(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
2020 年 12 月 30 日 (30.12.2020)

WIPO | PCT

(10) 国际公布号
WO 2020/258568 A1

(51) 国际专利分类号:
G06F 17/16 (2006.01)

(21) 国际申请号: PCT/CN2019/108928

(22) 国际申请日: 2019 年 9 月 29 日 (29.09.2019)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201910580367.8 2019年6月28日 (28.06.2019) CN

(71) 申请人: 苏州浪潮智能科技有限公司 (INSPUR SUZHOU INTELLIGENT TECHNOLOGY CO., LTD) [CN/CN]; 中国江苏省苏州市官浦路1号9幢, Jiangsu 215000 (CN)。

(72) 发明人: 梅国强 (MEI, Guoqiang); 中国江苏省苏州市官浦路1号9幢, Jiangsu 215000 (CN)。
郝锐 (HAO, Rui); 中国江苏省苏州市官浦路1号9幢, Jiangsu 215000 (CN)。

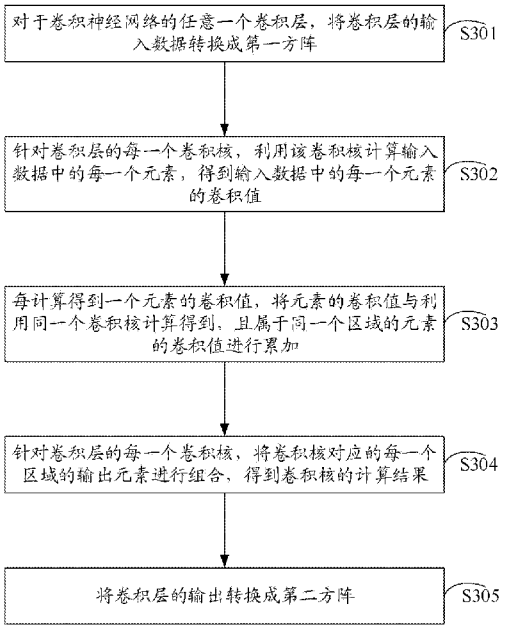
(74) 代理人: 北京集佳知识产权代理有限公司 (UNITALEN ATTORNEYS AT LAW); 中国北京市朝阳区建国门外大街22号赛特广场7层, Beijing 100004 (CN)。

(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, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX,

WO 2020/258568 A1

(54) **Title:** CONVOLUTIONAL NEURAL NETWORK-BASED DATA PROCESSING METHOD AND DEVICE

(54) 发明名称: 基于卷积神经网络的数据处理方法和装置



S301 For any convolutional layer in a convolutional neural network, convert data input to the convolutional layer into a first matrix

S302 For each convolution kernel in the convolutional layer, perform calculation on respective elements of the input data to obtain convoluted values of the respective elements of the input data

S303 After a convoluted value has been obtained from calculation, add together the convoluted value and convoluted values of elements in the same region obtained by calculation by the same convolution kernel

S304 For each convolution kernel in the convolutional layer, combine output elements of respective regions corresponding to the convolution kernel to obtain a calculation result of the convolution kernel

S305 Convert an output of the convolutional layer into a second matrix

图 3

(57) **Abstract:** A convolutional neural network-based data processing method. For any convolutional layer in a convolutional neural network, calculation is performed, by means of a convolution kernel of the convolutional layer, on elements of data input to the convolutional layer one by one so as to obtain convoluted values of the respective elements. Each calculation obtains a convoluted value, and this convoluted value and convoluted values of elements in the same region obtained by calculation by the same convolution kernel are added together to obtain an output element of the convolution kernel corresponding to the region. According to the data

[见续页]

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)。

本国际公布:
— 包括国际检索报告(条约第21条(3))。

processing method, in a process of calculating a convoluted value, each calculation obtains a convoluted value, and this convoluted value is added to a corresponding convoluted sum to directly obtain an element in an output of a convolutional layer. In this way, an output of a convolutional layer can be obtained after all convoluted values have been calculated without having to read convoluted values from a storage apparatus. Since interaction with a storage apparatus is effectively reduced in a process of calculating an output of a convolutional layer, data processing efficiency can be enhanced.

(57) 摘要：一种基于卷积神经网络的数据处理方法，对于所述卷积神经网络的任意一个卷积层，利用卷积层的卷积核逐个计算卷积层的输入数据中的元素，得到每个元素的卷积值，每计算得到一个卷积值，将卷积值与利用同一个卷积核计算得到，且属于同一个区域的元素的卷积值进行累加，得到一个卷积核对应一个区域的输出元素。上述数据处理方法，在计算卷积值的过程中，每计算得到一个卷积值，就将这个卷积值累加至对应的卷积和中，最后直接得到卷积层的输出中的元素，因此计算完所有卷积值就可以得到卷积层的输出，不必再读取存储设备中的卷积值进行计算，有效的减少了计算卷积层的输出的过程中与存储设备的交互，提高数据处理效率。

基于卷积神经网络的数据处理方法和装置

本申请要求于2019年06月28日提交中国专利局、申请号为201910580367.8、发明名称为“基于卷积神经网络的数据处理方法和装置”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本发明涉及深度学习技术，特别涉及一种基于卷积神经网络的数据处理方法和装置。

背景技术

随着深度学习技术的发展，卷积神经网络目前已经被广泛应用于生活中的多个领域，例如，利用卷积神经网络处理视频数据，音频数据或图像数据等，可以自动检测出相似的视频，相似的音频或相似的图像。

卷积神经网络一般包括多个卷积层，并且每一个卷积层均包括多个卷积核，现有的基于卷积神经网络的数据处理方法中，针对一个卷积层，一般是先利用卷积核根据该卷积层的输入计算多个对应的卷积值，每次计算得到的卷积值均被保存在存储设备中，计算所有卷积值后，在根据存储的卷积值计算得到这个卷积层的输出。因此，现有的方法在运行过程中需要频繁的对存储设备进行读操作和写操作，导致处理效率较低。

发明内容

基于上述现有技术的不足，本发明提出一种基于卷积神经网络的数据处理方法和装置，以提高数据处理效率。

本发明第一方面公开一种基于卷积神经网络的数据处理方法，包括：

对于所述卷积神经网络的任意一个卷积层，将所述卷积层的输入数据转换成第一方阵；其中，所述第一方阵是一个N阶方阵，所述N是根据所述卷积层的参数设置的正整数；所述输入数据包括多个输入矩阵，所述第一方阵分割为多个区域，每个所述区域包括的元素均具有相同的矩阵位置，所述元素的矩阵位置，指代所述元素在其对应的输入矩阵中的位置；

针对所述卷积层的每一个卷积核，利用所述卷积核计算所述输入数据中的

每一个元素，得到所述输入数据中的每一个所述元素的卷积值；其中，在利用所述卷积核计算所述输入数据中的每一个元素的过程中，每计算得到一个元素的卷积值，将所述元素的卷积值与利用同一个所述卷积核计算得到，且属于同一个区域的元素的卷积值进行累加，得到一个所述卷积核对应一个区域的输出元素；其中，所述区域为所述第一方阵的每一个区域；

针对所述卷积层的每一个卷积核，将所述卷积核对应的每一个区域的输出元素进行组合，得到所述卷积核的计算结果；所述卷积层的所有卷积核的计算结果，作为所述卷积层的输出。

可选的，利用所述卷积层的所有卷积核计算所述输入数据中的每一个元素的方式，包括：

将所述第一方阵输入多个乘法器，使所述多个乘法器同时利用自身对应的卷积核计算所述输入数据的每一个元素；其中，所述卷积层的所有卷积核被预先分配给所述多个乘法器。

可选的，所述每计算得到一个元素的卷积值，将所述元素的卷积值与利用同一个所述卷积核计算得到，且属于同一个区域的元素的卷积值进行累加，得到一个所述卷积核对应一个区域的输出元素，包括：

每计算得到一个元素的卷积值，利用加法器将所述卷积值与利用同一个所述卷积核计算得到，且属于同一个区域的元素的卷积值进行累加，得到一个所述卷积核对应一个区域的输出元素，所述输出元素保存在预设的寄存器中。

可选的，所述卷积层的每一个卷积核的计算结果均为一个输出矩阵，所述卷积层的所有卷积核的输出矩阵作为所述卷积层的输出；

其中，所述针对所述卷积层的每一个卷积核，将利用所述卷积核计算得到的所有输出元素组合成所述卷积核的计算结果之后，还包括：

将所述卷积层的输出转换成第二方阵；其中，所述第二方阵是一个N阶方阵，所述第二方阵分割为多个区域，每个所述区域包括的元素均具有相同的矩阵位置，所述元素的矩阵位置，指代所述元素在其对应的输出矩阵中的位置。

可选的，所述针对所述卷积层的每一个卷积核，将所述卷积核对应的每一个区域的输出元素进行组合，得到所述卷积核的计算结果；所述卷积层的所有

卷积核的计算结果，作为所述卷积层的输出之后，还包括：

利用所述池化层对所述卷积层的输出进行处理，得到所述卷积层的池化后的输出，所述卷积层的池化后的输出作为所述卷积层的下一个卷积层的输入数据。

本发明第二方面公开一种基于卷积神经网络的数据处理装置，包括：

转换单元，用于对于所述卷积神经网络的任意一个卷积层，将所述卷积层的输入数据转换成第一方阵；其中，所述第一方阵是一个N阶方阵，所述N是根据所述卷积层的参数设置的正整数；所述输入数据包括多个输入矩阵，所述第一方阵分割为多个区域，每个所述区域包括的元素均具有相同的矩阵位置，所述元素的矩阵位置，指代所述元素在其对应的输入矩阵中的位置；

计算单元，用于针对所述卷积层的每一个卷积核，利用所述卷积核计算所述输入数据中的每一个元素，得到所述输入数据中的每一个所述元素的卷积值；其中，在利用所述卷积核计算所述输入数据中的每一个元素的过程中，每计算得到一个元素的卷积值，将所述元素的卷积值与利用同一个所述卷积核计算得到，且属于同一个区域的元素的卷积值进行累加，得到一个所述卷积核对应一个区域的输出元素；其中，所述区域为所述第一方阵的每一个区域；

组合单元，用于针对所述卷积层的每一个卷积核，将所述卷积核对应的每一个区域的输出元素进行组合，得到所述卷积核的计算结果；所述卷积层的所有卷积核的计算结果，作为所述卷积层的输出。

可选的，所述计算单元包括多个乘法器；

所述计算单元利用所述卷积层的所有卷积核计算所述输入数据中的每一个元素，包括：

所述多个乘法器同时利用自身对应的卷积核计算所述输入数据的每一个元素；其中，所述卷积层的卷积核被预先分配给所述多个乘法器。

可选的，所述计算单元包括加法器和寄存器；

所述计算单元用于每计算得到一个元素的卷积值，将所述元素的卷积值与利用同一个所述卷积核计算得到，且属于同一个区域的元素的卷积值进行累加，得到一个所述卷积核对应一个区域的输出元素，包括：

每计算得到一个元素的卷积值，所述加法器将所述卷积值与利用同一个所述卷积核计算得到，且属于同一个区域的元素的卷积值进行累加，得到一个所述卷积核对应一个区域的输出元素；

所述寄存器用于保存所述输出元素。

可选的，所述卷积层的每一个卷积核的计算结果均为一个输出矩阵，所述卷积层的所有卷积核的输出矩阵作为所述卷积层的输出；

其中，所述转换单元还用于：

将所述卷积层的输出转换成第二方阵；其中，所述第二方阵是一个N阶方阵，所述第二方阵分割为多个区域，每个所述区域包括的元素均具有相同的矩阵位置，所述元素的矩阵位置，指代所述元素在其对应的输出矩阵中的位置。

可选的，所述数据处理装置，还包括：

池化单元，用于利用所述池化层对所述卷积层的输出进行处理，得到所述卷积层的池化后的输出，所述卷积层的池化后的输出作为所述卷积层的下一个卷积层的输入数据。

本发明提供一种基于卷积神经网络的数据处理方法和装置，对于所述卷积神经网络的任意一个卷积层，利用卷积层的卷积核逐个计算卷积层的输入数据中的元素，得到每个元素的卷积值，每计算得到一个卷积值，将卷积值与利用同一个卷积核计算得到，且属于同一个区域的元素的卷积值进行累加，得到一个卷积核对应一个区域的输出元素。本发明提供的数据处理方法，在计算卷积值的过程中，每计算得到一个卷积值，就将这个卷积值累加至对应的卷积和中，最后直接得到卷积层的输出中的元素，因此本发明计算完所有卷积值就可以得到卷积层的输出，不必再读取存储设备中的卷积值进行计算，有效的减少了计算卷积层的输出的过程中与存储设备的交互，提高数据处理效率。

附图说明

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，还可以根据提供的附图获得其他的附图。

图 1 为本发明实施例提供的一种卷积神经网络的模型结构示意图；

图 2a 为矩阵的卷积运算示意图；

图 2b 为对矩阵进行池化操作的示意图；

图 3 为本发明实施例提供的一种基于卷积神经网络的数据处理方法的流程图；

图 4 为本发明实施例提供的一种存储卷积神经网络的卷积层的输入数据的数据格式和卷积层的输出的数据格式示意图；

图 5 为本发明实施例提供的一种卷积神经网络的多个卷积层的数据格式示意图；

图 6 为用于实现本发明另一实施例提供的一种基于卷积神经网络的数据处理方法的器件配置图；

图 7 为本发明另一实施例提供的一种基于卷积神经网络的数据处理方法的流程图；

图 8 为本发明另一实施例提供的一种基于卷积神经网络的数据处理方法的输入信息示意图；

图 9 为本发明实施例提供的一种基于卷积神经网络的数据处理装置的结构示意图。

具体实施方式

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。

首先需要说明，本发明提供的基于卷积神经网络的数据处理方法与现有的处理方法的关键区别在于，现有的数据处理方法中每一个卷积层的输入数据均由多个特征矩阵（也可以认为是特征图）构成，并且不同卷积层之间特征图的数量和大小不一致；而本申请任一实施例提供的基于卷积神经网络的数据处理方法，针对组成卷积神经网络的每一个卷积层，将输入数据统一的转换成一个大小一致的特征矩阵（特征图），使得每个卷积层均根据同一形式的输入数据

进行数据处理。基于上述转换，在针对各个卷积层的数据处理过程中，可以对每个卷积层使用相同的或相似的数据存储格式，并且可以使用相同的或者相似的处理时序，因此在基于整个卷积神经网络的数据处理过程中可以不用调整数据存储格式和处理时序，从而有效的提高数据处理效率。

卷积神经网络是一种在深度学习领域被广泛使用的模型，利用训练好的卷积神经网络处理输入数据，可以获得输入数据的某一类特征，从而对输入数据进行分析。卷积神经网络主要包括由若干个卷积层，若干个池化层，一个全连接层，以及一个概率分类函数。卷积神经网络的输入首先经过第一个卷积层处理，然后第一个卷积层的输出又被输入至与第一个卷积层对应的第一个池化层，第一个池化层对输入数据进行池化操作后得到第一个池化层的输出，紧接着第一个池化层的输出又作为第二个卷积层的输入，然后是第二个池化层，第三个卷积层，等等，以此类推，直至得到全连接层的输入，全连接层的输入经过全连接层处理后得到全连接层的输出，最后再利用概率分类函数对全连接层的输出进行处理，就得到卷积神经网络的输出，也就是输入数据的特征。其中，一个卷积神经网络包括的卷积层的数量可以根据实际情况确定，如前文所述，卷积神经网络中池化层的数量等于卷积层的数量。

如图 1 所示，一种常见的卷积神经网络包括 5 个卷积层，5 个池化层，一个全连接层和一个概率分类函数。其中，每个卷积层均包括多个卷积核，具体的，在图 1 所示的常用的卷积神经网络中，第一卷积层包括 64 个卷积核，第二卷积层包括 128 个卷积核，第三卷积层包括 256 个卷积核，第四卷积层和第五卷积层分别包括 512 个卷积核。卷积层对输入数据进行处理，就是利用自身包括的卷积核对输入数据进行运算。

卷积神经网络中，卷积层和池化层的输入数据和输出数据均可以认为由一个或多个矩阵构成（构成输入数据的矩阵可以记为输入矩阵），例如，图 1 所示的卷积神经网络，就可以用 3 个 224 阶的方阵作为这个卷积神经网络的输入数据，或者说，作为这个卷积神经网络的第一卷积层的输入。第一卷积层对输入数据的处理，就是利用自身的 64 个卷积核分别对这 3 个 224 阶的方阵进行计算。

具体的，第一卷积层的64个卷积核各不相同，其中的每一个卷积核均需要对输入数据（即三个224阶的方阵）进行计算，得到这个卷积核对应的计算结果。也就是说，第一卷积层的卷积核1用于对输入数据进行计算，得到卷积核1的计算结果，第一卷积层的卷积核2也用于对输入数据进行计算，得到卷积核2的计算结果，以此类推，最后得到64个计算结果，这64个计算结果，就构成第一卷积层的输出。其他的卷积神经网络中的卷积层的数据过程与上述过程类似。

其中，卷积核对输入数据的计算，实际是利用这个卷积核自身的系数矩阵对输入矩阵进行卷积运算，然后将卷积运算的结果相加就得到这个卷积核的计算结果。具体的，一个卷积核包括若干个系数矩阵，系数矩阵的数量等于构成输入数据的输入矩阵的数量，例如，对于前述3个224阶方阵构成的输入数据，用于计算的卷积核就包括3个系数矩阵；系数矩阵一般是3阶方阵或5阶方阵，当然也可以根据实际需要增加系数矩阵的阶数。另外，一个卷积核中的系数矩阵与输入数据中的输入矩阵一一对应，即，上述3个系数矩阵中，每个系数矩阵对应一个输入矩阵。一个卷积核包括的所有系数矩阵中的元素的取值，在利用样本数据训练卷积神经网络的过程中确定。

一个系数矩阵对一个输入矩阵进行卷积运算，是指用这个系数矩阵计算输入矩阵的每一个元素的卷积值，这些卷积值按照对应的元素在输入矩阵中的位置排列，就构成本次卷积运算的结果。

参考图2a，对输入矩阵进行卷积运算，就是指，将位于系数矩阵的中间位置的元素（在图2a中，就是系数矩阵的第2行第2列的元素），与输入矩阵的一个元素对齐（例如，与输入矩阵的第1行第1列的元素对齐），使得系数矩阵的部分或全部元素与输入矩阵中的部分元素对应（如图2a所示，表示系数矩阵的元素的4个方框内，每个方框均包括一个表示输入矩阵的元素的点），然后将系数矩阵的元素与对应的元素相乘，得到多个乘积，例如，在图2a所示的情况下，系数矩阵的第2行第2列的元素（记为X22）与输入矩阵的第1行第1列的元素（记为Y11）相乘，元素X23与Y12相乘，元素X32与Y21相乘，元素X33与Y22相乘，得到四个乘积，这四个乘积相加，就得

到输入矩阵的第 1 行第 1 列的元素的与这个系数矩阵对应的一个卷积值。以此类推，按上述过程利用这个系数矩阵计算输入矩阵的每一个元素后，得到的卷积值按对应的元素在输入矩阵中的位置排列，就得到本次卷积运算的结果。也就是说，在上述卷积运算的结果（也是一个矩阵）中，输入矩阵的第 1 行第 1 列的元素计算得到的卷积值，也是卷积运算的结果中的第 1 行第 1 列的元素，其他卷积值以此类推。

需要说明的是，如图 2a 所示，对输入矩阵进行卷积运算时，可能出现系数矩阵中的部分元素未与输入矩阵的元素对应的情况，这种情况下，只考虑存在对应关系的元素的乘积即可，因此前述计算过程只对 4 个乘积求和，当然，若系数矩阵的 9 个元素均对应有输入矩阵的元素，则需要计算 9 个乘积，并对 9 个乘积进行求和。

还需要说明的是，前面已经指出，一个卷积层包括多个各不相同的卷积核，每个卷积核都需要用自身的系数矩阵对输入矩阵进行卷积运算。并且，对于输入数据中的每一个元素，一个卷积核中只有与这个元素所在的输入矩阵相对应的系数矩阵可以用于计算该元素。综上所述，对于输入数据中的每一个元素，这个元素会计算得到多个卷积值，其中的每一个卷积值均对应一个卷积核，卷积值的数量等于卷积层包括的卷积核的数量。

对于一个卷积核，这个卷积核的每个系数矩阵都和对应的输入矩阵完成卷积运算后，这些卷积运算的结果中的元素对应的相加，就得到这个卷积核的计算结果。

例如，一个卷积核中包括 3 个系数矩阵，那么这 3 个系数矩阵均完成对应的卷积运算后，就得到 3 个卷积运算的结果，也就是 3 个矩阵，这 3 个矩阵相加，就得到这个卷积核的计算结果，当然，一个卷积核的计算结果仍然是一个矩阵。

对于一个卷积层，这个卷积层的所有卷积核的计算结果的集合，就是这个卷积层的输出。

池化层对输入数据的池化操作，就是指对输入数据中的每一个输入矩阵进行池化操作。对一个矩阵的池化操作，如图 2b 所示，就是指，将一个矩阵分

割为多个2行2列的区域，各个区域互不重叠，然后分别提取该区域内的值最大的元素作为该区域的输出，最后各个区域的输出按对应的位置排列，就构成池化操作的结果。输入数据中每一个输入矩阵的池化操作的结果，就构成一个池化层的输出。

从前面对卷积神经网络的介绍可以发现，利用卷积神经网络处理数据，需要计算得到多个卷积值并计算这些卷积值的和。现有的基于卷积神经网络的数据处理方法中，一般是先计算得到一个卷积层的所有卷积值，然后将这些卷积值对应的组合成多个卷积运算的结果，最后再对这些卷积运算的结果求和得到各个卷积核的计算结果。显然，在计算过程中，每次计算得到的卷积值都需要保存在用于处理数据的计算机的内存中，所有卷积值计算完毕后，又需要从内存中读取这些卷积值用于求和，因此现有的基于卷积神经网络的数据处理方法需要多次对计算机内存进行读写，极大的降低了数据处理效率。

另外，现有的基于卷积神经网络的处理方法中，各个卷积层的输入数据分别是数量不同的多个矩阵（输入数据中包括的矩阵也可以认为是特征图），并且矩阵的行数和列数也大不相同，导致基于整个卷积神经网络的数据处理过程中需要根据输入数据的格式频繁的修改数据存储格式以及对应的处理时序，进一步降低了数据处理效率。

有鉴于此，本申请实施例提供一种基于卷积神经网络的数据处理方法，请参考图3，该方法包括以下步骤：

首先需要说明的是，卷积神经网络中的各个卷积层，除了用于处理数据的参数的数量和取值不同以外，处理数据的过程基本一致。而本申请实施例提供的数据处理方法，主要是通过改进卷积神经网络的各个卷积层的处理过程实现的。并且，在后续介绍本申请实施例的过程中会发现，本申请提供的方法对任意一个卷积层的改进，只需要调整相关参数，就可以直接应用于其他卷积层。因此，在介绍本申请实施例提供的基于卷积神经网络的数据处理方法的过程中，仅基于其中的一个卷积层进行介绍，根据一个卷积层的处理过程，本领域技术人员能够直接的将针对一个卷积层的数据处理方法扩展至任意一个卷积神经网络的任意一个卷积层，因此，通过结合多个卷积神经网络的多个卷积层

执行本申请提供的数据处理方法得到的实施例，也在本申请要求保护的范围内。

为了方便理解，本实施例基于下述例子进行介绍：

本实施例提供的数据处理方法主要应用于一个包括128个卷积核的卷积层，这个卷积层的输入数据包括64个输入矩阵，每一个输入矩阵均为一个112阶的方阵，对应的，结合前述对卷积神经网络的介绍，这个卷积层的每个卷积核均包括64个系数矩阵，每个系数矩阵都是一个3阶方阵。

S301、对于卷积神经网络的任意一个卷积层，将卷积层的输入数据转换成第一方阵。

其中，第一方阵是一个N阶方阵，N是根据卷积层的参数设置的正整数。

本申请实施例提供的数据处理方法，针对每一个卷积层，将输入数据统一的转换成一个行数和列数均固定为N的第一方阵，相当于将现有技术中卷积层的不同格式的多个特征图装换成一个相同大小的第一方阵（第一方阵也可以认为是一个大小为N×N的特征图）。基于上述转换，在针对各个卷积层的数据处理过程中，可以对每个卷积层使用相同的或相似的数据存储格式，并且可以使用相同的或者相似的处理时序，因此在基于整个卷积神经网络的数据处理过程中可以不用调整数据存储格式和处理时序，从而有效的提高数据处理效率。

需要说明的是，步骤S301所述的将输入数据转换成第一方阵，可以理解为以一个方阵的形式保存输入数据中的每一个元素。

第一方阵的阶数N主要根据卷积层的输入矩阵的行数，输入矩阵的列数，以及卷积层的卷积核的数量确定。可选的，结合前述例子，本实施例中可以将第一方阵设置为一个1792阶的方阵。

需要说明的是，第一方阵是一个预先划分为多个区域的方阵，区域的数量等于一个输入矩阵包含的元素的数量，也就是说，在本实施例中，第一方阵被划分为112的平方，12544个区域。具体的，若第一方阵设置为一个1792阶的方阵，那么第一方阵可以按照图4所示的方法划分。

图4中，每一个小方框表示一个16×16，即16行16列的方形区域，可

以发现，水平方向的一行方形区域包括1792除以16，即112个方形区域，竖直方向的一列方形区域也包括112个区域，也就是整个第一方阵包括12544个方形区域。

将输入数据转换成图4所示的1792阶的第一方阵的过程包括：

首先将输入数据包括的64个输入矩阵编号为1至64号，其中，具体哪一个输入矩阵分配哪一个编号不做限定，只需要确保64个编号恰好分配完，每一个输入矩阵均对应一个编号即可。

从输入矩阵1至输入矩阵64，依次获取每个输入矩阵的第一个元素，将这些元素按下述表1的形式填充在第一方阵的第一个方形区域，即第一行第一列的方形区域中（讨论方形区域的位置时，行和列以方形区域为间隔划分）。

表1

1	1	2	2	……	……	8	8
1	1	2	2	……	……	8	8
9	9	……	……	……	……	16	16
9	9	……	……	……	……	16	16
……	……	……	……	……	……	……	……
……	……	……	……	……	……	……	……
57	57	……	……	63	63	64	64
57	57	……	……	63	63	64	64

表1中的数字表示此处的元素属于对应编号的输入矩阵，也就是说，第一个方形区域中，第一行从左到右依次是两个输入矩阵1的第一个元素，两个输入矩阵2的第一个元素，两个输入矩阵3的第一个元素，以此类推，直至两个输入矩阵8的第一个元素为止，第二行与第一行完全相同，第三行则从输入矩阵9开始，按前一行的形式填充，直至输入矩阵16的第一个元素，第四行又与第三行完全相同。相当于第一个方形区域包括的是64个输入矩阵的第一个元素，每个元素均被复制为4份。

可选的，前述表1只是一种填充方式，在另一种填充方式中，可以将一个输入矩阵的第一元素的四个复制排列均排列在一行中。

需要说明的是，上述第一个元素是指矩阵的第一行第一列的元素，为了方便，本申请中涉及元素在输入矩阵中的位置时，直接将输入矩阵中的元素按先左后右，先上后下的顺序编号，因此，对于上述输入矩阵，第 2 个元素是指输入矩阵的第一行第二列的元素，第 113 个元素是指输入矩阵的第二行第一列的元素，对第一方阵中区域的位置的定义类似。

以前述第一个方形区域为例，填充第一矩阵的其他方形区域。最后得到的第一矩阵中，对于任意一个方形区域（如，第 i 个方形区域），区域内的元素为 64 个输入矩阵中，每一个输入矩阵的第 i 个元素，并且每个元素复制为 4 份。

S302、针对卷积层的每一个卷积核，利用该卷积核计算输入数据中的每一个元素，得到输入数据中的每一个元素的卷积值。

需要说明的是，步骤 S302 和步骤 S303 都是需要反复多次执行的步骤，并且，在步骤 S302 的执行过程中，每计算得到一个元素的卷积值，需要执行一次步骤 S303，然后在执行步骤 S302。

需要说明的是，步骤 S302 关键在于指出每一个卷积核都需要执行如步骤 S302 所述的计算输入数据的每一个元素的过程，而并未限定一次只能用一个卷积核进行计算。具体的，在本申请实施例提供的方法中，根据输入数据中的同一个元素在第一方阵中被复制的份数，以及用于计算卷积值的乘法器的数量，可以同时利用多个卷积核进行步骤 S302 所述的计算。

在本实施例中，假设只有一个乘法器用于执行步骤 S302，结合前文可以发现，输入数据中的每一个元素，在第一方阵中均被复制为 4 份，因此，本实施例中可以同时利用卷积层的四个卷积核进行前述计算。

也就是说，结合第一方阵，可以利用四个卷积核同时计算第一方阵中存储的输入矩阵 1 的第一个元素的 4 份复制，得到四个不同的卷积值，接着执行步骤 S303，步骤 S303 执行完后，又利用四个卷积核同时计算第一方阵中存储的输入矩阵 2 的第一个元素的 4 份复制，得到四个卷积值后再执行步骤 S303，以此类推。

如前文所述，利用卷积核计算输入数据中的元素的卷积值，实际是利用这

个卷积核中，与被计算的元素所属的输入矩阵相对应的系数矩阵进行卷积运算。

步骤 S302 中，用于计算的元素是从第一方阵中读取的，但是计算这个元素的卷积值时，使用的仍然是这个元素所属的输入矩阵中的元素，而不是直接将系数矩阵中的元素与第一方阵中的元素一一对应。

具体的，步骤 S302 中，计算一个元素的卷积值，首先需要从用于计算的卷积核中确定出这个元素所属的输入矩阵对应的系数矩阵，然后将这个系数矩阵的中心的元素与被计算的元素对应，接着按前文介绍的卷积运算的过程，从保存在内存中的第一方阵中，读取计算的这个元素所属的输入矩阵的其他元素，然后按照前述卷积运算过程进行计算。

例如，本实施例中，针对第一方阵存储的输入矩阵 1 的第一个元素，利用卷积核（假设为卷积核 A）计算这个元素的卷积值时，首先查找出卷积核 A 中与输入矩阵 1 对应的系数矩阵，然后用这个系数矩阵的中心的元素与输入矩阵 1 的第一个元素对应，根据前文对卷积运算的介绍，这种情况下这个系数矩阵的第二行第三列的元素（元素 X23）应该与输入矩阵 1 的第一行第二列的元素（元素 Y12）对应，元素 X32 与元素 Y21 对应，元素 X33 与元素 Y22 对应，所以从第一方阵中读取上述三个输入矩阵 1 的元素，然后利用读取的三个元素，前述被计算元素（即输入矩阵 1 的第一个元素）以及输入矩阵 1 对应的系数矩阵，计算得到输入矩阵 1 的第一个元素的卷积值，并且，如前文所述，这个卷积值是与卷积核 A 对应的卷积值。

S303、每计算得到一个元素的卷积值，将元素的卷积值与利用同一个卷积核计算得到，且属于同一个区域的元素的卷积值进行累加。

下面结合前述实例介绍步骤 S303 的具体实现过程：

以一个卷积核（记为卷积核 A）为例，步骤 S302 第一次执行时，卷积核 A 从第一方阵中读取输入矩阵 1 的第一个元素，然后计算得到该元素的卷积值，此时还未计算出其他元素的卷积值，故步骤 S303 中的累加，就是将卷积核 A 计算得到的输入矩阵 1 的第一个元素的卷积值保存；

保存后，步骤 S302 再次执行，从第一方阵的第一个方形区域中读取输入

矩阵 2 的元素，计算输入矩阵 2 的元素的卷积值，然后执行步骤 S303，步骤 S303 中，确认此次计算得到的卷积值，与前一个计算得到的卷积值，属于第一方阵的同一个方形区域（均位于第一个方形区域内），并且都是利用卷积核 A 计算得到的卷积值，所以将这个卷积值与前一个卷积值相加，得到一个卷积和，并删除计算得到的两个卷积值；

在后续计算过程中，每次出现利用卷积核 A 计算得到的卷积值，并且这个卷积值是第一方阵的第一个方形区域内存储的元素的卷积值时，就在步骤 S303 中将满足以上条件的卷积值累加至前述卷积和，保留累加后的卷积和并删除原有的卷积和以及卷积值，以此类推，直至第一方阵的第一个方形区域内保存的，分别属于 64 个输入矩阵的 64 个不同的元素均利用卷积核 A 计算得到对应的卷积值后，这 64 个卷积值累加得到的卷积和，就是卷积核 A 针对第一方阵的第一个方形区域计算得到的输出元素。

可选的，步骤 S303 所述的累加过程，可以利用加法器实现，累加得到的卷积核可以保存在寄存器中。

其他卷积核在第一方阵的其他方形区域内的计算过程与上述过程基本一致，不再赘述。

综上所述，通过反复执行步骤 S302 和步骤 S303，最终，对于卷积层的每一个卷积核，都能得到该卷积核对第一方阵的每一个方形区域的输出元素。具体的，本实施例中，卷积层包括 128 个卷积核，第一方阵中有 112×112 个方形区域，所以最后能得到 112×112×128 个输出元素。

S304、针对卷积层的每一个卷积核，将卷积核对应的每一个区域的输出元素进行组合，得到卷积核的计算结果。

卷积层的所有卷积核的计算结果，作为卷积层的输出。

前面已经指出，卷积核的计算结果也是一个矩阵。本实施例中，对于一个卷积核，最后能够计算得到 112×112 个输出元素，并且每一个输出元素均对应于第一方阵的一个方形区域，将这些输出元素，根据各自对应的方形区域在第一方阵内的位置组合成的矩阵，就是这个卷积核的计算结果。

例如，卷积核 A 的计算结果中，第一个元素是卷积核 A 的对应于第一方

阵的第一个方形区域的输出元素，第二个元素是对应于第一方阵的第二个方形区域的输出元素，以此类推，即可组合得到卷积核 A 的计算结果。可以看出，作为一个矩阵，卷积核的计算结果的元素的数量和排列方式，与第一方阵中各个方形区域的数量和排列方式一致，因此，一个卷积核的计算结果，是一个 112×112 的方阵。

本实施例中，卷积层包括 128 个卷积核，因此对应的有 128 个卷积核的计算结果，也就是说，128 个卷积核计算得到的 128 个 112 阶的方阵，就是这个卷积层的输出。

可选的，本实施例提供的方法还包括下述步骤：

S305、将卷积层的输出转换成第二方阵。

其中，第二方阵是一个 N 阶方阵，第二方阵分割为多个区域，每个区域包括的元素均具有相同的矩阵位置，元素的矩阵位置，指代元素在其对应的输出矩阵中的位置。

已知，本实施例中卷积层的输出是 128 个 112 阶的方阵，将这些方阵转换成第二方阵的过程与前述将卷积层的输入数据转换成第一方阵类似，此处不再赘述。

需要说明的是，本实施例中，针对 128 个 112 阶的方阵转换得到的第二方阵仍然以 16 行 16 列的方形区域划分，按照前述方法，第二方阵的第一个方形区域内填充 128 个输出矩阵的第一个元素，第二个方形区域内填充 128 个输出矩阵的第二个元素，以此类推。其中，一个方形区域内包括 128 个输出中的元素，因此，第二方阵中每个输出元素只复制一次。也就是说，第一个方形区域的第一行填充 8 个输出矩阵的第一个元素，每个元素复制为两份，第二行填充另外 8 个输出矩阵的第一个元素，同样的，每个元素复制为两份。其他区域以此类推。

从上述计算过程可以看出，本申请实施例提供的基于卷积神经网络的数据处理方法，在计算卷积核的输出元素时，基于步骤 S302 和步骤 S303 的过程，每次计算得到一个卷积值，就将这个卷积值与利用同一个卷积核计算得到的，属于第一方阵的同一个区域的元素的卷积值累加，并保存累加结果，通过这种

方式，本申请实施例提供的数据处理方法在处理一个卷积层的输入数据以获得输出的过程中，不必保存计算得到的每一个卷积值，能够有效的减少占用的存储空间。并且，基于这种方式，一个卷积核对输入数据中的每一个元素都进行计算后，就可以直接得到这个卷积核对应的所有输出元素，从而可以直接组合得到该卷积核的计算结果，而不必再次从内存中读取所有的卷积值进行计算，因此能够有效的减少计算卷积层的输出的过程中对内存的访问次数，提高数据处理效率。

进一步的，通过在第一方阵中复制输入数据的同一个元素，使得从第一方阵中读取用于计算的输入数据的元素时，能够利用多个卷积核同时读取输入数据的同一个元素，达到多个卷积核同时对输入数据的一个元素进行计算，得到多个卷积值的效果，有效提高数据处理速度。

前述实施例以一个卷积神经网络的一个卷积层为例介绍了本申请提供的数据处理方法，本领域技术人员能够将上述方法直接扩展至一个卷积神经网络的所有卷积层。下面以前述图 1 所示的卷积神经网络为例，介绍前述实施例中针对一个卷积层的方法扩展至整个卷积神经网络后的方法。

首先，获取第一卷积层的输入数据，假设为 3 个 224 阶的方阵，然后将这 3 个 224 阶的方阵按前述实施例的方法转换成 1792 阶的第一卷积层的第一方阵，其形式如图 5 所示，每个 8×8 的方形区域内均填充有 3 个 224 阶的方阵中的对应位置的元素。具体的，以第一个 8×8 的方形区域为例，可以将 3 个 224 阶的方阵的第一个元素分别记为 R1, G1, B1, 三个元素组合成{R1G1B1}，将{R1G1B1}复制为 64 份作为第一个 8×8 的方形区域内的 64 个元素。

然后对第一个卷积层，利用前一实施例提供的数据处理方法处理第一个卷积层的输入数据，得到第一个卷积层的输出，考虑到第一个卷积层包括 64 个卷积核，并且一个卷积层中，卷积核的计算结果的行数和列数与输入矩阵的行数和列数一致，因此，第一个卷积层的输出是 224×224×64，即 64 个 224 阶的方阵，第一个卷积层的输出可以转换成图 5 所示的对应的第二方阵。可以发现，第二方阵中的每个区域为 8 行 8 列，恰好能填充 64 个输出矩阵的对应位置的元素，因此，第一卷积层对应的第二方阵中，输出元素并未复制，每个输

出元素在第二方阵中均只有一个。

利用与第一卷积层连接的第一池化层对第一卷积层的输出进行池化操作，得到池化后的第一卷积层的输出，也就是64个112阶的方阵。池化后的第一卷积层的输出有作为第二卷积层的输入，第二卷积层的输入转换为图5所示的第二卷积层的第一方阵后，又利用本申请提供的数据处理方法进行处理，得到第二卷积层的输出，也就是128个112阶的方阵。转换得到的第二卷积层的对应的第二矩阵如图5所示。

第二卷积层的输出经过第二池化层后，得到池化后的第二卷积层的输出，即128个56阶的方阵，又输入至第三卷积层进行处理。第三卷积层根据输入数据转换得到的第一方阵包括56×56个方形区域，每个方形区域为32×32。第三卷积层的第一方阵中，每个方形区域填充有128个输入矩阵中对应位置的元素，并且每个元素被复制为8份。具体的，第一个32×32的方形区域中，第一行填充有四个输入矩阵的第一个元素，每个元素复制为8份，也就是说，第一个32×32的方形区域的第一行的前8个元素均为同一个输入矩阵的第一个元素，第9至第16个元素为另一个输入矩阵的第一个元素，以此类推。

第三卷积层对自身的输入数据基于前述实施例的方法进行处理后，由于第三卷积层包括256个卷积核，因此第三卷积层得到的输出是256个56阶的方阵，第三卷积层对应的第二矩阵仍然被划分为多个32×32的方形区域，其中，第三卷积层的输出中的元素在对应的第二矩阵中被复制为4份，也就是说，在第二矩阵中，第一个32×32的方形区域的第一行的前4个元素均为同一个输出矩阵的第一个元素，第5至第8个元素为另一个输出矩阵的第一个元素，以此类推。

第四卷积层的第一方阵和第二方阵，以及第五卷积层的第一方阵和第二方阵的形式如图5，具体计算过程可参考前文，不再赘述。

另外，本申请实施例提供的数据处理方法主要对每个卷积层的处理过程进行改进，全连接层和概率分类函数的处理过程可参考现有技术，不再详述。

总而言之，在确定每个卷积层对应的第一方阵和第二方阵均为1792阶的方阵的基础上，各个卷积层的第一方阵和第二方阵的区域划分方式根据对应的

卷积层的输入矩阵行数和列数确定。假设一个卷积层的输入矩阵的行数和列数均为 a，那么对应的第一方阵就划分为多个 b×b 的方形区域，其中，b 等于 1792 除以 a，同样的，这个卷积层对应的第二方阵也划分为多个 b×b 的方形区域。确定第一方阵和第二方阵的划分方式后，就可以根据方形区域与元素在输入矩阵或计算结果中的位置之间的对应关系向第一方阵和第二方阵填充元素，其中，涉及对元素进行复制时，复制的分数根据输入矩阵的数量，以及卷积层的卷积核的数量确定，例如，对于划分为多个 b×b 的方形区域的第一方阵，若输入数据包括的输入矩阵的数量为 c，那么输入数据中每个元素均被复制为 d 份，d 等于 b 的平方除以 c，对于划分为多个 b×b 的方形区域的第二方阵，若对应的卷积层包括的卷积核数量为 e，那么组成该卷积层的输出的每一个输出元素，在第二方阵中被复制为 f 份，f 等于 b 的平方除以 e。a，b，c，d，e 和 f 均为正整数。

前文已经提及，对于一个卷积层，计算该卷积层的输入数据中各个元素的卷积值具体可以由乘法器实现，并且，对卷积值的累积可以结合加法器和寄存器实现。

特别的，在针对一个卷积层的处理过程中，可以设置多个乘法器以及配套的加法器和寄存器，通过使上述器件同时工作，可以有效的提高该卷积层的数据处理效率。

一种可选的配置器件的方式可以是，针对一个卷积层，如果这个卷积层的第一方阵中，输入数据的每个元素被复制为 d 份，并且这个卷积层的卷积核的数量为 e，那么针对这个卷积层的数据处理过程，可以配置的乘法器的数量为 g，g 等于 e 除以 d，并且 g 为正整数。同时，每一个乘法器，均配置一个对应的加法器和一个对应的寄存器。

结合前述图 3 所示的实施例中的例子，对于一个包括 128 个卷积核的卷积层，输入数据为 64 个 112 阶的输入矩阵，如前文所述，在 1792 阶的第一方阵中，输入数据的每一个元素被复制为 4 份，因此，针对这个卷积层，可以配置 128 除以 4，也就是 32 个乘法器，同时对应的配置 32 个加法器和 32 个寄存器，这些器件的连接方式参考图 6，其中，图 6 的 RAM 表示内存，用于存储这个

卷积层的所有卷积核的计算结果，即该卷积层的输出。

结合图 6 所示的器件配置情况，参考图 7，本申请提供的数据处理方法针对上述卷积层的处理过程包括下述步骤：

S701、将卷积层的输入数据转换成第一方阵。

本实施例中，就是将 64 个 112 阶的输入矩阵转换成一个 1792 阶的第一方阵。

S702、将第一方阵输入多个乘法器，使多个乘法器同时利用自身对应的卷积核计算输入数据的每一个元素。

其中，卷积层的卷积核被预先分配个多个乘法器。在本实施例中，输入数据中的元素在第一方阵中被复制为 4 份，使得一个乘法器运行一次，可以同时利用四个卷积核计算得到四个卷积值，因此本实施例中，卷积层的 128 个卷积核被均分至 32 个乘法器，每个乘法器对应四个卷积核。

输入第一方阵后，32 个乘法器同时运行，每运行一次，每个乘法器就输出输入数据中的一个元素对应的四个卷积值。具体的，32 个乘法器首先计算输入矩阵 1 的第一个元素，于是，第一个乘法器得到输入矩阵 1 的第一个元素的四个卷积值，这四个卷积值分别利用第一个乘法器的四个卷积核计算得到，第二个乘法器也输出分别利用第二乘法器的四个卷积核计算得到的四个卷积值，以此类推，也就是说，32 个乘法器运行一次，就能得到输入数据的一个元素对应的 32×4，共 128 个卷积值。

可选的，第一方阵输入乘法器，可以将第一方阵的元素逐行输入乘法器，用于输入数据的输入信号可以参考图 8。其中的 clk 表示时钟信号，vsync 信号表示输入一个 1792 阶的方阵，其中，一个 1792 阶的方阵的输入信号图示的多个 de 信号，每个 de 信号对应 1792 阶的方阵的一行。

S703、乘法器每计算一次，得到的卷积值由加法器对应的累加，并将累加的结果保存在对应的寄存器中。

步骤 S703 是指，计算得到 128 个卷积值后，这 128 个卷积值输入对应的加法器，每个加法器接收其中的 4 个卷积值。

收到卷积值后，加法器从寄存器中读取之前存储的与这些卷积值对应的卷

积和，例如，一个乘法器利用自身的A、B、C和D共四个卷积核，计算第一方阵的第二个方形区域内存储的一个输入数据的元素，准确的说，是利用这四个卷积核分别计算这个元素的四个复制，一个卷积核对应一个复制，得到这个元素对应的四个卷积值。加法器获得这四个卷积值后，读取寄存器中的四个对应的卷积和，这四个卷积和都是根据第一方阵的第二个方形区域中存储的输入数据的元素计算得到的卷积和，并且四个卷积和分别利用A、B、C和D四个卷积核计算得到，然后将这四个卷积值与对应的四个卷积和相加，得到的结果作为新的卷积和，替代原本保存在寄存器中的四个卷积和。

加法器每完成一次累加，就继续执行一次步骤S702，再计算一个输入数据中的元素，再次得到128个卷积值。

通过重复上述过程，输入数据的所有元素均被计算后，寄存器中会保存各个卷积核的输出元素，具体的，一个寄存器对应四个卷积核，每个卷积核对应112×112个输出元素，所以一个寄存器中会保存112×112×4个输出元素。

如前文所述，步骤S702和步骤S703所述的过程，每执行一次，就完成一次对于输入数据的一个元素计算128个卷积值并累加的过程，因此，步骤S702和步骤S703均执行112×112×64次后，输入数据中的所有元素均被计算得到对应的卷积值，然后就可以执行步骤S704。

S704、输入数据中的所有元素均被计算后，卷积核的输出元素在内存中组合为卷积核的计算结果。

步骤S702和步骤S703的计算过程完全结束后，寄存器中保存的输出元素存入RAM，在RAM中按对应的位置组合得到所有卷积核的计算结果，从而得到卷积层的输出。

本实施例提供的方法，根据输入数据在第一方阵中被复制的情况，以及卷积层的卷积核的数量，配置对应数量的乘法器，加法器和寄存器，使得多个乘法器能够同时计算得到多个卷积值，并且同时利用多个加法器进行累加，从而进一步提高了本实施例的数据处理方法的效率。

具体的，通过使乘法器的数量与第一方阵中输入数据被复制的份数的乘积等于卷积层的卷积核的数量，本实施例能够保证，对于任意一个卷积层，只需

要将输入数据的每个元素遍历一次，就可以得到这个卷积层的输出，从而达到一种流水线处理的效果。

当然，前面针对一个卷积层配置的器件，可以直接用于卷积神经网络的其他卷积层的数据处理过程，也可以根据具体适用的卷积层调整器件的数量后，再应用于其他卷积层的处理过程。

结合前述实施例提供的基于卷积神经网络的数据处理方法，本申请另一实施例还提供一种基于卷积神经网络的数据处理装置，如图 9 所示，该装置包括以下结构：

转换单元 901，用于对于卷积神经网络的任意一个卷积层，将卷积层的输入数据转换成第一方阵。

其中，第一方阵是一个 N 阶方阵，N 是根据卷积层的参数设置的正整数；输入数据包括多个输入矩阵，第一方阵分割为多个区域，每个区域包括的元素均具有相同的矩阵位置，元素的矩阵位置，指代元素在其对应的输入矩阵中的位置。

计算单元 902，用于针对卷积层的每一个卷积核，利用卷积核计算输入数据中的每一个元素，得到输入数据中的每一个元素的卷积值，并且，用于在利用卷积核计算输入数据中的每一个元素的过程中，每计算得到一个元素的卷积值，将元素的卷积值与利用同一个卷积核计算得到，且属于同一个区域的元素的卷积值进行累加，得到一个卷积核对应一个区域的输出元素。

其中，上述区域为第一方阵的每一个区域。

组合单元 903，用于针对卷积层的每一个卷积核，将卷积核对应的每一个区域的输出元素进行组合，得到卷积核的计算结果。

卷积层的所有卷积核的计算结果，作为卷积层的输出。

可选的，计算单元 902 包括多个乘法器。

计算单元 902 利用卷积层的所有卷积核计算输入数据中的每一个元素，包括：

多个乘法器同时利用自身对应的卷积核计算输入数据的每一个元素；其

中，卷积层的卷积核被预先分配给多个乘法器。

计算单元902包括加法器和寄存器。

计算单元902用于每计算得到一个元素的卷积值，将元素的卷积值与利用同一个卷积核计算得到，且属于同一个区域的元素的卷积值进行累加，得到一个卷积核对应一个区域的输出元素，包括：

每计算得到一个元素的卷积值，加法器将卷积值与利用同一个卷积核计算得到，且属于同一个区域的元素的卷积值进行累加，得到一个卷积核对应一个区域的输出元素。

寄存器用于保存输出元素。

卷积层的每一个卷积核的计算结果均为一个输出矩阵，卷积层的所有卷积核的输出矩阵作为卷积层的输出。

其中，转换单元901还用于：

将卷积层的输出转换成第二方阵；其中，第二方阵是一个N阶方阵，第二方阵分割为多个区域，每个区域包括的元素均具有相同的矩阵位置，元素的矩阵位置，指代元素在其对应的输出矩阵中的位置。

可选的，数据处理装置，还包括：

池化单元904，用于利用池化层对卷积层的输出进行处理，得到卷积层的池化后的输出，卷积层的池化后的输出作为卷积层的下一个卷积层的输入数据

本发明提供一种基于卷积神经网络的数据处理装置，对于所述卷积神经网络的任意一个卷积层，计算单元902利用卷积层的卷积核逐个计算卷积层的输入数据中的元素，得到每个元素的卷积值，每计算得到一个卷积值，将卷积值与利用同一个卷积核计算得到，且属于同一个区域的元素的卷积值进行累加，得到一个卷积核对应一个区域的输出元素。本发明提供的数据处理方法，在计算卷积值的过程中，每计算得到一个卷积值，就将这个卷积值累加至对应的卷积和中，最后直接得到卷积层的输出中的元素，因此本发明计算完所有卷积值就可以得到卷积层的输出，不必再读取存储设备中的卷积值进行计算，有效的减少了计算卷积层的输出的过程中与存储设备的交互，提高数据处理效率。

专业技术人员能够实现或使用本申请。对这些实施例的多种修改对本领域

的专业技术人员来说将是显而易见的，本文中所定义的一般原理可以在不脱离本申请的精神或范围的情况下，在其它实施例中实现。因此，本申请将不会被限制于本文所示的这些实施例，而是要符合与本文所公开的原理和新颖特点相一致的最宽的范围。

权 利 要 求

1、一种基于卷积神经网络的数据处理方法，其特征在于，包括：

对于所述卷积神经网络的任意一个卷积层，将所述卷积层的输入数据转换成第一方阵；其中，所述第一方阵是一个N阶方阵，所述N是根据所述卷积层的参数设置的正整数；所述输入数据包括多个输入矩阵，所述第一方阵分割为多个区域，每个所述区域包括的元素均具有相同的矩阵位置，所述元素的矩阵位置，指代所述元素在其对应的输入矩阵中的位置；

针对所述卷积层的每一个卷积核，利用所述卷积核计算所述输入数据中的每一个元素，得到所述输入数据中的每一个所述元素的卷积值；其中，在利用所述卷积核计算所述输入数据中的每一个元素的过程中，每计算得到一个元素的卷积值，将所述元素的卷积值与利用同一个所述卷积核计算得到，且属于同一个区域的元素的卷积值进行累加，得到一个所述卷积核对应一个区域的输出元素；其中，所述区域为所述第一方阵的每一个区域；

针对所述卷积层的每一个卷积核，将所述卷积核对应的每一个区域的输出元素进行组合，得到所述卷积核的计算结果；所述卷积层的所有卷积核的计算结果，作为所述卷积层的输出。

2、根据权利要求1所述的数据处理方法，其特征在于，利用所述卷积层的所有卷积核计算所述输入数据中的每一个元素的方式，包括：

将所述第一方阵输入多个乘法器，使所述多个乘法器同时利用自身对应的卷积核计算所述输入数据的每一个元素；其中，所述卷积层的所有卷积核被预先分配给所述多个乘法器。

3、根据权利要求1所述的数据处理方法，其特征在于，所述每计算得到一个元素的卷积值，将所述元素的卷积值与利用同一个所述卷积核计算得到，且属于同一个区域的元素的卷积值进行累加，得到一个所述卷积核对应一个区域的输出元素，包括：

每计算得到一个元素的卷积值，利用加法器将所述卷积值与利用同一个所述卷积核计算得到，且属于同一个区域的元素的卷积值进行累加，得到一个所述卷积核对应一个区域的输出元素，所述输出元素保存在预设的寄存器中。

4、根据权利要求 1 所述的数据处理方法，其特征在于，所述卷积层的每一个卷积核的计算结果均为一个输出矩阵，所述卷积层的所有卷积核的输出矩阵作为所述卷积层的输出；

其中，所述针对所述卷积层的每一个卷积核，将利用所述卷积核计算得到的所有输出元素组合成所述卷积核的计算结果之后，还包括：

将所述卷积层的输出转换成第二方阵；其中，所述第二方阵是一个 N 阶方阵，所述第二方阵分割为多个区域，每个所述区域包括的元素均具有相同的矩阵位置，所述元素的矩阵位置，指代所述元素在其对应的输出矩阵中的位置。

5、根据权利要求 1 所述的数据处理方法，其特征在于，所述针对所述卷积层的每一个卷积核，将所述卷积核对应的每一个区域的输出元素进行组合，得到所述卷积核的计算结果；所述卷积层的所有卷积核的计算结果，作为所述卷积层的输出之后，还包括：

利用所述池化层对所述卷积层的输出进行处理，得到所述卷积层的池化后的输出，所述卷积层的池化后的输出作为所述卷积层的下一个卷积层的输入数据。

6、一种基于卷积神经网络的数据处理装置，其特征在于，包括：

转换单元，用于对于所述卷积神经网络的任意一个卷积层，将所述卷积层的输入数据转换成第一方阵；其中，所述第一方阵是一个 N 阶方阵，所述 N 是根据所述卷积层的参数设置的正整数；所述输入数据包括多个输入矩阵，所述第一方阵分割为多个区域，每个所述区域包括的元素均具有相同的矩阵位置，所述元素的矩阵位置，指代所述元素在其对应的输入矩阵中的位置；

计算单元，用于针对所述卷积层的每一个卷积核，利用所述卷积核计算所述输入数据中的每一个元素，得到所述输入数据中的每一个所述元素的卷积值；其中，在利用所述卷积核计算所述输入数据中的每一个元素的过程中，每计算得到一个元素的卷积值，将所述元素的卷积值与利用同一个所述卷积核计算得到，且属于同一个区域的元素的卷积值进行累加，得到一个所述卷积核对应一个区域的输出元素；其中，所述区域为所述第一方阵的每一个区域；

组合单元，用于针对所述卷积层的每一个卷积核，将所述卷积核对应的每

一个区域的输出元素进行组合，得到所述卷积核的计算结果；所述卷积层的所有卷积核的计算结果，作为所述卷积层的输出。

7、根据权利要求 6 所述的数据处理装置，其特征在于，所述计算单元包括多个乘法器；

所述计算单元利用所述卷积层的所有卷积核计算所述输入数据中的每一个元素，包括：

所述多个乘法器同时利用自身对应的卷积核计算所述输入数据的每一个元素；其中，所述卷积层的卷积核被预先分配给所述多个乘法器。

8、根据权利要求 6 所述的数据处理装置，其特征在于，所述计算单元包括加法器和寄存器；

所述计算单元用于每计算得到一个元素的卷积值，将所述元素的卷积值与利用同一个所述卷积核计算得到，且属于同一个区域的元素的卷积值进行累加，得到一个所述卷积核对应一个区域的输出元素，包括：

每计算得到一个元素的卷积值，所述加法器将所述卷积值与利用同一个所述卷积核计算得到，且属于同一个区域的元素的卷积值进行累加，得到一个所述卷积核对应一个区域的输出元素；

所述寄存器用于保存所述输出元素。

9、根据权利要求 6 所述的数据处理装置，其特征在于，所述卷积层的每一个卷积核的计算结果均为一个输出矩阵，所述卷积层的所有卷积核的输出矩阵作为所述卷积层的输出；

其中，所述转换单元还用于：

将所述卷积层的输出转换成第二方阵；其中，所述第二方阵是一个 N 阶方阵，所述第二方阵分割为多个区域，每个所述区域包括的元素均具有相同的矩阵位置，所述元素的矩阵位置，指代所述元素在其对应的输出矩阵中的位置。

10、根据权利要求 6 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：

池化单元，用于利用所述池化层对所述卷积层的输出进行处理，得到所述卷积层的池化后的输出，所述卷积层的池化后的输出作为所述卷积层的下一个

卷积层的输入数据。

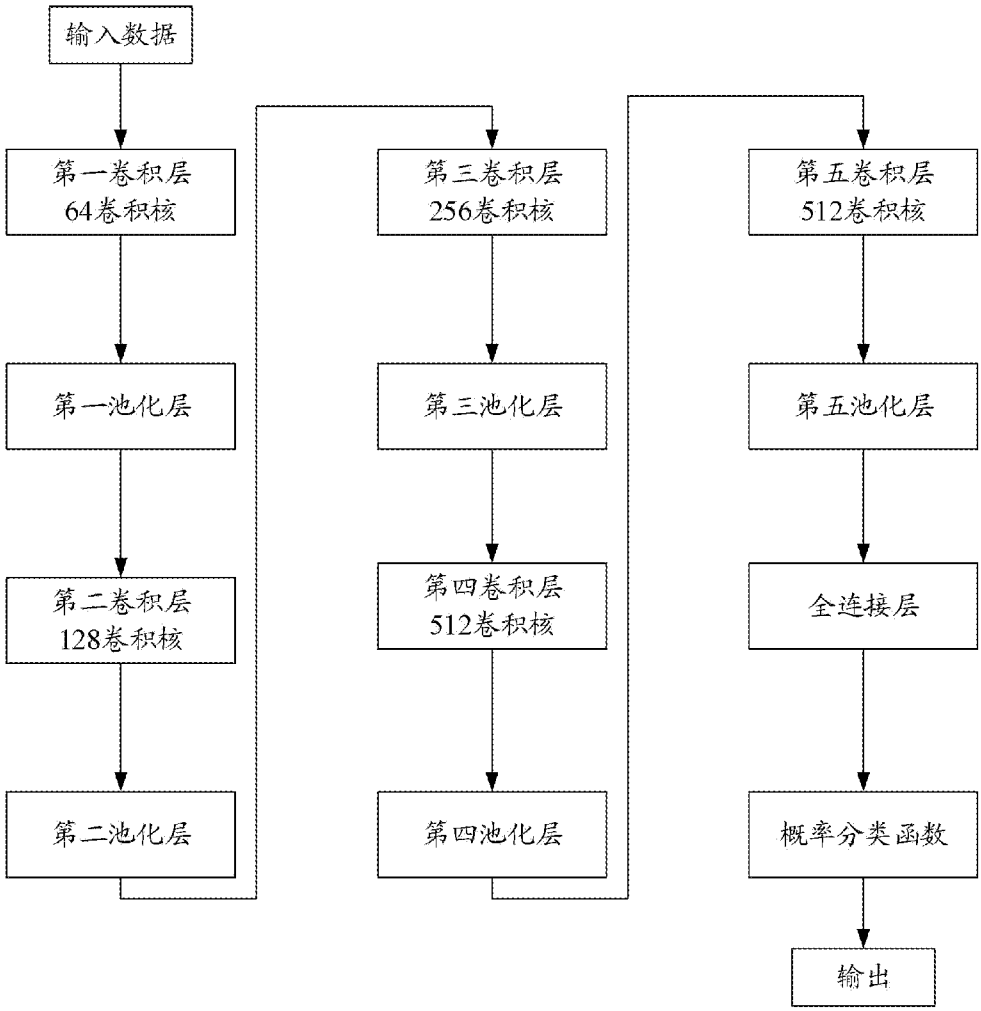

输入数据
第一卷积层
64卷积核
第一池化层
第二卷积层
128卷积核
第二池化层
第三卷积层
256卷积核
第三池化层
第四卷积层
512卷积核
第四池化层
第五卷积层
512卷积核
第五池化层
全连接层
概率分类函数
输出

图 1

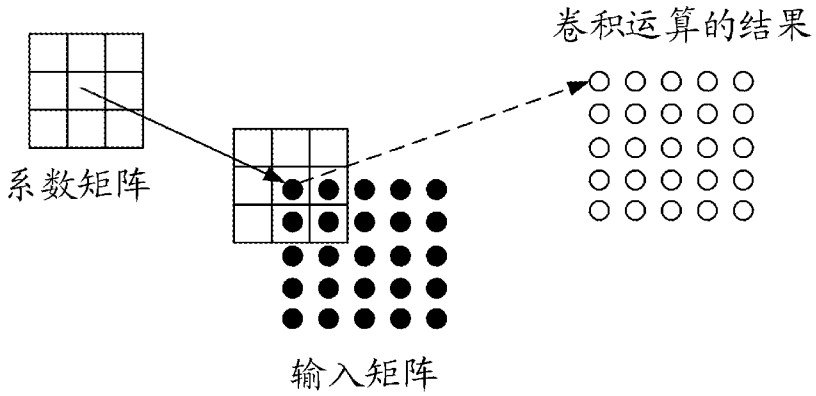


图 2a

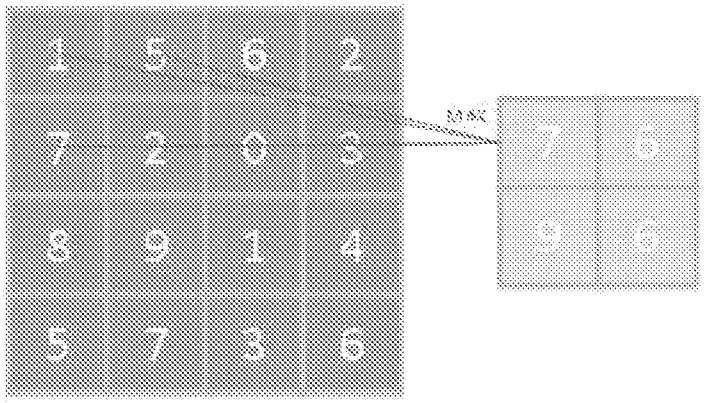


图 2b

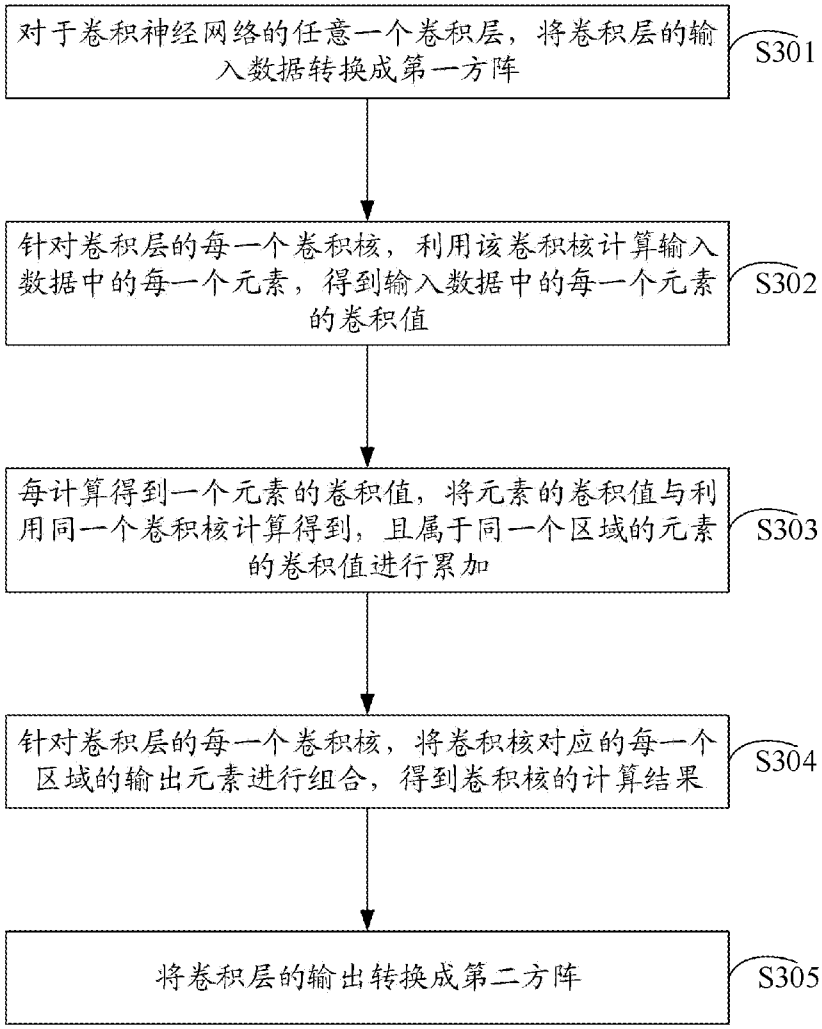
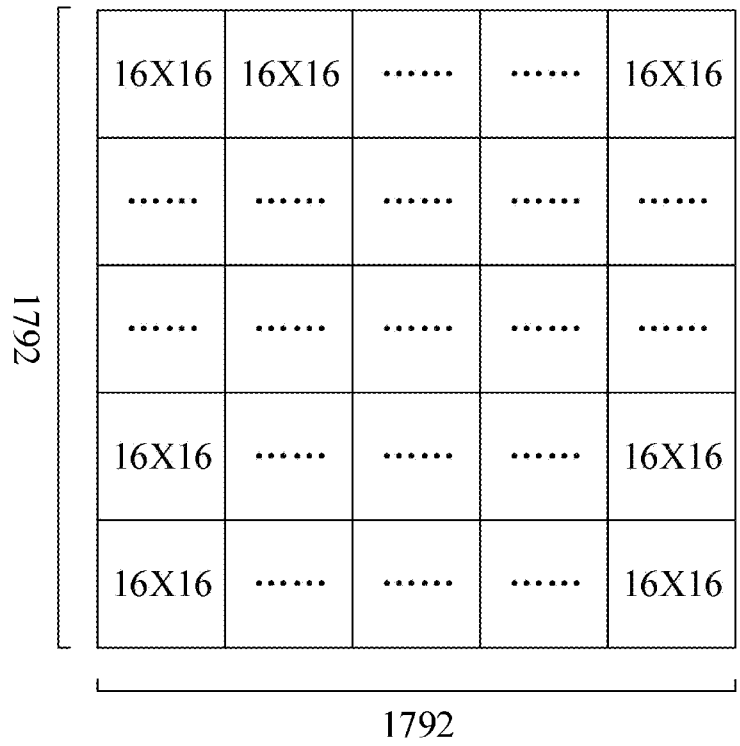

对于卷积神经网络的任意一个卷积层，将卷积层的输入数据转换成第一方阵
S301
针对卷积层的每一个卷积核，利用该卷积核计算输入数据中的每一个元素，得到输入数据中的每一个元素的卷积值
S302
每计算得到一个元素的卷积值，将元素的卷积值与利用同一个卷积核计算得到，且属于同一个区域的元素的卷积值进行累加
S303
针对卷积层的每一个卷积核，将卷积核对应的每一个区域的输出元素进行组合，得到卷积核的计算结果
S304
将卷积层的输出转换成第二方阵
S305

图 3



数据格式示意

图 4

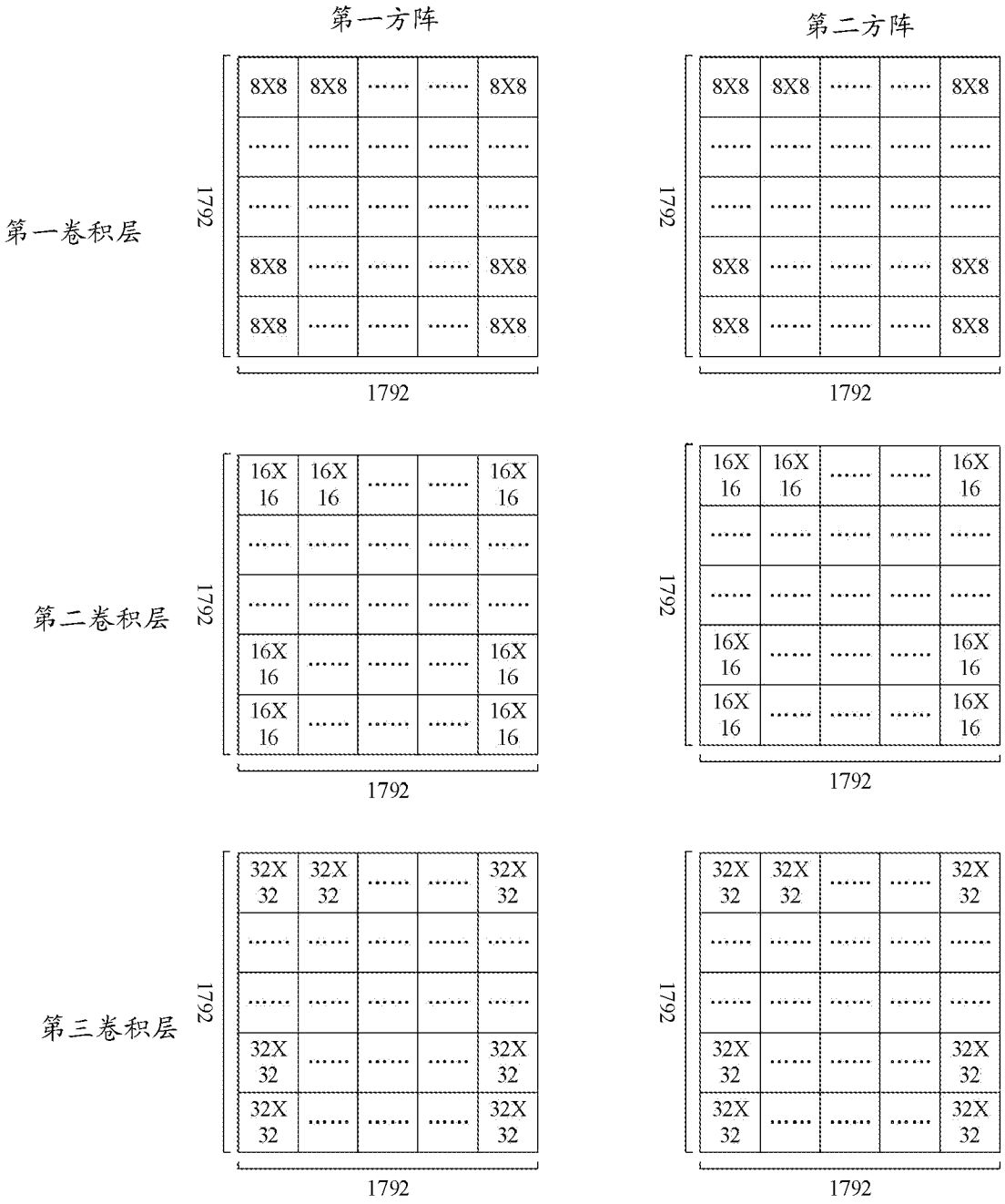


图 5

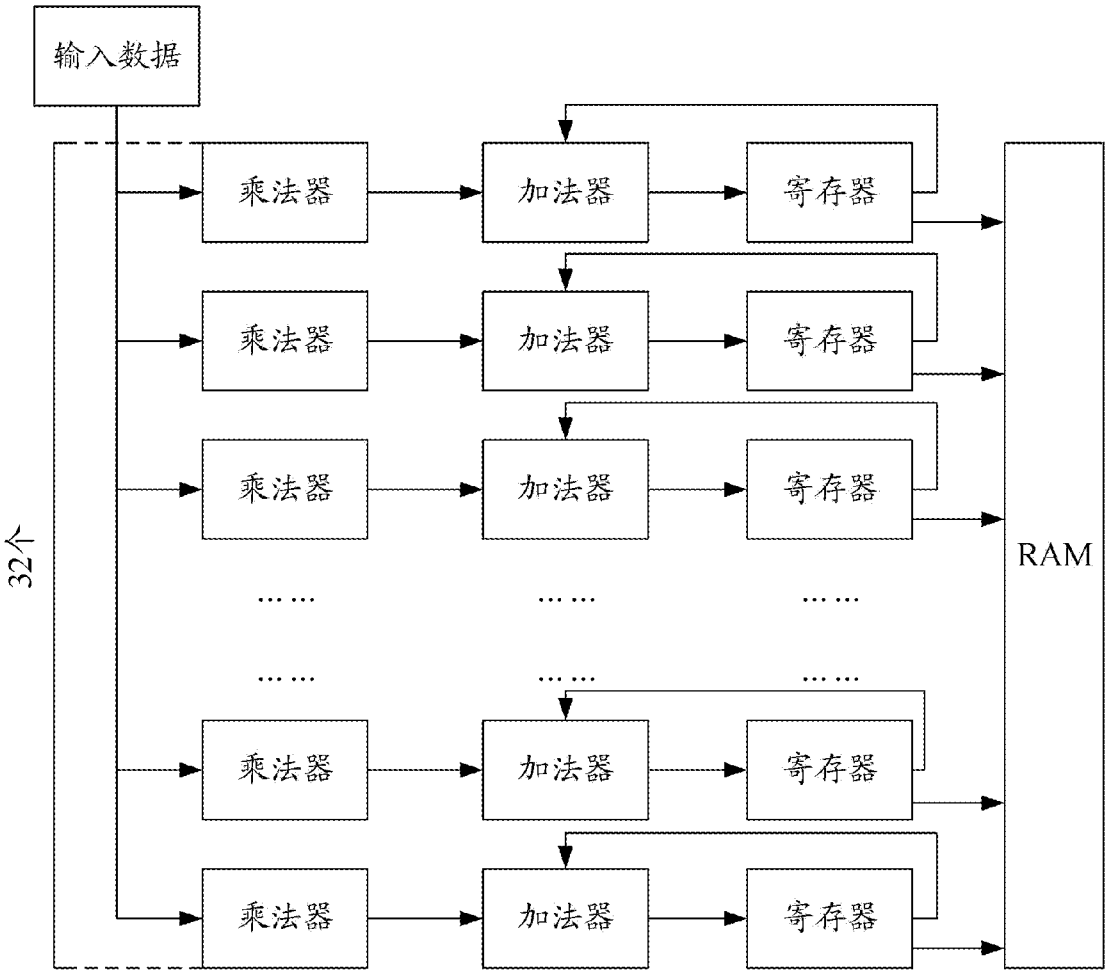

输入数据
乘法器
加法器
寄存器
乘法器
加法器
寄存器
乘法器
加法器
寄存器
……
……
……
……
……
……
乘法器
加法器
寄存器
乘法器
加法器
寄存器
32个
RAM

图 6

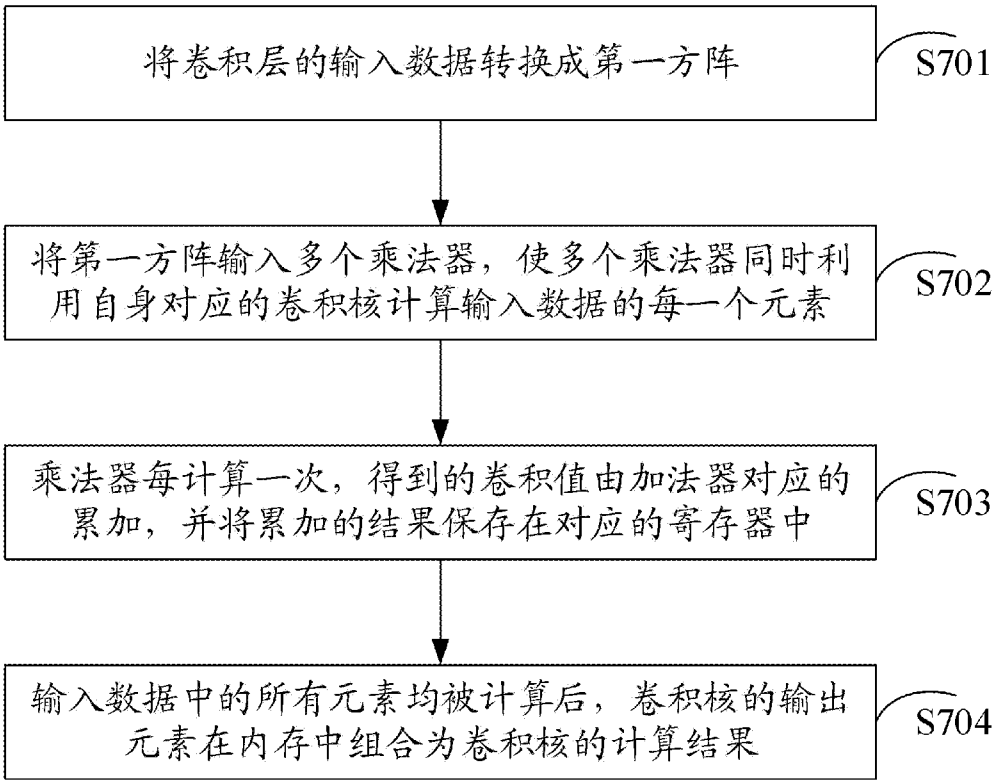

将卷积层的输入数据转换成第一方阵
S701
将第一方阵输入多个乘法器，使多个乘法器同时利用自身对应的卷积核计算输入数据的每一个元素
S702
乘法器每计算一次，得到的卷积值由加法器对应的累加，并将累加的结果保存在对应的寄存器中
S703
输入数据中的所有元素均被计算后，卷积核的输出元素在内存中组合为卷积核的计算结果
S704

图 7

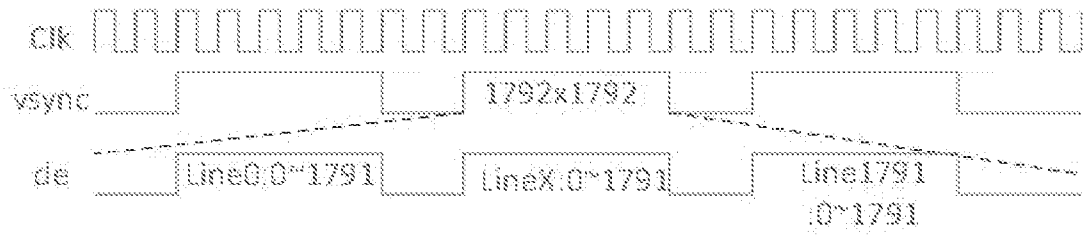

clk
vsync
1792x1792
de
Line0 0~1791
LineX 0~1791
Line1791 0~1791

图 8

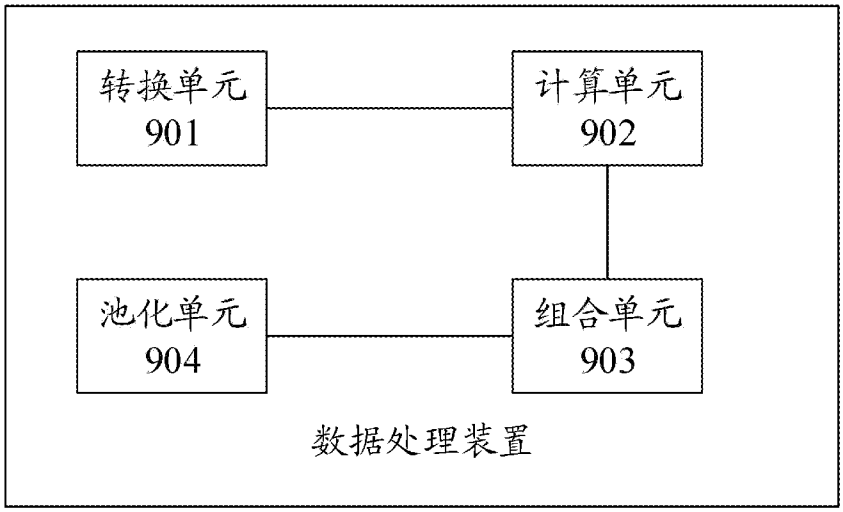


图 9

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2019/108928

A. CLASSIFICATION OF SUBJECT MATTER

G06F 17/16(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)

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; EPODOC; CNKI; CNPAT: 卷积, 神经网络, 方阵, 矩阵, 加, 乘, 核, 元素, convolution, neural, network, matrix, add, sum, accumulate, multiply, kernel, element

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 109086244 A (NATIONAL UNIVERSITY OF DEFENSE TECHNOLOGY OF PLA) 25 December 2018 (2018-12-25) description, paragraphs [0035]-[0048], and figures 1-5	1-10
A	CN 106951395 A (SHANGHAI KELU INFORMATION TECHNOLOGY CO., LTD.) 14 July 2017 (2017-07-14) entire document	1-10
A	CN 108205702 A (NATIONAL UNIVERSITY OF DEFENSE TECHNOLOGY OF PLA) 26 June 2018 (2018-06-26) entire document	1-10
A	US 2017208081 A1 (TATA CONSULTANCY SERVICES LIMITED) 20 July 2017 (2017-07-20) entire document	1-10

☐ 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	Date of mailing of the international search report
16 March 2020	**27 March 2020**
Name and mailing address of the ISA/CN **China National Intellectual Property Administration (ISA/CN)** **No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088** **China**	Authorized officer
Facsimile No. **(86-10)62019451**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2019/108928

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 109086244 A	25 December 2018	None	
CN 106951395 A	14 July 2017	None	
CN 108205702 A	26 June 2018	None	
US 2017208081 A1	20 July 2017	IN 201621000095 A	10 November 2017

国际检索报告

国际申请号 PCT/CN2019/108928

A. 主题的分类

G06F 17/16(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)

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:卷积，神经网络，方阵，矩阵，加，乘，核，元素，convolution,neural,network,matrix,add,sum,accumulate,multiply,kernel,element

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
A	CN 109086244 A (中国人民解放军国防科技大学) 2018年 12月 25日 (2018-12-25) 说明书第[0035]-[0048]段、图1-5	1-10
A	CN 106951395 A (上海客鹭信息技术有限公司) 2017年 7月 14日 (2017-07-14) 全文	1-10
A	CN 108205702 A (中国人民解放军国防科技大学) 2018年 6月 26日 (2018-06-26) 全文	1-10
A	US 2017208081 A1 (TATA CONSULTANCY SERVICES LIMITED) 2017年 7月 20日 (2017-07-20) 全文	1-10

☐ 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	国际检索报告邮寄日期
2020年 3月 16日	2020年 3月 27日
ISA/CN的名称和邮寄地址 中国国家知识产权局(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 冯婷霆 电话号码 86-(10)-53961364

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2019/108928

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）
CN 109086244 A	2018年 12月 25日	无	
CN 106951395 A	2017年 7月 14日	无	
CN 108205702 A	2018年 6月 26日	无	
US 2017208081 A1	2017年 7月 20日	IN 201621000095 A	2017年 11月 10日