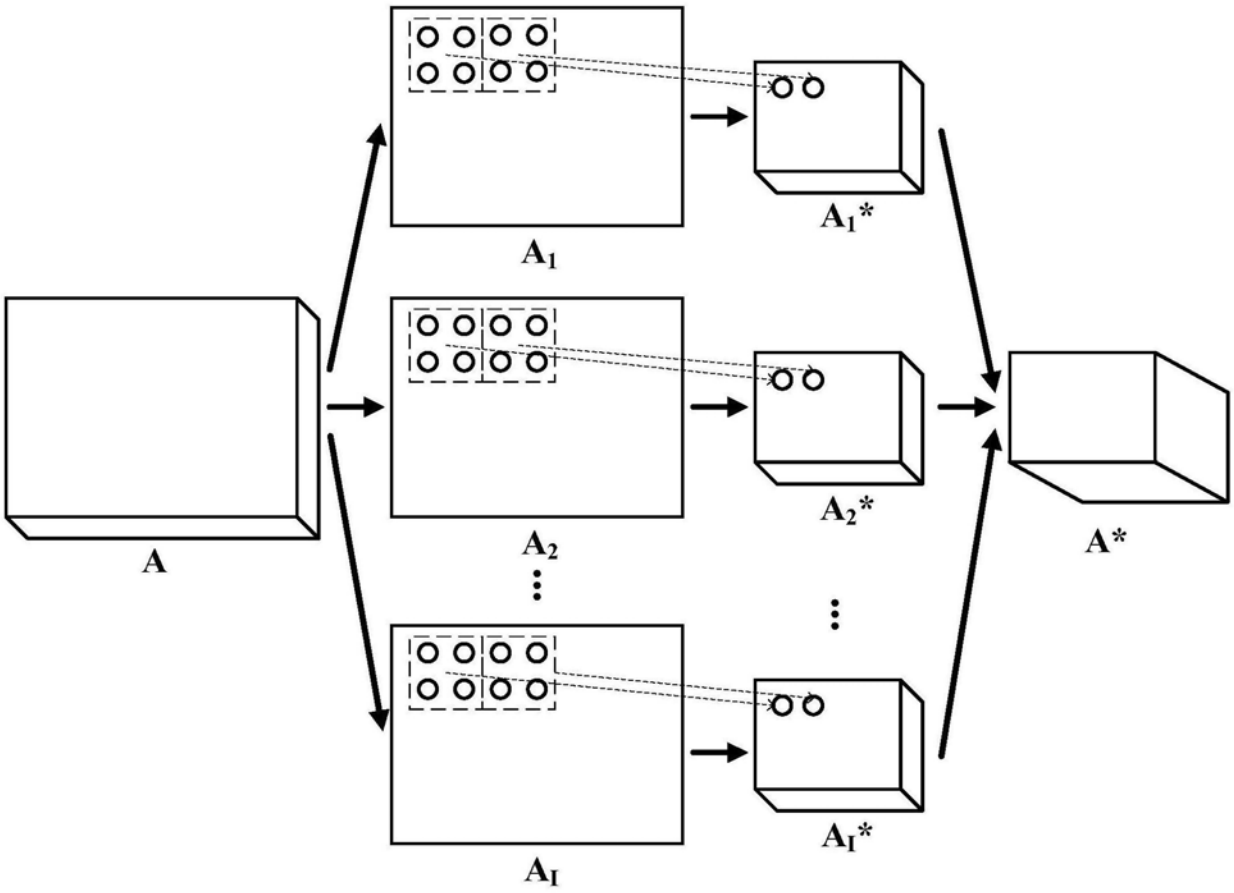

A
A1
A2
AI
A1*
A2*
AI*
A*

图3

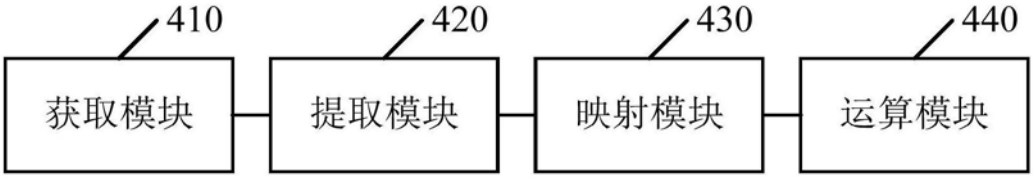

410
420
430
440
获取模块
提取模块
映射模块
运算模块

图4


510
520
530
540
550
获取模块
划分模块
提取模块
映射模块
运算模块

图5

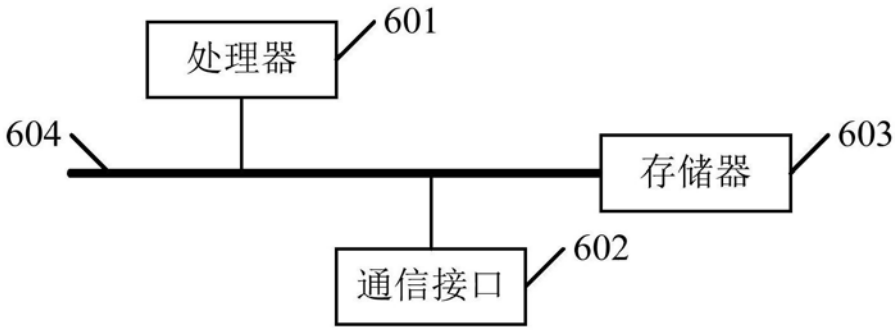


图6