说　　明　　书

基于分块卷积的深度优先数据调度方法、系统及设备

技术领域

[0001]　本发明属于卷积神经网络领域，具体涉及了一种基于分块卷积的深度优先数据调度方法、系统及设备。

背景技术

[0002]　随着深度学习技术的不断发展，以卷积神经网络为代表的一系列模型在图像分类、目标检测等领域取得了良好的效果，并在生活中得到了广泛应用。但卷积神经网络中各卷积层的特征图通常较大，采用逐层卷积的方式会占用大量的内存，而全硬件设备通常内存有限，这使得卷积模型难以在全硬件设备上进行部署，在一定程度上限制了卷积神经网络的应用。另外，逐层卷积方式只有在前一层卷积结束后才能进行下一层卷积，灵活性较低，在全硬件设备上可能造成一定的资源浪费。

[0003]　目前采用模型剪枝、量化等方法在一定程度上能够减少模型前向推理时的内存占用，但在模型较大时仍可能出现内存不足的情况。因此，需要设计一种针对全硬件设备的卷积和调度方法，以实现卷积模型在资源有限的全硬件设备上高效运行。

发明内容

[0004]　为了解决现有技术中的上述问题，即现有的全硬件设备的数据调度需要同时对整个图像进行卷积，对图像的处理占用内存过大不宜部署在全硬件设备中的问题，本发明提供了一种基于分块卷积的深度优先数据调度方法，包括：

[0005] 步骤 S100，将第 0 层特征图 feature0 分为 m*n 个预设尺寸为 B 的 block，并设定坐标索引（X，Y），初始化（X，Y）=（0,0），特征图的层数 j=0；

[0006] 步骤 S200，若第 j 层的特征图 featurej 存在未处理的 block，将 featurej 的 block（X，Y）输入下一层网络进行运算；所述下一层网络为卷积层或最大池化层中任一种；若所述下一层网络为卷积层，则生成尺寸为 B 的 block，即尺寸为 B 的 featurej+1 的 block（X，Y）；若所述下一层网络为最大池化层，则生成尺寸为$\frac{1}{4}$B的 block，即尺寸为$\frac{1}{4}$B的 featurej+1 的 block（X，Y）；

[0007] 若 featurej 不存在未处理的 block，则对第 0 层特征图 feature0 的下一个 block（X，Y）进行步骤 S200 的操作，即令 X=NX，Y=NY，j=0，并转入步骤 S200；

[0008] 若 feature0 不存在未处理的 block，,则完成数据调度，最深层的特征图即为前向推理结果；

[0009] 步骤S300，若第j+1层的特征图featurej+1的总尺寸$B_{j+1} < \mathrm{B}$，则对featurej的下一个block（X，Y）重复步骤S200的操作，即令X=NX，Y=XY转入步骤S200；若第j+1层的特征图featurej+1的总尺寸$B_{j+1} = \mathrm{B}$，则将featurej+1的尺寸为B的特征图设置为1个block重复步骤S200的操作，即令j=j+1转入步骤S200。在一些优选的实施方式中，将第0层特征图像feature0的尺寸设为$\mathrm{H} * \mathrm{W} = 2^{M*k} * 2^{N*\mathrm{k}}$，block大小预设为$\mathrm{B} = 2^{k} * 2^{k}$，B小于feature0的尺寸。

[0010] 在一些优选的实施方式中，所述NX_j和NY_j，其获得方法包括：

[0011] X 配置为二进制表示的 m 个 bit 位，从低位至高位为x_0，x_1，x_2，……，x_{m-1}；Y 配置为二进制表示的 m 个 bit 位，从低位至高位为y_0，y_1，y_2，……，y_{m-1}；将坐标 X 和坐标 Y 的二进制表示进行二维混排生成二维混排坐标x_0，y_0，x_1，y_1，x_2，y_2，……，x_{m-1}，y_{m-1}，令所述二

维混排坐标加 1，再进行二维解混排生成坐标 NX 和坐标 NY,每一层的 X 和 Y 相互独立计算，调取顺序均由 NX 和 NY 的获得方法生成。

[0012] 在一些优选的实施方式中，每次经过卷积层之前还包括：对各block进行边缘补零；

[0013] 对于单通道的特征图，边缘补零的尺寸计算公式为：

$$\mathrm{out}_h = \frac{（\mathrm{in}_h + 2 * \mathrm{pad}_h - \mathrm{kernel}_h）}{\mathrm{stride}_h} + 1$$

$$\mathrm{out}_w = \frac{（\mathrm{in}_w + 2 * \mathrm{pad}_w - \mathrm{kernel}_w）}{\mathrm{stride}_w} + 1$$

[0014] 其中，$(\mathrm{in}_w，\mathrm{in}_h)$表示输入特征图的尺寸，$(\mathrm{out}_w，\mathrm{out}_h)$和表示卷积层输出特征图的尺寸，$(\mathrm{kernel}_w，\mathrm{kernel}_h)$表示卷积核尺寸，$(\mathrm{stride}_w，\mathrm{stride}_h)$表示卷积核在宽和高在两个方向上的步长，$(\mathrm{pad}_w，\mathrm{pad}_h)$表示边缘补零的尺寸。

[0015] 在一些优选的实施方式中，每当featurej的总尺寸B_{j+1}=B时，继续进行卷积或最大池化，不保存featurej的结果。

[0016] 在一些优选的实施方式中，本发明的方法只要数据准备完毕，便可以直接调用卷积核进行卷积。

[0017] 本发明的另一方面，提出了一种基于分块卷积的深度优先数据调度系统，包括图像划分模块、特征图像卷积模块和深度优先调用模块；

[0018] 所述图像划分模块，将第 0 层特征图 feature0 分为 m*n 个预设尺寸为 B 的 block，并设定坐标索引（X，Y），初始化（X，Y）=（0,0），特征图的层数 j=0；

[0019] 所述特征图像卷积模块，若第 j 层的特征图 featurej 存在未处理的 block，将 featurej 的 block（X，Y）输入下一层网络进行运算；所述下一层网络为卷积层或最大池化层中任一种；若所述下一层网络为

卷积层，则生成尺寸为 B 的 block，即尺寸为 B 的 featurej+1 的 block（X，Y）；若所述下一层网络为最大池化层，则生成尺寸为$\frac{1}{4}$B的 block，即尺寸为$\frac{1}{4}$B的 featurej+1 的 block（X，Y）；

[0020] 若 featurej 不存在未处理的 block，则对第 0 层特征图 feature0 的下一个 block（X，Y）进行特征图像卷积模块的功能，即令 X=NX，Y=NY，j=0，并转入特征图像卷积模块；

[0021] 若 feature0 不存在未处理的 block，则完成数据调度，最深层的特征图即为前向推理结果；

[0022] 所述深度优先调用模块，配置为若第 j+1 层的特征图 featurej+1 的总尺寸$B_{j+1} < \mathrm{B}$，则对 featurej 的下一个 block（X，Y）重复特征图像卷积模块的功能，即令 X=NX，Y=XY 转入特征图像卷积模块；若第 j+1 层的特征图 featurej+1 的总尺寸$B_{j+1} = \mathrm{B}$，则将 featurej+1 的特征图设置为 1 个 block 重复特征图像卷积模块的功能，即令 j=j+1 转入特征图像卷积模块。

[0023] 本发明的第三方面，提出了一种电子设备，其特征在于，包括：至少一个处理器；以及与至少一个所述处理器通信连接的存储器；其中，所述存储器存储有可被所述处理器执行的指令，所述指令用于被所述处理器执行以实现上述的基于分块卷积的深度优先数据调度方法。

[0024] 本发明的第四方面，提出了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存储有计算机指令，所述计算机指令用于被所述计算机执行以实现上述的基于分块卷积的深度优先数据调度方法。

[0025] 本发明的有益效果：

[0026] （1）本发明基于分块卷积的深度优先数据调度方法，通过对分块卷积的进行深度优先的调度方法替代了现有的必须将图像进行逐

层卷积的调用方法，避免了存储大量卷积层中间结果所带来的内存消耗，提高了卷积模型在全硬件设备上的推理效率。

附图说明

[0027] 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：

[0028] 图 1 是本发明基于分块卷积的深度优先数据调度方法实施例的流程示意图；

[0029] 图 2 是本发明基于分块卷积的深度优先数据调度方法实施例中将特征图分块补零的示意图；

[0030] 图 3 是本发明基于分块卷积的深度优先数据调度方法实施例的深度优先调度的原理示意图；

[0031] 图 4 是本发明基于分块卷积的深度优先数据调度方法实施例实施例中二维混排和解混排的原理示意图；

[0032] 图 5 是用于实现本申请方法、系统、设备实施例的服务器的计算机系统的结构示意图。

具体实施方式

[0033] 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。

[0034] 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。

[0035] 本发明的一种基于分块卷积的深度优先数据调度方法，包括：

[0036] 步骤 S100，将第 0 层特征图 feature0 分为 m*n 个预设尺寸为 B 的 block，并设定坐标索引（X，Y），初始化（X，Y）=（0,0），特征图的层数 j=0；

[0037] 步骤 S200，若第 j 层的特征图 featurej 存在未处理的 block，将 featurej 的 block（X，Y）输入下一层网络进行运算；所述下一层网络为卷积层或最大池化层中任一种；若所述下一层网络为卷积层，则生成尺寸为 B 的 block，即尺寸为 B 的 featurej+1 的 block（X，Y）；若所述下一层网络为最大池化层，则生成尺寸为$\frac{1}{4}$B的 block，即尺寸为$\frac{1}{4}$B的 featurej+1 的 block（X，Y）；

[0038] 若 featurej 不存在未处理的 block，则对第 0 层特征图 feature0 的下一个 block（X，Y）进行步骤 S200 的操作，即令 X=NX，Y=NY，j=0，并转入步骤 S200；

[0039] 若 feature0 不存在未处理的 block，则完成数据调度，最深层的特征图即为前向推理结果；

[0040] 步骤 S300，若第 j+1 层的特征图 featurej+1 的总尺寸$B_{j+1} < \mathrm{B}$，则对 featurej 的下一个 block（X，Y）重复步骤 S200 的操作，即令 X=NX，Y=XY 转入步骤 S200；若第 j+1 层的特征图 featurej+1 的总尺寸$B_{j+1} = \mathrm{B}$，则将 featurej+1 的尺寸为 B 的特征图设置为 1 个 block 重复步骤 S200 的操作，即令 j=j+1 转入步骤 S200。

[0041] 为了更清晰地对本发明基于分块卷积的深度优先数据调度方法进行说明，下面结合图 1 对本发明实施例中各步骤展开详述。

[0042] 本发明第一实施例的基于分块卷积的深度优先数据调度方法，包括步骤 S100-步骤 S300，各步骤详细描述如下：

[0043] 步骤 S100，将第 0 层特征图 feature0 分为 m*n 个预设尺寸为 B 的 block，并设定坐标索引（X，Y），初始化（X，Y）=（0,0），特征图的层数 j=0；

[0044] 在本实施例中，如图 2 所示，在特征图每次经过卷积层之前，对各 block 进行边缘补零，图 2 左侧是一个 H*W 的特征图示意图，根据预设的分块大小对特征图进行分块，图 2 右侧是对分块后的每一个 block 进行边缘补零的示意图，分块后，特征图由原来的一次加载整张特征图变为分四次加载一个 block 进行卷积，也可根据实际情况将特征图分其他块数卷积，在此不做具体的限定。

[0045] 对于单通道的特征图，边缘补零的尺寸计算公式为：

$$\mathrm{out}_h = \frac{（\mathrm{in}_h + 2 * \mathrm{pad}_h - \mathrm{kernel}_h）}{\mathrm{stride}_h} + 1$$

$$\mathrm{out}_w = \frac{（\mathrm{in}_w + 2 * \mathrm{pad}_w - \mathrm{kernel}_w）}{\mathrm{stride}_w} + 1$$

[0046] 其中，$(\mathrm{in}_w，\mathrm{in}_h)$表示输入特征图的尺寸，$(\mathrm{out}_w，\mathrm{out}_h)$和表示卷积层输出特征图的尺寸，$(\mathrm{kernel}_w，\mathrm{kernel}_h)$表示卷积核尺寸，$(\mathrm{stride}_w，\mathrm{stride}_h)$表示卷积核在宽和高在两个方向上的步长，$(\mathrm{pad}_w，\mathrm{pad}_h)$表示边缘补零的尺寸。

[0047] 对于每个 block，需要设置其输入卷积的特征图尺寸与卷积输出的特征图尺寸相同，即$(\mathrm{out}_w，\mathrm{out}_h) = (\mathrm{in}_w，\mathrm{in}_h)$，已知卷积核大小和卷积步长后，可通过边缘补零的尺寸计算公式计算得到宽高两个方向上的边缘补零尺寸。以图 2 为例，若每个 block 的$(\mathrm{in}_w，\mathrm{in}_h) = (5,5)$，设卷积核大小$(\mathrm{kernel}_w，\mathrm{kernel}_h) = (3,3)$及卷积步长$(\mathrm{stride}_w，\mathrm{stride}_h) = (1,1)$，若要满足卷积输出特征图尺寸$(\mathrm{out}_w，\mathrm{out}_h) = (5,5)$，则通过所述边缘补零的尺寸计算公式计算得到$(\mathrm{pad}_w，\mathrm{pad}_h) = (1,1)$，即在宽度方向上两边各补一列零，在高度方向上两遍各补一列零。

[0048] 在本实施例中，比如，将第 0 层特征图像 feature0 的尺寸设为$\mathrm{H} * \mathrm{W} = 2^{M*k} * 2^{N*\mathrm{k}}$，block 大小预设为$\mathrm{B} = 2^k * 2^k$，B 小于 feature0 的尺寸。特征图经过卷积层后保持大小不变，最大池化层使特征图边长变

为池化前特征图的一半，因此，在实际运算过程中，只可能出现两种情况；第一种情况，特征图的边长大于 B，则一定可按 B 均分成若干个 block；第二种情况，特征图边长小于 B，则按照实际大小继续进行卷积。

[0049] 若第 j 层的特征图 featurej 存在未处理的 block，将 featurej 的 block（X，Y）输入下一层网络进行运算；所述下一层网络为卷积层或最大池化层中任一种；若所述下一层网络为卷积层，则生成尺寸为 B 的 block，即尺寸为 B 的 featurej+1 的 block（X，Y）；若所述下一层网络为最大池化层，则生成尺寸为$\frac{1}{4}$B的 block，即尺寸为$\frac{1}{4}$B的 featurej+1 的 block（X，Y）；

[0050] 若 featurej 不存在未处理的 block，则对第 0 层特征图 feature0 的下一个 block（X，Y）进行步骤 S200 的操作，即令 X=NX，Y=NY，j=0，并转入步骤 S200；

[0051] 若 feature0 不存在未处理的 block，则完成数据调度，最深层的特征图即为前向推理结果；

[0052] 步骤 S300，若第 j+1 层的特征图 featurej+1 的总尺寸$B_{j+1} < \mathrm{B}$，则对 featurej 的下一个 block（X，Y）重复步骤 S200 的操作，即令 X=NX，Y=XY 转入步骤 S200；若第 j+1 层的特征图 featurej+1 的总尺寸$B_{j+1} = \mathrm{B}$，则将 featurej+1 的尺寸为 B 的特征图设置为 1 个 block 重复步骤 S200 的操作，即令 j=j+1 转入步骤 S200。

[0053] 根据输入图像的大小和模型中最大化池化层的个数，最后一层特征图的大小可能大于、等于或小于预设的 block 大小 B。

[0054] 在本实施例中，所述NX_j和NY_j可以通过预设的排序表获得，也可根据本申请特别提出的二维调度方法获得，所述二维调度方法可通过设置二维调度装置的全硬件方法执行，如图 4 所示，具体为：

[0055] X 配置为二进制表示的 m 个 bit 位，从低位至高位为x_0，x_1，x_2，……，x_{m-1}；Y 配置为二进制表示的 m 个 bit 位，从低位至高位为y_0，

y_1，y_2，……，y_{m-1}；将坐标 X 和坐标 Y 的二进制表示进行二维混排生成二维混排坐标x_0，y_0，x_1，y_1，x_2，y_2，……，x_{m-1}，y_{m-1}，令所述二维混排坐标加 1，再进行二维解混排生成坐标 NX 和坐标 NY,每一层的 X 和 Y 相互独立计算，调取顺序均由 NX 和 NY 的获得方法生成。

[0056] 以第 0 层特征图为例，如图 3 所示，图 3 左为 feature0 是一个由 m*n 个 block 组成的特征图；首先调用 block(0,0)进行卷积，得到 feature1 (0,0)， feature1 (0,0)刚好为一个 block 大小，继续向下一层执行。图 3 中表示，下一层为 maxpool 层，执行后得到 feature2 的 block(0,0)为一个 block 的四分之一大小，此时将不再继续向下一层执行，重新回到 feature0；

通过依次调度 feature0 的 block(0,0)、block(1,0)、block(0,1)和 block(1,1),分别执行卷积和 maxpool 后，使得 feature2 的 block(0,0)满足一个 block 大小，如图 3 右所示，此时不再回到 feature0 进行调度，而是调用 feature2 的 block(0,0)继续向下执行，直到遇到 maxpool 层使 feature 小于一个 block 大小时，再回溯到 feature0 继续调用 block(2,0)、block(3,0)、block(2,1)、block(3,1)，以此类推，直到完成前向推理过程；

[0057] 在本实施例中，只要数据准备完毕，便可以直接调用卷积核进行卷积。

[0058] 本实施例中，每当 featurej 的总尺寸B_{j+1}=B 时，继续进行卷积或最大池化，不保存 featurej 的结果。

[0059] 现有的特征图处理方法，通常对 feature0 全部进行卷积得到完整的下一层特征图 feature1，再进行下一层的卷积。现有的卷积方法需要占用较大的内存来存储特征图 feature1，在池化过程中也存在需要占用内存来存储特征图 feature2 的问题。而本申请不向内存中保存 feature1 的结果而是直接将其作为中间结果进行下一层运算。若下一层仍为卷积层，可得到 feature2 的 block（0,0），由于其满足一个 block 大小（或未经过

最大池化操作）可继续向下一层卷积，而不将其保存。若下一层为最大池化层，可得到 feature2 的 block（0,0），其大小只有 block 的四分之一，此时回溯至特征图 feature0 的 block（0,1）进行卷积，重复回溯过程直至 feature2 的尺寸为一个 block 大小。

[0060] 本发明第二实施例的基于分块卷积的深度优先数据调度系统，包括：图像划分模块、特征图像卷积模块和深度优先调用模块；

[0061] 所述图像划分模块，将第 0 层特征图 feature0 分为 m*n 个预设尺寸为 B 的 block，并设定坐标索引（X，Y），初始化（X，Y）=（0,0），特征图的层数 j=0；；

[0062] 在本实施例中，所述系统还包括图像补零模块，配置为在特征图每次经过卷积层之前，将各block进行边缘补零，对于单通道的特征图，边缘补零的尺寸计算公式为：

$$\mathrm{out}_h = \frac{（\mathrm{in}_h + 2 * \mathrm{pad}_h - \mathrm{kernel}_h）}{\mathrm{stride}_h} + 1$$

$$\mathrm{out}_w = \frac{（\mathrm{in}_w + 2 * \mathrm{pad}_w - \mathrm{kernel}_w）}{\mathrm{stride}_w} + 1$$

[0063] 其中，$(\mathrm{in}_w，\mathrm{in}_h)$表示输入特征图的尺寸，$(\mathrm{out}_w，\mathrm{out}_h)$和表示卷积层输出特征图的尺寸，$(\mathrm{kernel}_w，\mathrm{kernel}_h)$表示卷积核尺寸，$(\mathrm{stride}_w，\mathrm{stride}_h)$表示卷积核在宽和高在两个方向上的步长，$(\mathrm{pad}_w，\mathrm{pad}_h)$表示边缘补零的尺寸。

[0064] 所述特征图像卷积模块，若第 j 层的特征图 featurej 存在未处理的 block，将 featurej 的 block（X，Y）输入下一层网络进行运算；所述下一层网络为卷积层或最大池化层中任一种；若所述下一层网络为卷积层，则生成尺寸为 B 的 block，即尺寸为 B 的 featurej+1 的 block（X，Y）；若所述下一层网络为最大池化层，则生成尺寸为$\frac{1}{4}$B的 block，即尺寸为$\frac{1}{4}$B的 featurej+1 的 block（X，Y）；

[0065] 若 featurej 不存在未处理的 block，则对第 0 层特征图 feature0 的下一个 block（X，Y）进行特征图像卷积模块的功能，即令 X=NX，Y=NY，j=0，并转入特征图像卷积模块；

[0066] 若 feature0 不存在未处理的 block，则完成数据调度，最深层的特征图即为前向推理结果；

所述深度优先调用模块，配置为若第 j+1 层的特征图 featurej+1 的总尺寸$B_{j+1} < \mathrm{B}$，则对 featurej 的下一个 block（X，Y）重复特征图像卷积模块的功能，即令 X=NX，Y=XY 转入特征图像卷积模块；若第 j+1 层的特征图 featurej+1 的总尺寸$B_{j+1} = \mathrm{B}$，则将 featurej+1 的尺寸为 B 的特征图设置为 1 个 block 重复特征图像卷积模块的功能，即令 j=j+1 转入特征图像卷积模块。

[0067] 所属技术领域的技术人员可以清楚地了解到，为描述的方便和简洁，上述描述的系统的具体工作过程及有关说明，可以参考前述方法实施例中的对应过程，在此不再赘述。

[0068] 需要说明的是，上述实施例提供的基于分块卷积的深度优先数据调度系统，仅以上述各功能模块的划分进行举例说明，在实际应用中，可以根据需要而将上述功能分配由不同的功能模块来完成，即将本发明实施例中的模块或者步骤再分解或者组合，例如，上述实施例的模块可以合并为一个模块，也可以进一步拆分成多个子模块，以完成以上描述的全部或者部分功能。对于本发明实施例中涉及的模块、步骤的名称，仅仅是为了区分各个模块或者步骤，不视为对本发明的不当限定。

[0069] 本发明第三实施例的一种电子设备，其特征在于，包括：至少一个处理器；以及与至少一个所述处理器通信连接的存储器；其中，所述存储器存储有可被所述处理器执行的指令，所述指令用于被所述处理器执行以实现上述的基于分块卷积的深度优先数据调度方法。

[0070] 本发明第四实施例的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存储有计算机指令，所述计算机指令用于被所述计算机执行以实现上述的基于分块卷积的深度优先数据调度方法。

[0071] 所属技术领域的技术人员可以清楚地了解到，为描述的方便和简洁，上述描述的存储装置、处理装置的具体工作过程及有关说明，可以参考前述方法实施例中的对应过程，在此不再赘述。

[0072] 术语“第一”、“第二”等是用于区别类似的对象，而不是用于描述或表示特定的顺序或先后次序。

[0073] 术语“包括”或者任何其它类似用语旨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/装置不仅包括那些要素，而且还包括没有明确列出的其它要素，或者还包括这些过程、方法、物品或者设备/装置所固有的要素。

[0074] 至此，已经结合附图所示的优选实施方式描述了本发明的技术方案，但是，本领域技术人员容易理解的是，本发明的保护范围显然不局限于这些具体实施方式。在不偏离本发明的原理的前提下，本领域技术人员可以对相关技术特征做出等同的更改或替换，这些更改或替换之后的技术方案都将落入本发明的保护范围之内。

权　利　要　求　书

1、一种基于分块卷积的深度优先数据调度方法，其特征在于，所述方法包括：

步骤S100，将第0层特征图feature0分为m*n个预设尺寸为B的block，并设定坐标索引（X，Y），初始化（X，Y）=（0,0），特征图的层数j=0；

步骤S200，若第j层的特征图featurej存在未处理的block，将featurej的block（X，Y）输入下一层网络进行运算；所述下一层网络为卷积层或最大池化层中任一种；若所述下一层网络为卷积层，则生成尺寸为B的block，即尺寸为B的featurej+1的block（X，Y）；若所述下一层网络为最大池化层，则生成尺寸为$\frac{1}{4}$B的block，即尺寸为$\frac{1}{4}$B的featurej+1的block（X，Y）；

若featurej不存在未处理的block，则对第0层层特征图feature0的下一个block（X，Y）进行步骤S200的操作，即令X=NX，Y=NY，j=0，并转入步骤S200；

若feature0不存在未处理的block，则完成数据调度，最深层的特征图即为前向推理结果；

步骤S300，若第j+1层的特征图featurej+1的总尺寸$B_{j+1} < \mathrm{B}$，则对featurej的下一个block（X，Y）重复步骤S200的操作，即令X=NX，Y=XY转入步骤S200；若第j+1层的特征图featurej+1的总尺寸$B_{j+1} = \mathrm{B}$，则将featurej+1的尺寸为B的特征图设置为1个block重复步骤S200的操作，即令j=j+1转入步骤S200。

2、根据权利要求1所述的基于分块卷积的深度优先数据调度方法，其特征在于，还包括，将第0层特征图像feature0的尺寸设为$\mathrm{H} * \mathrm{W} = 2^{M*k} * 2^{N*\mathrm{k}}$，block大小预设为$\mathrm{B} = 2^{k} * 2^{k}$，B小于feature0的尺寸。

3、根据权利要求 1 所述的基于分块卷积的深度优先数据调度方法，其特征在于，所述 NX 和 NY，其获得方法包括：

X 配置为二进制表示的 m 个 bit 位，从低位至高位为x_0，x_1，x_2，……，x_{m-1}；Y 配置为二进制表示的 m 个 bit 位，从低位至高位为y_0，y_1，y_2，……，y_{m-1}；将坐标 X 和坐标 Y 的二进制表示进行二维混排生成二维混排坐标x_0，y_0，x_1，y_1，x_2，y_2，……，x_{m-1}，y_{m-1}，令所述二维混排坐标加 1，再进行二维解混排生成坐标 NX 和坐标 NY,每一层的 X 和 Y 相互独立计算，调取顺序均由 NX 和 NY 的获得方法生成。

4、根据权利要求 1 所述的基于分块卷积的深度优先数据调度方法，其特征在于，在特征图每次经过卷积层之前还包括：对各 block 进行边缘补零；

对于单通道的特征图，边缘补零的尺寸计算公式为：

$$\mathrm{out}_h = \frac{（\mathrm{in}_h + 2 * \mathrm{pad}_h - \mathrm{kernel}_h）}{\mathrm{stride}_h} + 1$$

$$\mathrm{out}_w = \frac{（\mathrm{in}_w + 2 * \mathrm{pad}_w - \mathrm{kernel}_w）}{\mathrm{stride}_w} + 1$$

其中，$(\mathrm{in}_w，\mathrm{in}_h)$表示输入特征图的尺寸，$(\mathrm{out}_w，\mathrm{out}_h)$和表示卷积层输出特征图的尺寸，$(\mathrm{kernel}_w，\mathrm{kernel}_h)$表示卷积核尺寸，$(\mathrm{stride}_w，\mathrm{stride}_h)$表示卷积核在宽和高在两个方向上的步长，$(\mathrm{pad}_w，\mathrm{pad}_h)$表示边缘补零的尺寸。

5、根据权利要求 1 所述的基于分块卷积的深度优先数据调度方法，其特征在于，每当 featurej 的总尺寸B_{j+1}=B 时，继续进行卷积或最大池化，不保存 featurej 的结果。

6、根据权利要求 1 所述的基于分块卷积的深度优先数据调度方法，其特征在于，本发明的方法只要数据准备完毕，便可以直接调用卷积核进行卷积。

7、一种基于分块卷积的深度优先数据调度系统，其特征在于，所述系统包括：图像划分模块、特征图像卷积模块和深度优先调用模块；

所述图像划分模块，将第 0 层特征图 feature0 分为 m*n 个预设尺寸为 B 的 block，并设定坐标索引（X，Y），初始化（X，Y）=（0,0），特征图的层数 j=0；

所述特征图像卷积模块，若第 j 层的特征图 featurej 存在未处理的 block，将 featurej 的 block（X，Y）输入下一层网络进行运算；所述下一层网络为卷积层或最大池化层中任一种；若所述下一层网络为卷积层，则生成尺寸为 B 的 block，即尺寸为 B 的 featurej+1 的 block（X，Y）；若所述下一层网络为最大池化层，则生成尺寸为$\frac{1}{4}$B的 block，即尺寸为$\frac{1}{4}$B 的 featurej+1 的 block（X，Y）；

若 featurej 不存在未处理的 block，则对第 0 层特征图 feature0 的下一个 block（X，Y）进行特征图像卷积模块的功能，即令 X=NX，Y=NY，j=0，并转入特征图像卷积模块；

若 feature0 不存在未处理的 block，则完成数据调度，最深层的特征图即为前向推理结果；

所述深度优先调用模块，配置为若第 j+1 层的特征图 featurej+1 的总尺寸$B_{j+1} < \mathrm{B}$，则对 featurej 的下一个 block（X，Y）重复特征图像卷积模块的功能，即令 X=NX，Y=XY 转入特征图像卷积模块；若第 j+1 层的特征图 featurej+1 的总尺寸$B_{j+1} = \mathrm{B}$，则将 featurej+1 的尺寸为 B 的特征图设置为 1 个 block 重复特征图像卷积模块的功能，即令 j=j+1 转入特征图像卷积模块。

8、根据权利要求 7 所述的基于分块卷积的深度优先数据调度系统，其特征在于，所述系统还包括图像补零模块，配置为在特征图每次经过卷积层之前将feature0分为 m*n 个 block 之后，将各 block 进行边缘补零；

对于单通道的特征图，边缘补零的尺寸计算公式为：

$$\mathrm{out}_h = \frac{（\mathrm{in}_h + 2 * \mathrm{pad}_h - \mathrm{kernel}_h）}{\mathrm{stride}_h} + 1$$

$$\mathrm{out}_w = \frac{（\mathrm{in}_w + 2 * \mathrm{pad}_w - \mathrm{kernel}_w）}{\mathrm{stride}_w} + 1$$

其中，$(\mathrm{in}_w，\mathrm{in}_h)$表示输入特征图的尺寸，$(\mathrm{out}_w，\mathrm{out}_h)$和表示卷积层输出特征图的尺寸，$(\mathrm{kernel}_w，\mathrm{kernel}_h)$表示卷积核尺寸，$(\mathrm{stride}_w，\mathrm{stride}_h)$表示卷积核在宽和高在两个方向上的步长，$(\mathrm{pad}_w，\mathrm{pad}_h)$表示边缘补零的尺寸。

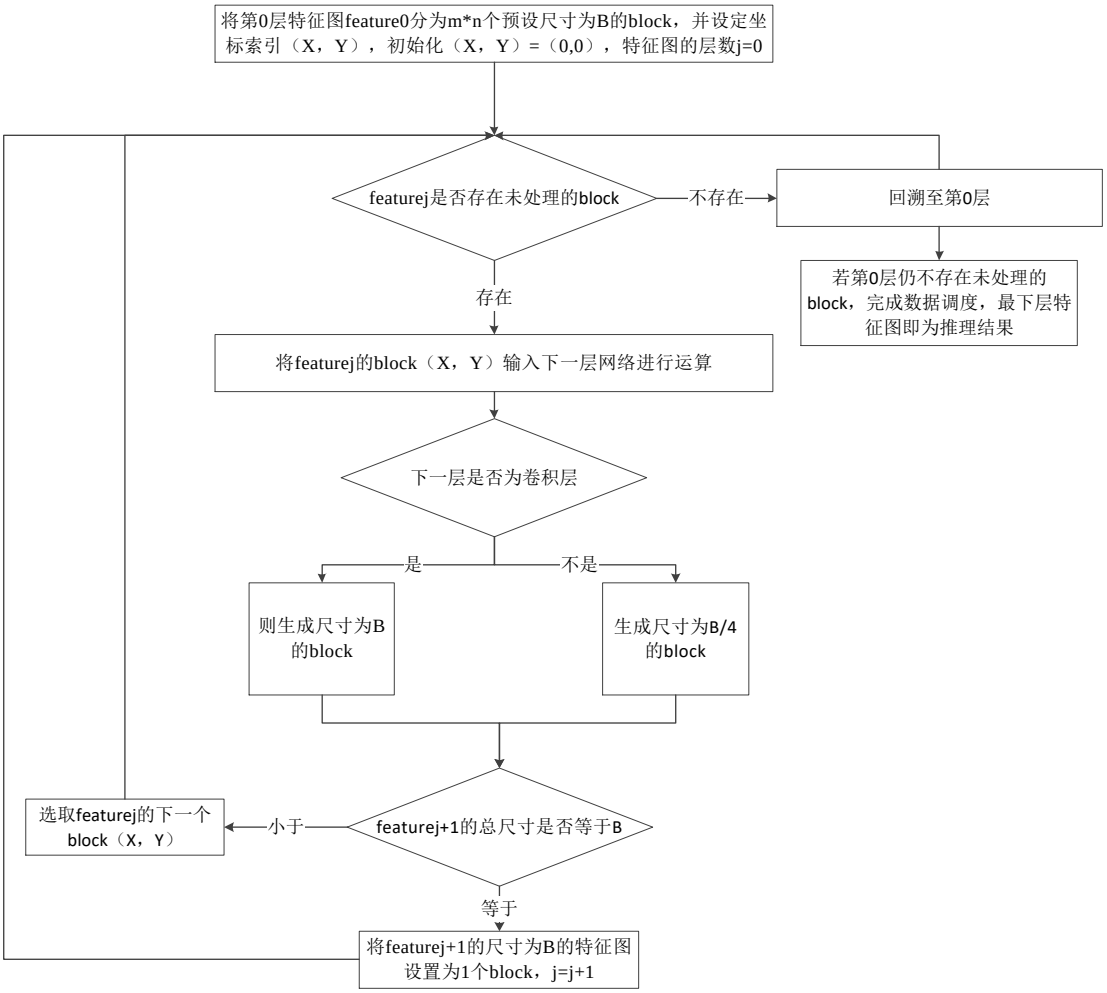
9、一种电子设备，其特征在于，包括：至少一个处理器；以及与至少一个所述处理器通信连接的存储器；其中，所述存储器存储有可被所述处理器执行的指令，所述指令用于被所述处理器执行以实现权利要求1-6任一项所述的基于分块卷积的深度优先数据调度方法。

10、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存储有计算机指令，所述计算机指令用于被所述计算机执行以实现权利要求1-6任一项所述的基于分块卷积的深度优先数据调度方法。

说　明　书　摘　要

本发明属于卷积神经网络领域，具体涉及了一种基于分块卷积的深度优先数据调度方法、系统及设备，旨在解决现有的卷积模型计算方法需要逐层进行计算，保存中间结果特征图需要占用大量内存而不宜部署在全硬件设备中的问题。本发明包括：将输入的特征图像分为多个block，逐个调用各block进行卷积或最大池化生成下一层特征图，若下一层特征图达到预设block大小继续向下一层调用获得更深层的特征图，若下一层特征图小于预设block大小则返回第0层进行调用，直至完成推理过程。本发明避免了存储大量卷积层中间结果所带来的内存消耗，提高卷积模型在全硬件设备上的推理效率。

摘　要　附　图


将第0层特征图feature0分为m*n个预设尺寸为B的block，并设定坐标索引（X，Y），初始化（X，Y）=（0,0），特征图的层数j=0
featurej是否存在未处理的block
不存在
回溯至第0层
若第0层仍不存在未处理的block，完成数据调度，最下层特征图即为推理结果
存在
将featurej的block（X，Y）输入下一层网络进行运算
下一层是否为卷积层
是
不是
则生成尺寸为B的block
生成尺寸为B/4的block
选取featurej的下一个block（X，Y）
小于
featurej+1的总尺寸是否等于B
等于
将featurej+1的尺寸为B的特征图设置为1个block，j=j+1

说　明　书　附　图

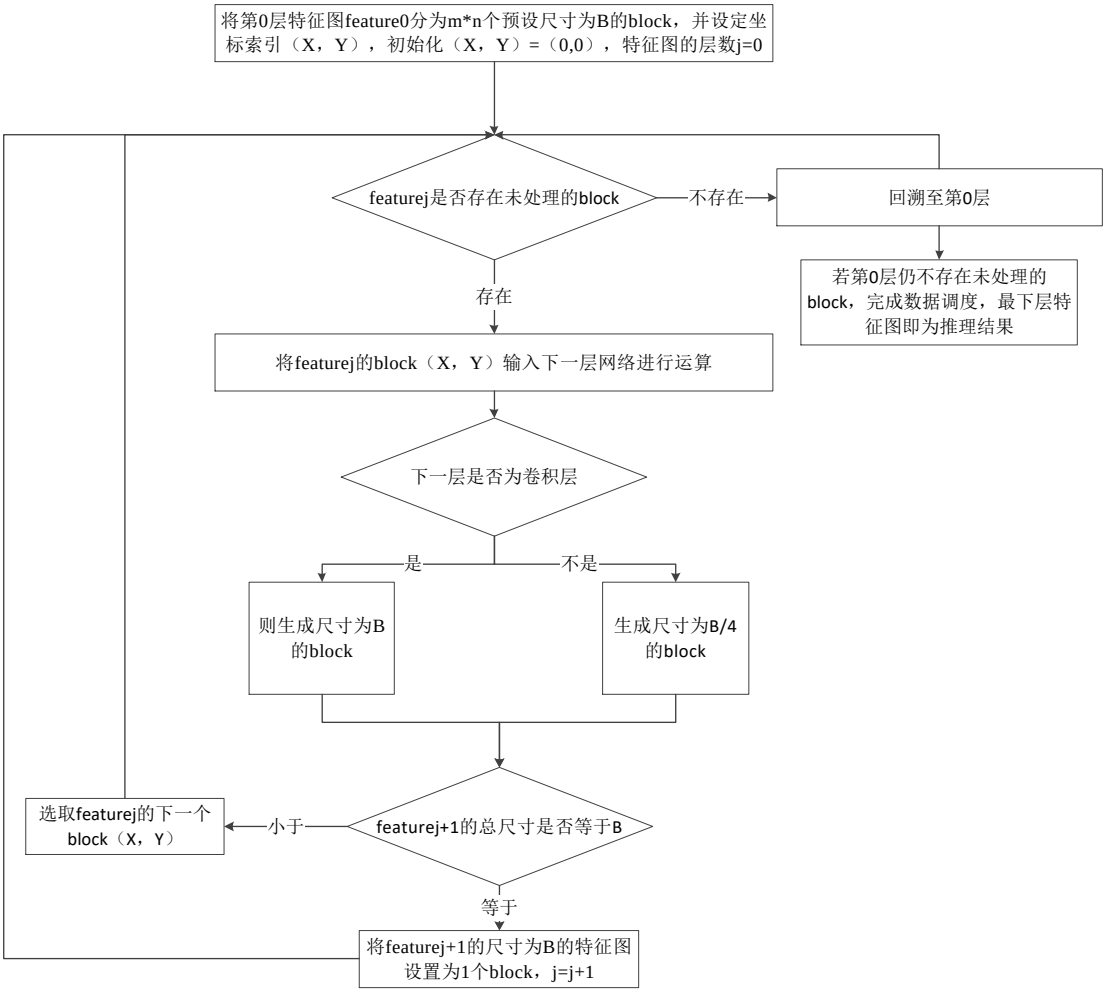

将第0层特征图feature0分为m*n个预设尺寸为B的block，并设定坐标索引（X，Y），初始化（X，Y）=（0,0），特征图的层数j=0
featurej是否存在未处理的block
不存在
回溯至第0层
若第0层仍不存在未处理的block，完成数据调度，最下层特征图即为推理结果
存在
将featurej的block（X，Y）输入下一层网络进行运算
下一层是否为卷积层
是
不是
则生成尺寸为B的block
生成尺寸为B/4的block
选取featurej的下一个block（X，Y）
小于
featurej+1的总尺寸是否等于B
等于
将featurej+1的尺寸为B的特征图设置为1个block，j=j+1

图1

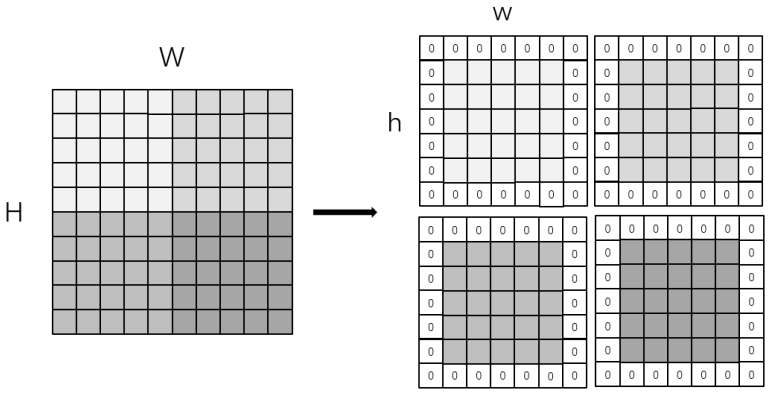

W
H
w
h

图2

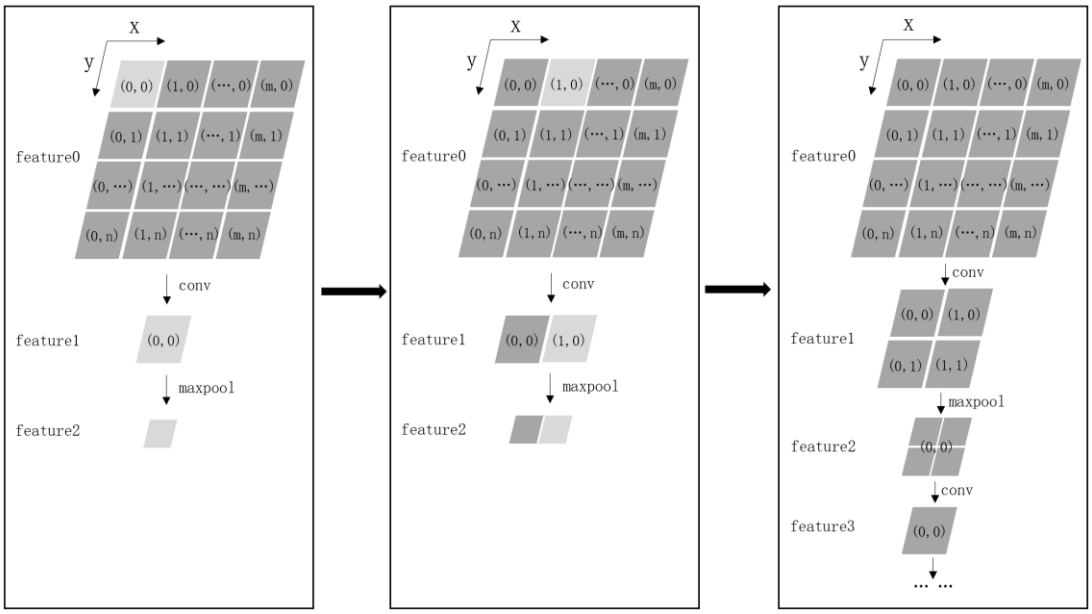

x
y
(0, 0)
(1, 0)
(…, 0)
(m, 0)
(0, 1)
(1, 1)
(…, 1)
(m, 1)
(0, …)
(1, …)
(…, …)
(m, …)
(0, n)
(1, n)
(…, n)
(m, n)
feature0
conv
feature1
maxpool
feature2
feature3
… …

图3

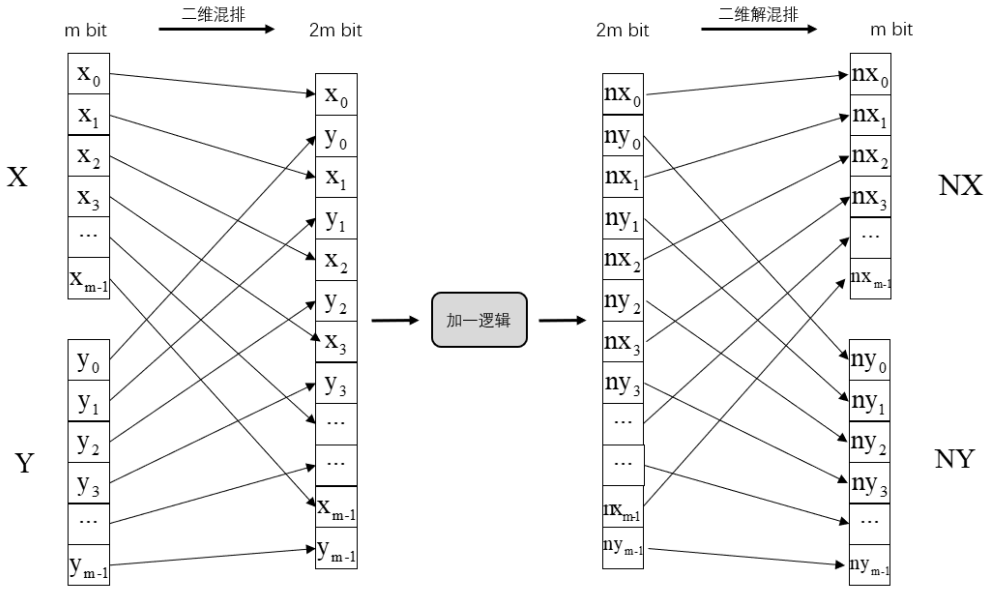

m bit
二维混排
2m bit
X
Y
加一逻辑
2m bit
二维解混排
m bit
NX
NY

图4