(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
2019 年 10 月 17 日 (17.10.2019)

WIPO | PCT

(10) 国际公布号
WO 2019/196223 A1

(51) 国际专利分类号:
G06N 3/04 (2006.01)

(21) 国际申请号: PCT/CN2018/095365

(22) 国际申请日: 2018 年 7 月 12 日 (12.07.2018)

(25) 申请语言: 中文

(26) 公布语言: 中文

(30) 优先权:
201810306577.3 2018年4月8日 (08.04.2018) CN

(71) 申请人: 清 华 大 学 (TSINGHUA UNIVERSITY) [CN/CN]; 中 国 北 京 市 海 淀 区 清 华 大 学, Beijing 100084 (CN)。

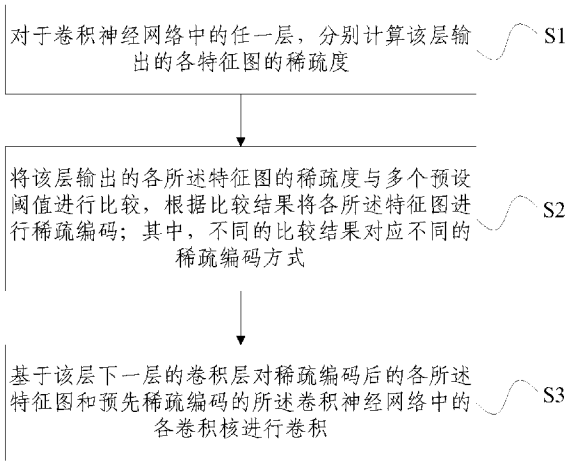
(72) 发明人: 刘勇攀(LIU, Yongpan); 中国北京市海淀区清华大学, Beijing 100084 (CN)。 袁哲(YUAN, Zhe); 中国北京市海淀区清华大学, Beijing 100084 (CN)。 岳金山(YUE, Jinshan); 中国北京市海淀区清华大学, Beijing 100084 (CN)。 杨华中(YANG, Huazhong); 中国北京市海淀区清华大学, Beijing 100084 (CN)。 李学清(LI, Xueqing); 中国北京市海淀区清华大学, Beijing 100084 (CN)。 王智博(WANG, Zhibo); 中国北京市海淀区清华大学, Beijing 100084 (CN)。

(74) 代理人: 北京路浩知识产权代理有限公司(CN-KNOWHOW INTELLECTUAL PROPERTY AGENT LIMITED); 中国北京市海淀区丹棱街3号中国电子大厦B座18层, Beijing 100080 (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, 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,

(54) **Title:** ACCELERATION METHOD AND ACCELERATOR USED FOR CONVOLUTIONAL NEURAL NETWORK

(54) **发明名称:** 一种应用于卷积神经网络的加速方法和加速器



S1 FOR ANY LAYER OF THE CONVOLUTIONAL NEURAL NETWORK, CALCULATE THE DENSITY OF VARIOUS FEATURE MAPS OUTPUT BY THIS LAYER

S2 COMPARE THE DENSITY OF THE VARIOUS FEATURE MAPS OUTPUT BY THIS LAYER TO A PLURALITY OF PRESET THRESHOLDS, AND PERFORM, ACCORDING TO THE COMPARISON RESULTS, SPARSE CODING TO THE VARIOUS FEATURE MAPS, DIFFERENT COMPARISON RESULTS CORRESPONDING TO DIFFERENT SPARSE CODING SCHEMES

S3 PERFORM, ON THE BASIS OF THE CONVOLUTION LAYER OF A LAYER FOLLOWING THIS LAYER, CONVOLUTION TO THE VARIOUS SPARSE CODED FEATURE MAPS AND VARIOUS PRE-SPARSE CODED CONVOLUTION KERNELS IN THE CONVOLUTIONAL NEURAL NETWORK

图 1

(57) **Abstract:** The present invention provides an acceleration method and accelerator used for a convolutional neural network, the method comprising: S1, for any layer of the convolutional neural network, calculating the density of various feature maps output by this layer; S2, comparing the density of the various feature maps output by this layer to a plurality of preset thresholds, and performing, according to the comparison results, sparse coding to the various feature maps, different comparison results corresponding to different sparse coding schemes; and S3, performing, on the basis of the convolution layer of a layer following this layer, convolution to the various sparse coded feature maps and various pre-sparse coded convolution kernels in the convolutional neural network. The present invention reduces the calculated amount of the convolution operation in the convolutional neural network and improves operation speed.

WO 2019/196223 A1

[见续页]

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))。

(57) 摘要：本发明提供一种应用于卷积神经网络的加速方法和加速器，所述方法包括：S1，对于卷积神经网络中的任一层，分别计算该层输出的各特征图的稠密度；S2，将该层输出的各所述特征图的稠密度与多个预设阈值进行比较，根据比较结果将各所述特征图进行稀疏编码；其中，不同的比较结果对应不同的稀疏编码方式；S3，基于该层下一层的卷积层对稀疏编码后的各所述特征图和预先稀疏编码的所述卷积神经网络中的各卷积核进行卷积。本发明减少卷积神经网络中卷积运算的计算量，提高了运算速度。

一种应用于卷积神经网络的加速方法和加速器

交叉引用

本申请引用于2018年04月08日提交的专利名称为“一种应用于卷积神经网络的加速方法和加速器”的第2018103065773号中国专利申请，其通过引用被全部并入本申请。

技术领域

本发明属于运算优化技术领域，更具体地，涉及一种应用于卷积神经网络的加速方法和加速器。

背景技术

卷积神经网络（Convolutional Neural Network, CNN）是一种前馈神经网络，其人工神经元可以响应一部分覆盖范围内的周围单元，适用于对大型图像的处理。卷积神经网络广泛应用于图像识别、语音识别等领域，但计算量非常大。

由于卷积神经网络中的激活函数ReLU（Rectified linear unit，修正线性单元）会造成大量稀疏的特征图（feature map）；同时，采用剪枝等方法训练卷积神经网络会造成大量稀疏的权重数据（weight data）。利用特征图和权重数据的稀疏性可以大幅提高卷积神经网络的计算效率。目前，已有很多方法基于卷积神经网络中特征图和权重数据的稀疏性提高计算速度。这些方法大致可以分为两类，一类着眼于跳过0值。例如有的方法去除输入中的0值，从而减少输入为0的无效计算。另一类采取忽略零值的方法。例如有的方法在输入数据为0时，不执行乘法操作，从而减少运算。但这些方法都着眼于处理稀疏神经网络本身，假定神经网络稀疏是前提。然而实际上卷积神经网络中各层输出特征图可能是稀疏的，可能是非稀疏的。实际运用中卷积神经网络各层的权重数据和特征图的稠密度一般在5%-90%之间分布。

稀疏矩阵是指数值为0的元素数目远远多于非0元素的数目，并且非0元素分布没有规律的矩阵。现有技术中一方面只能对稀疏的卷积神经网

络进行处理，在卷积神经网络不是稀疏的情况下计算量很大，运算速度低；另一方面现有技术只能处理卷积神经中权重数据或特征图是稀疏的情况，不能处理权重数据和特征图都是稀疏的情况。

发明内容

为克服上述卷积神经网络运算速度低的问题或者至少部分地解决上述问题，本发明提供了一种应用于卷积神经网络的加速方法和加速器。

根据本发明的第一方面，提供一种应用于卷积神经网络的加速方法，包括：

S1，对于卷积神经网络中的任一层，分别计算该层输出的各特征图的稠密度；

S2，将该层输出的各所述特征图的稠密度与多个预设阈值进行比较，根据比较结果将各所述特征图进行稀疏编码；其中，不同的比较结果对应不同的稀疏编码方式；

S3，基于该层下一层的卷积层对稀疏编码后的各所述特征图和预先稀疏编码的所述卷积神经网络中的各卷积核进行卷积。

具体地，所述步骤 S1 具体包括：

对于任一所述特征图，统计该特征图中非 0 元素的个数和该特征图中所有元素的总个数；

将该特征图中非 0 元素的个数与该特征图中所有元素的总个数之间的比值作为该特征图的稠密度。

具体地，所述预设阈值包括第一预设阈值和第二预设阈值；其中，所述第一预设阈值小于所述第二预设阈值；

相应地，所述步骤 S2 具体包括：

若各所述特征图的稠密度小于所述第一预设阈值，则将各所述特征图编码为稀疏矩阵存储格式；

若各所述特征图的稠密度大于或等于所述第一预设阈值，且小于所述第二预设阈值，则将各所述特征图中的 0 元素进行标记；

若各所述特征图的稠密度大于或等于所述第二预设阈值，则不对各所述特征图进行稀疏编码。

具体地，所述步骤 S3 之前还包括：

计算训练好的卷积网络中各卷积核的稠密度；

若各所述卷积核的稠密度小于所述第一预设阈值，则将各所述卷积核编码为稀疏矩阵存储格式；

若各所述卷积核的稠密度大于或等于所述第一预设阈值，且小于所述第二预设阈值，则将各所述卷积核中的 0 元素进行标记；

若各所述卷积核的稠密度大于或等于所述第二预设阈值，则不对各所述卷积核进行稀疏编码。

具体地，所述步骤 S3 具体包括：

当各所述特征图或各所述卷积核中存在所述标记时，对各所述特征图或各所述卷积核中标记对应的元素不进行计算。

根据本发明另一方面提供一种应用于卷积神经网络的加速器，包括：神经网络计算阵列模块和动态稀疏调整模块；

其中，所述动态稀疏调整模块用于计算卷积神经网络各层输出的各特征图的稠密度，将各所述特征图的稠密度与多个预设阈值进行比较，根据比较结果对各所述特征图进行稀疏编码；其中，不同的比较结果对应的不同稀疏编码方式；

所述神经网络计算阵列模块用于对稀疏编码后的各所述特征图和预先稀疏编码的所述卷积神经网络中的各卷积核进行卷积操作。

具体地，所述动态稀疏调整模块包括线上稠密度识别模块、输出临时寄存模块、动态编码模块和动态稀疏控制模块；

其中，所述线上稠密度识别模块用于对于任一所述特征图，统计该特征图中 0 元素的个数和该特征图中所有元素的总个数；将该特征图中 0 元素的个数与该特征图中所有元素的总个数之间的比值作为该特征图的稠密度；

所述输出临时寄存模块用于存储卷积神经网络中各层输出的各所述特征图；

所述动态稀疏控制模块用于将所述线上稠密度识别模块输出的各所述特征图的稠密度与多个预设阈值进行比较；

所述动态编码模块用于根据比较结果将所述输出临时寄存模块中各所述特征图进行稀疏编码。

具体地，所述预设阈值包括第一预设阈值和第二预设阈值；其中，所述第一预设阈值小于所述第二预设阈值；

相应地，所述动态编码模块具体用于：

若各所述特征图的稠密度小于所述第一预设阈值，则将各所述特征图编码为稀疏矩阵存储格式；

若各所述特征图的稠密度大于或等于所述第一预设阈值，且小于所述第二预设阈值，则将各所述特征图中的 0 元素进行标记；

若各所述特征图的稠密度大于或等于所述第二预设阈值，则不对各所述特征图进行稀疏编码。

具体地，所述动态编码模块还用于：

若预先计算的各所述卷积核的稠密度小于所述第一预设阈值，则将各所述卷积核编码为稀疏矩阵存储格式；

若各所述卷积核的稠密度大于或等于所述第一预设阈值，且小于所述第二预设阈值，则将各所述卷积核中的 0 元素进行标记；

若各所述卷积核的稠密度大于或等于所述第二预设阈值，则不对各所述卷积核进行稀疏编码。

具体地，所述神经网络计算阵列模块具体用于：

当各所述特征图或各所述卷积核中存在所述标记时，对各所述特征图或各所述卷积核中标记对应的元素不进行计算。

本发明提供一种应用于卷积神经网络的加速方法和加速器，该方法通过将卷积神经网络中各层输出的各特征图的稠密度与多个预设阈值进行比较，获取各所述特征图的稀疏状态，将不同稀疏状态的特征图进行不同方式的稀疏编码，然后基于各层下一层的卷积层对稀疏编码后的各特征图和预先稀疏编码的卷积神经网络中的卷积核进行卷积操作，减少卷积神经网络中卷积运算的计算量，提高运算速度。

附图说明

图 1 为本发明实施例提供的应用于卷积神经网络的加速方法整体流程示意图；

图 2 为本发明实施例提供的应用于卷积神经网络的加速器整体结构示意图；

图 3 为本发明实施例提供的应用于卷积神经网络的加速器中极限能量效率测试结果示意图;

图 4 为本发明实施例提供的应用于卷积神经网络的加速器中极限能量效率测试结果对比示意图。

具体实施方式

下面结合附图和实施例，对本发明的具体实施方式作进一步详细描述。以下实施例用于说明本发明，但不用来限制本发明的范围。

在本发明的一个实施例中提供一种应用于卷积神经网络的加速方法，图 1 为本发明实施例提供的应用于卷积神经网络的加速方法整体流程示意图，该方法包括:

S1，对于卷积神经网络中的任一层，分别计算该层输出的各特征图的稠密度;

具体地，所述卷积神经网络中可以包括池化层或不包括池化层。首先对所述卷积神经网络进行训练，训练完成后卷积神经网络中的卷积核不再变动，因此卷积神经网络中的卷积核不需要线上动态稀疏编码，直接线下稀疏编码一次即可。所述线上是指位于加速器的芯片上，所述线下是指不位于所述加速器的芯片上。在每次卷积运算时直接读取稀疏编码的卷积核进行卷积计算。当输入原始图像数据时，对原始图像数据进行稀疏编码，然后将稀疏编码的原始数据和稀疏编码的卷积核输入所述卷积神经网络的第一层卷积层进行卷积计算。由于原始图像数据一般不是稀疏的，也可以不对所述原始图像数据进行稀疏编码，直接输入所述原始图像数据。所述稀疏编码是将数据以稀疏格式存储。

S1 中，由于卷积神经网络中每层输出的各特征图的稠密度不一样，不同层输出的特征图也是动态变化的，因此稠密度也是动态变化的。所述稠密度表示各所述特征图的稀疏程度。为了更好地提高所述卷积神经网络的运算速度，计算每一层输出的各特征图的稠密度，以根据每一层输出的各特征图的稠密度对每一层输出的各特征图进行稀疏编码。

S2，将该层输出的各所述特征图的稠密度与多个预设阈值进行比较，根据比较结果将各所述特征图进行稀疏编码; 其中，不同的比较结果对应不同的稀疏编码方式;

S2 中，现有技术中将各层输出的特征图全部进行稀疏编码，计算量大。本实施例根据所述预设阈值获取该层输出的各所述特征图的稀疏状态。从而将不同稀疏状态的特征图进行不同形式的稀疏编码。

S3，基于该层下一层的卷积层对稀疏编码后的各所述特征图和预先稀疏编码的所述卷积神经网络中的各卷积核进行卷积。

S3 中，将稀疏编码后的各所述特征图和预先稀疏编码的所述卷积神经网络中的各卷积核作为该层下一层的卷积层的输入，进行卷积操作。然后，将所述卷积操作的结果作为所述卷积层下一层的输入，将所述卷积层下一层输出的特征图继续执行上述稀疏编码和卷积操作，直到所述卷积神经网络中最后一层输出各特征图为止。本实施例中不限于卷积核的稀疏编码方式。

本实施例将卷积神经网络中各层输出的各特征图的稠密度与多个预设阈值进行比较，获取各所述特征图的稀疏状态，将不同稀疏状态的特征图进行不同方式的稀疏编码，然后基于各层下一层的卷积层对稀疏编码后的各特征图和预先稀疏编码的卷积神经网络中的卷积核进行卷积操作，减少卷积神经网络中卷积运算的计算量，提高运算速度。

在上述实施例的基础上，本实施例中，所述步骤 S1 具体包括：对于任一所述特征图，统计该特征图中非 0 元素的个数和该特征图中所有元素的总个数；将该特征图中非 0 元素的个数与该特征图中所有元素的总个数之间的比值作为该特征图的稠密度。

具体地，各特征图的稠密度为各特征图中非 0 元素的个数与各特征图中所有元素的总个数之间的比值。例如，一个特征图中非 0 元素的个数为 10，该特征图中所有元素的总个数为 100，则该特征图的稠密度为 0.1。

在上述实施例的基础上，本实施例中所述预设阈值包括第一预设阈值和第二预设阈值；所述预设阈值包括第一预设阈值和第二预设阈值；相应地，所述步骤 S2 具体包括：若各所述特征图的稠密度小于所述第一预设阈值，则将各所述特征图编码为稀疏矩阵存储格式；若各所述特征图的稠密度大于或等于所述第一预设阈值，且小于所述第二预设阈值，则将各所述特征图中的 0 元素进行标记；若各所述特征图的稠密度大于或等于所述第二预设阈值，则不对各所述特征图进行稀疏编码。

具体地，本实施例中所述预设阈值包括第一预设阈值 th1 和第二预设阈值 th2。根据所述第一预设阈值和所述第二预设阈值将各所述特征图的特征状态 AS 分为三种状态，即将稠密度小于所述第一预设阈值的特征图分为完全稀疏状态 S，将稠密度大于或等于所述第一预设阈值，且小于所述第二预设阈值的特征图分为中等稀疏状态 M，将稠密度大于或等于所述第二预设阈值的特征图分为完全非稀疏状态 D。若各特征图为稀疏状态 S，则将各所述特征图编码为稀疏矩阵存储格式，所述稀疏矩阵存储格式包括各所述特征图中的非 0 数据 activ 和稀疏索引 index，例如坐标编码和压缩稀疏行编码。通过将特征图编码为稀疏矩阵存储格式，可以节省大量存储空间，同时节省大量计算时间。若各特征图为中等稀疏状态 M，则将各特征图中的 0 元素添加标记 guard，所述标记用于标识 0 元素。对于标记的元素可以不参与计算和存储，从而降低功耗。将各特征图中的 0 元素进行标记也是一种稀疏编码方式。若各特征图为完全非稀疏状态 D，则不需要动态编码，直接输出各特征图的非稀疏数据。

在上述实施例的基础上，本实施例中所述步骤 S3 之前还包括：计算训练好的卷积网络中各卷积核的稠密度；若各所述卷积核的稠密度小于所述第一预设阈值，则将各所述卷积核编码为稀疏矩阵存储格式；若各所述卷积核的稠密度大于或等于所述第一预设阈值，且小于所述第二预设阈值，则将各所述卷积核中的 0 元素进行标记；若各所述卷积核的稠密度大于或等于所述第二预设阈值，则不对各所述卷积核进行编码。

具体地，各卷积核的稠密度为各卷积核中非 0 元素的个数与各卷积核中所有元素的总个数之间的比值。将各卷积核的状态 WS 与特征图一样分为三种状态。每种状态对应不同的稀疏编码方式。由于特征图和卷积核各存在三种状态，组合后共有 9 种状态，从而对卷积神经网络的稠密度进行更细粒度的划分。

在上述各实施例的基础上，本实施例中所述步骤 S3 具体包括：当各所述特征图或各所述卷积核中存在所述标记时，对各所述特征图或各所述卷积核中标记对应的元素不进行计算。

具体地，当各所述特征图或各所述卷积核为完全稀疏状态 S 时，在输入前去除 0，减少存储空间，同时不用对 0 元素进行计算；当各所述特征

图或各所述卷积核为中等稀疏状态 M 时，各所述特征图或各所述卷积核中的 0 元素虽然存储，但对标记对应的元素不进行计算，从而减少计算。

在本发明的另一个实施例中提供一种应用于卷积神经网络的加速器，图 2 为本发明实施例提供的应用于卷积神经网络的加速器整体结构示意图，包括：神经网络计算阵列模块和动态稀疏调整模块；其中，所述动态稀疏调整模块用于计算卷积神经网络中各层输出的各特征图的稠密度，将各所述特征图的稠密度与多个预设阈值进行比较，根据比较结果对各所述特征图进行稀疏编码；其中，不同的比较结果对应的不同稀疏编码方式；所述神经网络计算阵列模块用于对稀疏编码后的各所述特征图和预先稀疏编码的所述卷积神经网络中的各卷积核进行卷积操作。

具体地，所述卷积神经网络中可以包括池化层或不包括池化层。首先对卷积神经网络进行训练，训练完成后卷积神经网络中的卷积核不再变动，因此卷积神经网络中的卷积核不需要线上动态稀疏编码，直接线下稀疏编码一次即可。在每次卷积运算时所述神经网络计算阵列模块直接读取线下稀疏编码的卷积核进行卷积计算。当卷积神经网络输入原始图像数据时，动态稀疏调整模块对原始图像数据进行稀疏编码，然后神经网络计算阵列模块根据稀疏编码的原始数据和稀疏编码的卷积核进行卷积计算。由于原始图像数据一般不是稀疏的，也可以不对所述原始图像数据进行稀疏编码，直接输入所述原始图像数据。所述稀疏编码是将数据以稀疏格式存储。

由于卷积神经网络中各层输出的各特征图的稠密度不一样，不同层输出的特征图也是动态变化的，因此稠密度也是动态变化的。所述稠密度表示各所述特征图的稀疏程度。为了更好地提高所述卷积神经网络的运算速度，所述动态稀疏调整模块计算每一层输出的各特征图的稠密度，以根据每一层输出的各特征图的稠密度对每一层输出的各特征图进行稀疏编码。

所述动态稀疏调整模块根据多个预设阈值获取该层输出的各所述特征图的稀疏状态。从而将不同稀疏状态的特征图进行不同形式的稀疏编码，而不仅限于一种稀疏编码。现有技术中将各层输出的特证图全部进行稀疏编码，计算量大。

所述神经网络计算阵列模块根据稀疏编码后的各所述特征图和预先

稀疏编码的所述卷积神经网络中的各卷积核进行卷积操作。如果包括池化模块，所述池化模块将所述卷积操作的结果进行池化操作。此外，所述加速器还包括中间数据存储器模块、主芯片控制器和芯片上下数据交流模块。其中，所述主控制器控制加速器整个芯片的运行动作与时序。芯片上下数据交流模块用于从所述芯片外部存储器读取数据或将芯片计算好的数据写入外部存储。例如，初始化后，芯片在所述主控制器的控制下通过所述芯片上下数据交流模块从外部存储器读取原始图像数据和初始卷积核。所述中间数据存储器模块用于存储所述神经网络计算阵列模块计算过程中的中间结果。

本实施例动态稀疏调整模块将卷积神经网络各层输出的各特征图的稠密度与多个预设阈值进行比较，获取各所述特征图的稀疏状态，将不同稀疏状态的特征图进行不同方式的稀疏编码，以供神经网络计算阵列模块对稀疏编码后的各特征图和预先稀疏编码的卷积神经网络中的卷积核进行卷积操作，一方面，减少卷积神经网络中卷积运算的计算量，提高运算速度；另一方面，根据稀疏状态的不同，动态切换加速器的处理状态，提高了加速器的灵活性。

在上述实施例的基础上，本实施例中所述动态稀疏调整模块包括线上稠密度识别模块、输出临时寄存模块、动态编码模块和动态稀疏控制模块；其中，所述线上稠密度识别模块用于对于任一所述特征图，统计该特征图中 0 元素的个数和该特征图中所有元素的总个数；将该特征图中 0 元素的个数与该特征图中所有元素的总个数之间的比值作为该特征图的稠密度；所述输出临时寄存模块用于存储卷积神经网络中各层输出的各所述特征图；所述动态稀疏控制模块用于将所述线上稠密度识别模块输出的各所述特征图的稠密度与多个预设阈值进行比较；所述动态编码模块用于根据比较结果将所述输出临时寄存模块中各所述特征图进行稀疏编码。

具体地，所述动态稀疏调整模块具体包括四个模块。所述线上稠密度识别模块用于统计计算过程中各特征图中非 0 元素的个数，以计算各特征图的稠密度。所述输出临时寄存模块用于将卷积神经网络中各层输出的特征图以非稀疏格式进行暂存。所述动态稀疏控制模块用于通过预设的多个预设阈值控制所述特征图的稀疏状态。所述动态编码模块根据各所述特征

图的稀疏状态将所述输出临时寄存模块中各所述特征图进行稀疏编码，从而提高卷积运算的速度。

在上述实施例的基础上，本实施例中所述预设阈值包括第一预设阈值和第二预设阈值；所述预设阈值包括第一预设阈值和第二预设阈值；相应地，所述动态编码模块具体用于：若各所述特征图的稠密度小于所述第一预设阈值，则将各所述特征图编码为稀疏矩阵存储格式；若各所述特征图的稠密度大于或等于所述第一预设阈值，且小于所述第二预设阈值，则将各所述特征图中的 0 元素进行标记；若各所述特征图的稠密度大于或等于所述第二预设阈值，则不对各所述特征图进行编码。

具体地，本实施例中所述预设阈值包括第一预设阈值 th1 和第二预设阈值 th2。所述动态稀疏控制模块根据所述第一预设阈值和所述第二预设阈值将各所述特征图的特征状态 AS 分为三种状态，即将稠密度小于所述第一预设阈值的特征图分为完全稀疏状态 S，将稠密度大于或等于所述第一预设阈值，且小于所述第二预设阈值的特征图分为中等稀疏状态 M，将稠密度大于或等于所述第二预设阈值的特征图分为完全非稀疏状态 D。

若各特征图为稀疏状态 S，则所述动态编码模块将所述输出临时寄存模块中各所述特征图编码为稀疏矩阵存储格式，所述稀疏矩阵存储格式包括各所述特征图中的非 0 数据 activ 和稀疏索引 index，例如坐标编码和压缩稀疏行编码。通过将特征图编码为稀疏矩阵存储格式，可以节省大量存储空间，同时节省大量计算时间。若各特征图为中等稀疏状态 M，则所述动态编码模块将所述输出临时寄存模块中各特征图中的 0 元素添加标记 guard，对于标记的元素可以不参与计算和存储，从而降低功耗。若各特征图为完全非稀疏状态 D，则不需要动态编码，所述动态编码模块直接输出各特征图的非稀疏数据。

在上述实施例的基础上，本实施例中所述动态编码模块还用于：若预先计算的各所述卷积核的稠密度小于所述第一预设阈值，则将各所述卷积核编码为稀疏矩阵存储格式；若各所述卷积核的稠密度大于或等于所述第一预设阈值，且小于所述第二预设阈值，则将各所述卷积核中的 0 元素进行标记；若各所述卷积核的稠密度大于或等于所述第二预设阈值，则不对各所述卷积核进行编码。

具体地，各卷积核的稠密度为各卷积核中非 0 元素的个数与各卷积核中所有元素的总个数之间的比值。各卷积核的状态 WS 与特征图一样有三种状态。每种状态对应不同的稀疏编码方式。由于特征图和卷积核各存在三种状态，组合后共有 9 种状态，从而对卷积神经网络的稠密度进行更细粒度的划分。

在上述各实施例的基础上，本实施例中所述神经网络计算阵列模块具体用于：当各所述特征图或各所述卷积核中存在所述标记时，对各所述特征图或各所述卷积核中标记对应的元素不进行计算。

具体地，当各所述特征图或各所述卷积核为完全稀疏状态 S 时，在将各所述特征图或各所述卷积核输入所述神经网络计算阵列模块前去除 0，减少存储空间，同时不用对 0 元素进行计算；当各所述特征图或各所述卷积核为中等稀疏状态 M 时，各所述特征图或各所述卷积核中的 0 元素虽然存储，但对标记对应的元素不进行计算，从而减少计算。

例如，采用台积电 65nm 工艺制作所述加速器的芯片，所述芯片的面积为 3mm*4mm，运行频率为 20-200MHz，功耗为 20.5-248.4 毫瓦。本实施例中极限能量效率会随着特征图和卷积核稠密度的下降而快速上升，如图 3 所示。当特征图和卷积核的稠密度均为 5%时，极限能量效率可达到 62.1TOPS/W，是不采用本实施例加速器时极限能量效率的 6.2 倍。如图 4 所示，相比于只支持特征数据稀疏的实现，本实施例能量效率可提升 4.3 倍。相比于无自适应稀疏控制的实现，本发明能量效率可提升 2.8 倍。相比于无稠密度控制但量化精度可变的实现，本发明能量效率可提升 2 倍。

最后，本申请的方法仅为较佳的实施方案，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，均应包含在本发明的保护范围之内。

权 利 要 求 书

1、一种应用于卷积神经网络的加速方法，其特征在于，包括:

S1，对于卷积神经网络中的任一层，分别计算该层输出的各特征图的稠密度;

S2，将该层输出的各所述特征图的稠密度与多个预设阈值进行比较，根据比较结果将各所述特征图进行稀疏编码;其中，不同的比较结果对应不同的稀疏编码方式;

S3，基于该层下一层的卷积层对稀疏编码后的各所述特征图和预先稀疏编码的所述卷积神经网络中的各卷积核进行卷积。

2、根据权利要求 1 所述的方法，其特征在于，所述步骤 S1 具体包括:

对于任一所述特征图，统计该特征图中非 0 元素的个数和该特征图中所有元素的总个数;

将该特征图中非 0 元素的个数与该特征图中所有元素的总个数之间的比值作为该特征图的稠密度。

3、根据权利要求 1 所述的方法，其特征在于，所述预设阈值包括第一预设阈值和第二预设阈值;其中，所述第一预设阈值小于所述第二预设阈值;

相应地，所述步骤 S2 具体包括:

若各所述特征图的稠密度小于所述第一预设阈值，则将各所述特征图编码为稀疏矩阵存储格式;

若各所述特征图的稠密度大于或等于所述第一预设阈值，且小于所述第二预设阈值，则将各所述特征图中的 0 元素进行标记;

若各所述特征图的稠密度大于或等于所述第二预设阈值，则不对各所述特征图进行稀疏编码。

4、根据权利要求 3 所述的方法，其特征在于，所述步骤 S3 之前还包括:

计算训练好的卷积网络中各卷积核的稠密度;

若各所述卷积核的稠密度小于所述第一预设阈值，则将各所述卷积核编码为稀疏矩阵存储格式;

若各所述卷积核的稠密度大于或等于所述第一预设阈值，且小于所述

第二预设阈值，则将各所述卷积核中的 0 元素进行标记；

若各所述卷积核的稠密度大于或等于所述第二预设阈值，则不对各所述卷积核进行稀疏编码。

5、根据权利要求 3 或 4 所述的方法，其特征在于，所述步骤 S3 具体包括：

当各所述特征图或各所述卷积核中存在所述标记时，对各所述特征图或各所述卷积核中标记对应的元素不进行计算。

6、一种应用于卷积神经网络的加速器，其特征在于，包括：神经网络计算阵列模块和动态稀疏调整模块；

其中，所述动态稀疏调整模块用于计算卷积神经网络中各层输出的各特征图的稠密度，将各所述特征图的稠密度与多个预设阈值进行比较，根据比较结果对各所述特征图进行稀疏编码；其中，不同的比较结果对应的不同稀疏编码方式；

所述神经网络计算阵列模块用于对稀疏编码后的各所述特征图和预先稀疏编码的所述卷积神经网络中的各卷积核进行卷积操作。

7、根据权利要求 6 所述的加速器，其特征在于，所述动态稀疏调整模块包括线上稠密度识别模块、输出临时寄存模块、动态编码模块和动态稀疏控制模块；

其中，所述线上稠密度识别模块用于对于任一所述特征图，统计该特征图中 0 元素的个数和该特征图中所有元素的总个数；将该特征图中 0 元素的个数与该特征图中所有元素的总个数之间的比值作为该特征图的稠密度；

所述输出临时寄存模块用于存储卷积神经网络中各层输出的各所述特征图；

所述动态稀疏控制模块用于将所述线上稠密度识别模块输出的各所述特征图的稠密度与多个预设阈值进行比较；

所述动态编码模块用于根据比较结果将所述输出临时寄存模块中各所述特征图进行稀疏编码。

8、根据权利要求 7 所述的加速器，其特征在于，所述预设阈值包括第一预设阈值和第二预设阈值；其中，所述第一预设阈值小于所述第二预

设阈值;

相应地，所述动态编码模块具体用于:

若各所述特征图的稠密度小于所述第一预设阈值，则将各所述特征图编码为稀疏矩阵存储格式;

若各所述特征图的稠密度大于或等于所述第一预设阈值，且小于所述第二预设阈值，则将各所述特征图中的 0 元素进行标记;

若各所述特征图的稠密度大于或等于所述第二预设阈值，则不对各所述特征图进行稀疏编码。

9、根据权利要求 8 所述的加速器，其特征在于，所述动态编码模块还用于:

若预先计算的各所述卷积核的稠密度小于所述第一预设阈值，则将各所述卷积核编码为稀疏矩阵存储格式;

若各所述卷积核的稠密度大于或等于所述第一预设阈值，且小于所述第二预设阈值，则将各所述卷积核中的 0 元素进行标记;

若各所述卷积核的稠密度大于或等于所述第二预设阈值，则不对各所述卷积核进行稀疏编码。

10、根据权利要求 7 或 8 所述的加速器，其特征在于，所述神经网络计算阵列模块具体用于:

当各所述特征图或各所述卷积核中存在所述标记时，对各所述特征图或各所述卷积核中标记对应的元素不进行计算。

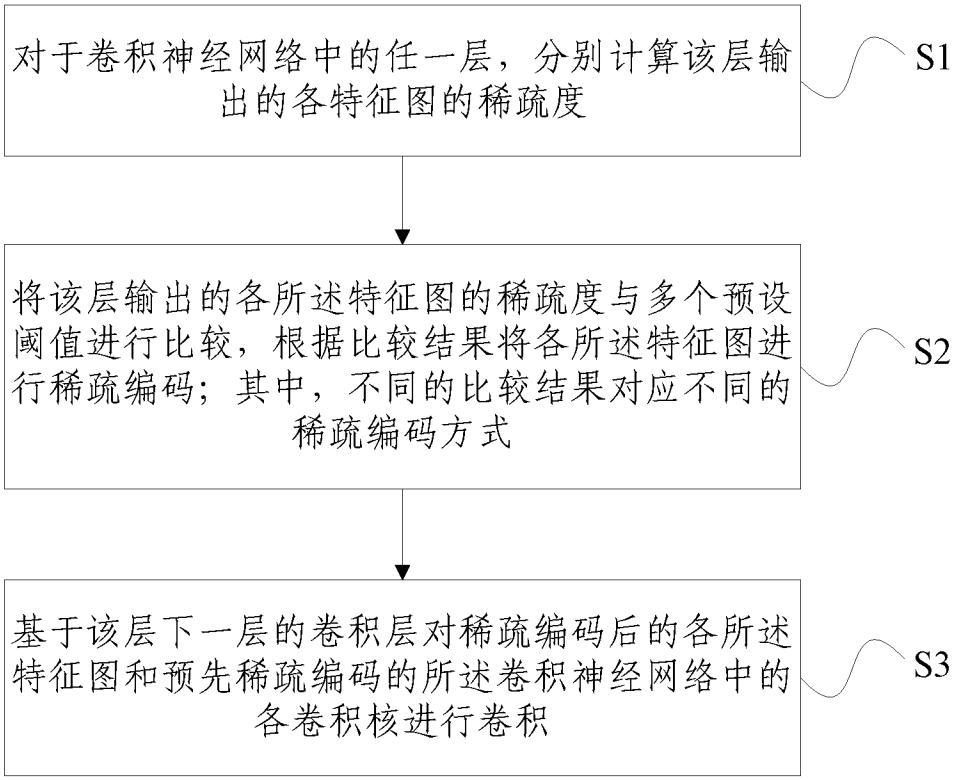

对于卷积神经网络中的任一层，分别计算该层输出的各特征图的稀疏度
S1
将该层输出的各所述特征图的稀疏度与多个预设阈值进行比较，根据比较结果将各所述特征图进行稀疏编码；其中，不同的比较结果对应不同的稀疏编码方式
S2
基于该层下一层的卷积层对稀疏编码后的各所述特征图和预先稀疏编码的所述卷积神经网络中的各卷积核进行卷积
S3

图 1

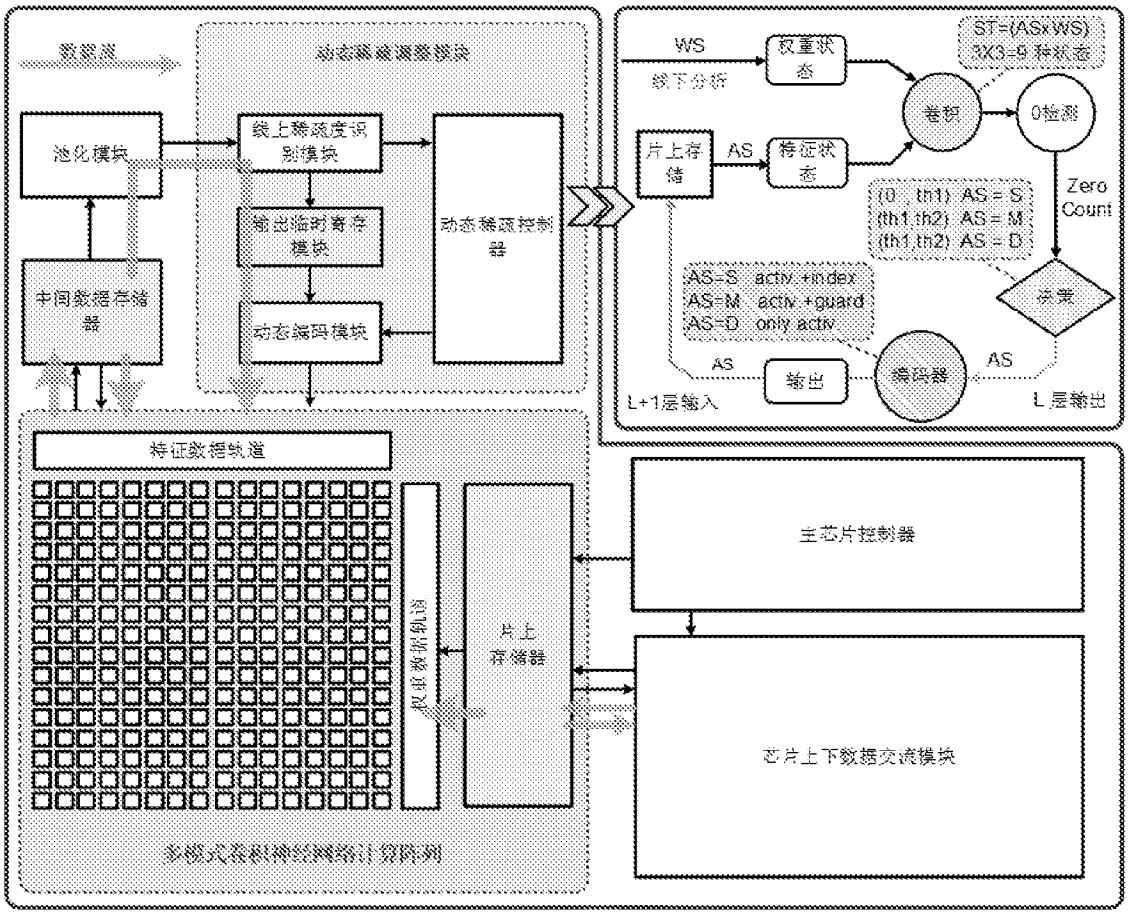

数据流
动态稀疏调整模块
池化模块
线上稀疏度识别模块
输出缓时寄存模块
动态稀疏控制器
中间数据存储器
动态编码模块
特征数据轨道
权重数据轨道
片上存储器
多模式卷积神经网络计算阵列
WS
线下分析
权重状态
ST=(ASxWS)
3X3=9 种状态
卷积
0检测
片上存储
AS
特征状态
Zero Count
(0, th1) AS = S
(th1,th2) AS = M
(th1,th2) AS = D
AS=S activ +index
AS=M activ +guard
AS=D only activ
结果
AS
输出
编码器
AS
L+1层输入
L 层输出
主芯片控制器
芯片上下数据交流模块

图 2

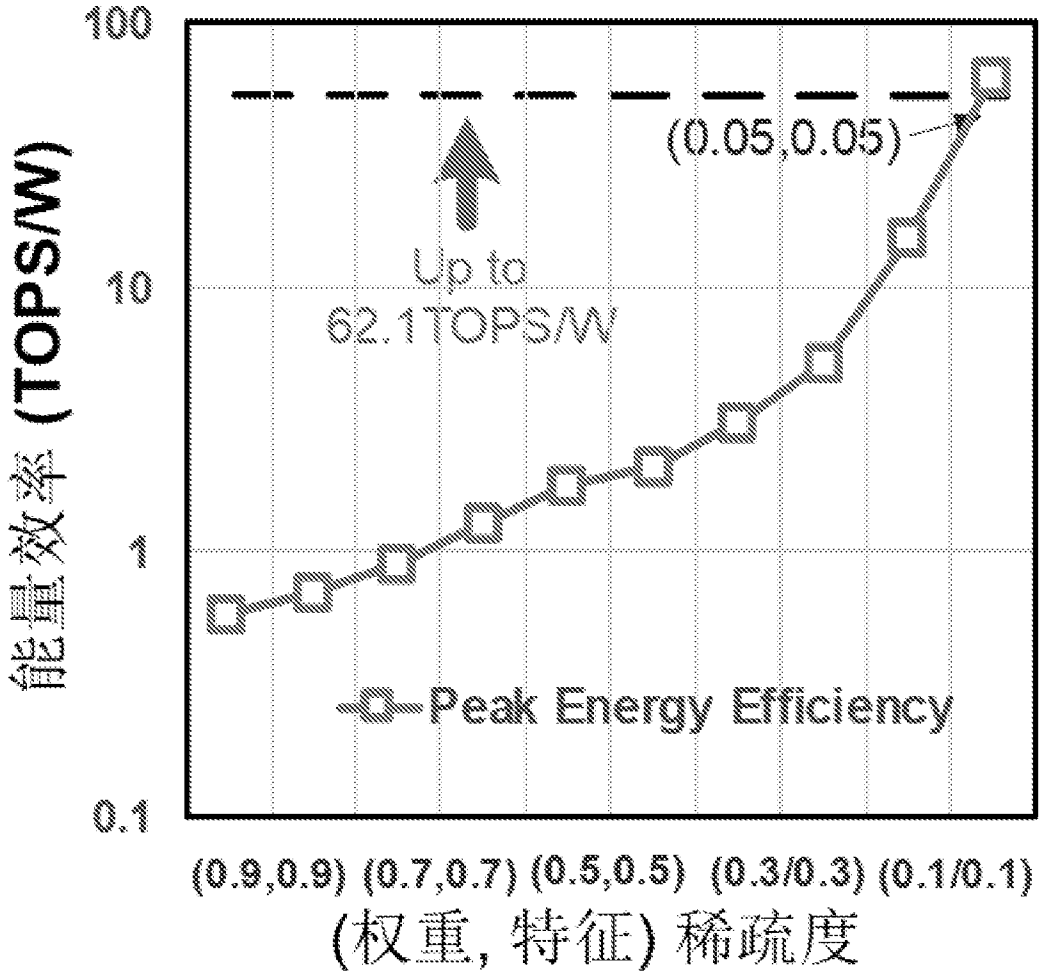

100
10
1
0.1
能量效率 (TOPS/W)
(0.05,0.05)
Up to
62.1TOPS/W
Peak Energy Efficiency
(0.9,0.9) (0.7,0.7) (0.5,0.5) (0.3/0.3) (0.1/0.1)
(权重, 特征) 稀疏度

图 3

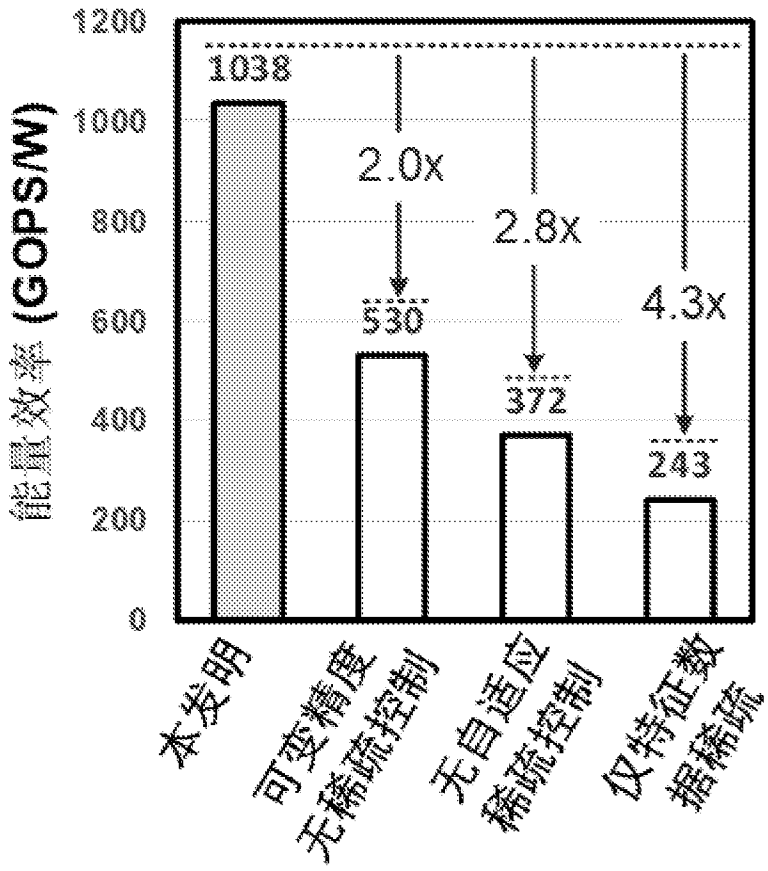

能量效率 (GOPS/W)
1200
1000
800
600
400
200
0
1038
2.0x
2.8x
4.3x
530
372
243
本发明
可变精度无稀疏控制
无自适应稀疏控制
仅特征数据稀疏

图 4

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2018/095365

A. CLASSIFICATION OF SUBJECT MATTER

G06N 3/04(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)

G06N3/-

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)

USTXT; EPTXT; CNTXT; CNABS; WOTXT; VEN; CNKI: 卷积核, 特征图, 稠密度, 稀疏编码, 稀疏矩阵存储格式, 标记, 卷积神经网络, convolutional kernel, feature map, dense degree, pre-sparse encoding, sparse matrix storage format, mark, convolutional neural network

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	CN 107239825 A (BEIJING DEEPHI INTELLIGENT TECHNOLOGY CO., LTD.) 10 October 2017 (2017-10-10) description, paragraphs 22, 103, 115-133, and figures 1-3, 8-11	1-3, 5-8, 10
Y	CN 107239824 A (BEIJING DEEPHI INTELLIGENT TECHNOLOGY CO., LTD.) 10 October 2017 (2017-10-10) description, paragraphs 12 and 53-56, and figures 2 and 4	1-3, 5-8, 10
A	CN 105184362 A (INSTITUTE OF AUTOMATION, CHINESE ACADEMY OF SCIENCES) 23 December 2015 (2015-12-23) entire document	1-10
A	CN 107609641 A (TSINGHUA UNIVERSITY) 19 January 2018 (2018-01-19) entire document	1-10
A	US 2017103309 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 13 April 2017 (2017-04-13) 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
29 November 2018	**07 January 2019**

Name and mailing address of the ISA/CN	Authorized officer
State Intellectual Property Office of the P. R. China **No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing 100088** **China**	
Facsimile No. **(86-10)62019451**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2018/095365

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
CN	107239825	A	10 October 2017	CN	107689948	A	13 February 2018
				CN	107679617	A	09 February 2018
				CN	107689224	A	13 February 2018
CN	107239824	A	10 October 2017	US	2018157969	A1	07 June 2018
CN	105184362	A	23 December 2015	CN	105184362	B	02 February 2018
CN	107609641	A	19 January 2018		None		
US	2017103309	A1	13 April 2017	US	2017103308	A1	13 April 2017

国际检索报告

国际申请号
PCT/CN2018/095365

A. 主题的分类

G06N 3/04(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)

G06N3/-

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））

USTXT;EPTXT;CNTXT;CNABS;WOTXT;VEN;CNKI:卷积核，特征图，稠密度，稀疏编码，稀疏矩阵存储格式，标记，卷积神经网络，convolutional kernel, feature map, dense degree, pre-sparse encoding, sparse matrix storage format, mark, convolutional neural network

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
Y	CN 107239825 A (北京深鉴智能科技有限公司) 2017年 10月 10日 (2017-10-10) 说明书第22、103、115-133段及图1-3、8-11	1-3、5-8、10
Y	CN 107239824 A (北京深鉴智能科技有限公司) 2017年 10月 10日 (2017-10-10) 说明书第12、53-56段及图2、4	1-3、5-8、10
A	CN 105184362 A (中国科学院自动化研究所) 2015年 12月 23日 (2015-12-23) 全文	1-10
A	CN 107609641 A (清华大学) 2018年 1月 19日 (2018-01-19) 全文	1-10
A	US 2017103309 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 2017年 4月 13日 (2017-04-13) 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
2018年 11月 29日	2019年 1月 7日
ISA/CN的名称和邮寄地址 中华人民共和国国家知识产权局(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 夏雪 电话号码 86-(20)-28950718

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2018/095365

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	107239825	A	2017年 10月 10日	CN	107689948	A	2018年 2月 13日
				CN	107679617	A	2018年 2月 9日
				CN	107689224	A	2018年 2月 13日
CN	107239824	A	2017年 10月 10日	US	2018157969	A1	2018年 6月 7日
CN	105184362	A	2015年 12月 23日	CN	105184362	B	2018年 2月 2日
CN	107609641	A	2018年 1月 19日		无		
US	2017103309	A1	2017年 4月 13日	US	2017103308	A1	2017年 4月 13日