申请人：浙江大学

发明人：沈海斌 朱超阳 黄科杰

第一发明人身份证号：321102196711040430

联系人电话及邮箱：13357506534 616527227@qq.com

寄信地址及邮编：浙江大学玉泉校区老生仪楼3楼    邮编310027

**专利申请提交的技术交底材料**

　　一、名称

一种基于结构化剪枝的稀疏神经网络加速器  
　　二、技术领域

本发明公开了一种稀疏卷积神经网络加速器，涉及稀疏神经网络推理加速技术。  
　　三、背景技术

卷积神经网络被广泛应用于计算机视觉、语音识别、自然语言处理、定向广告推荐和网络搜索等领域。虽然卷积神经网络能提供出色的效果，但是它运算参数量大，计算复杂度高。传统处理器性能和功耗的限制，使得大规模的卷积神经网络很难在嵌入式或终端设备上进行实现。因此，在保证卷积神经网络预测准确度的前提下，利用网络的稀疏性对网络模型进行压缩和加速成为了目前研究的热点。

网络稀疏性体现在激励稀疏和权重稀疏两个方面。激励稀疏的原因为：很多神经网络使用了线性整流函数（ReLu）作为激活函数，导致小于零的激励被置为零；部分神经网络（如对抗生成网络）包含了使用零值插入对输入特征图进行向上采样的译码层。权重的稀疏性来自于剪枝。剪枝通过某种度量标准确定需要被剪去的权重。激励稀疏和权重稀疏都可以被利用以提升硬件加速器的处理速度和能效。

四、发明内容  
　　本发明旨在压缩稀疏权重和激励以减少存储空间的需求和数据搬移，并跳过零值权重与激励的读取和运算，提升加速器加速效果，降低加速器功耗。

本发明采用软硬件协同优化，对卷积神经网络进行结构化剪枝，并基于此设计硬件加速器。

本发明一种基于结构化剪枝的稀疏神经网络加速器的加速方法流程如下：

S01：对神经网络进行结构化剪枝，得到压缩模型和表征压缩权重位置的索引集；

S02：对压缩模型进行分析，结合网络各层参数完成压缩网络到硬件的映射，以指令流的形式表示映射关系；

S03：加速器主控制器载入指令流，并控制加速器从片外存储器读取输入激励和权重到对应的缓存器；

S04：计算核根据主控制器的控制信号实现稀疏卷积运算、线性整流运算和池化运算，产生输出激励；

S05：判断当前层是否为最后一层，如果不是，则进行输出激励的压缩编码，并将输出激励和索引对存储到输出激励缓存器，否则直接输出到片外存储器，完成整个神经网络的推理加速。

具体技术方案如下：

一种基于结构化剪枝的稀疏神经网络加速器，包括：主控制器，用于控制网络运算的数据流并产生各模块控制信号；计算核，用于完成稀疏卷积的乘加运算、激活运算和池化运算；输出激励编码器，对输出激励进行压缩并编码，输出激励和索引成对存储；输入激励缓存器，缓存从片外存储器读取的输入激励和索引；输出激励缓存器，缓存输出激励编码器产生的激励索引对；权重缓存器，缓存从片外存储器读取的权重和索引；片外存储器，动态随机存取存储器（DRAM），存储原始图片、权重、神经网络计算的中间结果和最终输出的特征图。

主控制器负责解析指令集，产生输入激励、权重的读取地址和输出激励的存储地址，以及对运算流程进行控制的控制信号。主控制器包含一个指令集解码器、一个地址生成器、一个控制向量生成器和七个计数器。指令集解码器对指令集进行解码，生成各计数器的判断上限和下限。七个计数器依次级联，前一级上溢信号overflow作为下一级增信号inc，七个计数器的输出{kw,kh,ow,oh,ic,oc,l}表示加速进程。地址生成器根据计数器的输出生成下一个时钟周期计算所需的输入激励和权重的读取地址，以及当前输出激励的存储地址。控制向量生成器读入计数器的输出，生成控制信号ctrl，控制计算核。

计算核处理压缩激励和权重。为保证只有非零数据被读取，输入激励的读取需要在权重读取之前。为了处理数据读取的这种依赖并保持吞吐率，计算核内部署了六级流水。前两级流水用于读取非零激励。激励索引寄存器存储输入激励的简介索引，该间接索引作为输出激励寄存器的读地址。在一个非零激励被读取后，接下来的两级流水用于读取对应的非零权重。首先，权重地址指针寄存器根据输入激励索引输出对应的权重地址的头指针和尾指针。接着权重寄存器根据地址指针输出M个权重索引对。最后两级流水用于乘加运算。乘加运算后的部分和存储在部分和寄存器中。当输出激励的整个卷积运算完成后，输出激励经过ReLu模块和池化模块，从计算核输出。

DRAM，神经网络的参数数量巨大，对于VGG-16等大型网络而言，即使经过模型压缩，也难以把权重完全存在缓存器内。同时网络部分层的输出特征图尺寸巨大，同样需要DRAM。

缓存器，包括输入激励缓存器、输出激励缓存器和权重缓存器。网络分块进行计算，需要重复使用的数据存在缓存器中，以减少对DRAM的重复读取，降低读取功耗。

所述基于结构化剪枝的稀疏神经网络加速器的进一步设计于，所述映射具体为：每个计算核并行地计算多张输出特征图，因此共享相同的输入激励。受益于结构化剪枝，每个计算核的负载均衡。计算核内的每个乘法器共享相同的输入激励，并且由于卷积运算窗口滑动的特性，输入激励可以在时间维度上复用。

本发明的优点如下：

本发明通过结构化剪枝和数据流设计，保证输入激励和权重数据的高复用率，以及运算核的负载均衡；结合计算核的六级流水线，完全跳过了零值激励与权重的读取和运算，最大化地利用了神经网络的稀疏特性，进一步提升了卷积神经网络加速器的吞吐率和能效。  
　　五、附图说明

图1加速器工作流程图

图2结构化剪枝示例图,(a)卷积核的结构化剪枝；(b)权重矩阵的结构化剪枝

图3硬件加速器顶层架构

图4主控制器结构图

图5 映射示意图

图6计算核结构图

图7权重编码和索引示意图

图8稀疏加速示意图  
　　六、具体实施方式

以下结合附图对本发明做进一步说明。

如图1所示为基于结构化剪枝的稀疏神经网络加速器加速方法流程示意图。首先，对卷积网络模型进行结构化剪枝，得到压缩模型和表征压缩权重位置的索引集；然后结合网络各层参数完成压缩网络到硬件的映射，生成指令流；接着加速器主控制器载入指令流，并控制加速器从片外存储器读取输入激励和权重到对应的缓存器；计算核根据主控制器的控制信号实现稀疏卷积运算、线性整流运算和池化运算，产生输出激励；最后判断当前层是否为最后一层，如果不是，则进行输出激励的压缩编码，并将输出激励和索引对存储到输出激励缓存器，否则直接输出到片外存储器，完成整个神经网络的推理加速。

结构化剪枝的策略如下：对于初始的密集神经网络，其卷积层四维卷积核尺寸为R×R×C×F，把R×R×C的卷积核称为一个三维卷积核，共计F个三维卷积核。对三维卷积核进行分组，每个小组的内的三维卷积核采用相同的剪枝方式；其全连接层权重矩阵的尺寸为IC×OC，由OC个列向量组成，将多个列向量归为一组，组内的列向量采用相同的剪枝方式。如图2（a）所示为将所有三维卷积核归位一组时的剪枝结果，（b）为将所有列向量归位一组时的剪枝结果。当剪枝后准确度下降明显时，会调整分组情况，将卷积核/列向量分成更多组。

如图3所示为硬件加速器结构示意图。整体结构包括了：主控制器，如图4，解析从上位机CPU接收的指令流，控制网络运算的数据流并产生各模块控制信号；计算核，根据主控制器控制信号完成稀疏卷积的乘加运算、激活运算和池化运算，产生输出激励；输出激励编码器，对输出激励进行压缩并进行CSC（按列压缩）编码，输出激励和索引成对存储；输入激励缓存器，缓存从片外存储器读取的输入激励和索引；输出激励缓存器，缓存输出激励编码器产生的激励索引对；权重缓存器，缓存从片外存储器读取的权重和索引；片外存储器，动态随机存取存储器（DRAM），存储原始图片、权重、神经网络计算的中间结果和最终输出的特征图。

基于结构化剪枝的稀疏神经网络加速器的映射步骤为：首先，将输入特征图和卷积核分块，并将数据分发到对应计算核中；接着，每块数据计算完成时，优先沿输入通道方向更新输入激励和权重；然后沿着卷积窗口滑动的轨迹更新输入激励和权重，直到一层卷积层运算完成。因为不同计算核并行地计算输出特征图的不同输出通道，所以计算核共享相同的输入激励。图5显示了每次分发至计算核的输入激励和权重。其中，输入激励被展开成一个向量，M0个卷积核的权重被展开成一个二维矩阵，矩阵高M0，宽C0×R×R。因为输入激励和权重是稀疏的，且激励和权重均经过压缩编码，因此实际只传输非零的数据和对应的索引。

如图6所示为计算核结构图。只计算稀疏编码后的激励和权重的前提是解决数据读取的依赖。数据读取依赖来自两个方面：编码后，读取压缩权重需要先读取地址指针；读取权重前需要先读取非零激励值，再根据输入激励索引确定权重的地址指针，从而读取权重。所以为了保证计算核的吞吐量，计算核内部署了六级流水线。前两级流水线用于读取输入激励，因为不同输入激励对应地非零权重数量可能不同，因此每个输入激励的有效时间可能不同。计算核会根据主控制器的使能信号决定第一级流水线中输入激励索引的生命周期。第三、四级流水线分别用于读取权重地址指针和非零权重索引对。最后两级流水线完成MAC乘加运算。MAC运算的结果根据权重索引存入部分和寄存器中。当一组输出激励计算完成后，计算核对其进行线性整流运算和池化运算，并将结果输出。

输出激励进行CSC编码，权重矩阵压缩编码方式如图7所示。由于神经网络进行了结构化剪枝，卷积核按照实施例所示映射（图5）后，权重矩阵呈现列稀疏性，即权重矩阵的某些列元素全为零。索引表示非零权重列与前一个非零权重列之间的列数，地址指针表示每一列元素的个数。网络模型压缩阶段已完成对权重的编码，非零权重和索引被存储在独立的DRAM中。所述基于结构化剪枝的稀疏神经网络加速器通过AXI总线与该DRAM相连。

图8为稀疏加速过程，示例中，R等于3，M0等于4.第一个非零激励对应的R×R列权重中仅有四列为非零值。加速开始时，计算核的激励索引寄存器收到主控制器使能信号，输出第一个索引“0”。输入激励寄存器根据该索引读出第一个非零激励“X0”。接着该索引“0”向后级传输。权重地址指针寄存器根据该索引输出第一组权重的地址指针“0”和“4”。权重寄存器根据地址指针并行输出第一组权重{a,b,c,d}。然后该组权重并行地和X0进行MAC运算，并将部分和存入部分和寄存器的第一个地址（由权重索引决定）。因为X0的生命周期为四个时钟周期，所以第五个时钟周期才有新的使能信号来到计算核。计算核开始进行非零激励X2对应的运算。如果X2没有对应的非零权重，则X2不会往下级传输。整个过程由六级流水线实现，保证了计算核的吞吐量。

综上所述，本发明通过结构化剪枝和数据流设计，保证输入激励和权重数据的高复用率，以及运算核的负载均衡；结合计算核的六级流水线，完全跳过了零值激励与权重的读取和运算，最大化地利用了神经网络的稀疏特性，提升了卷积神经网络加速器的吞吐率和能效。

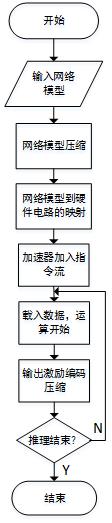


图 1

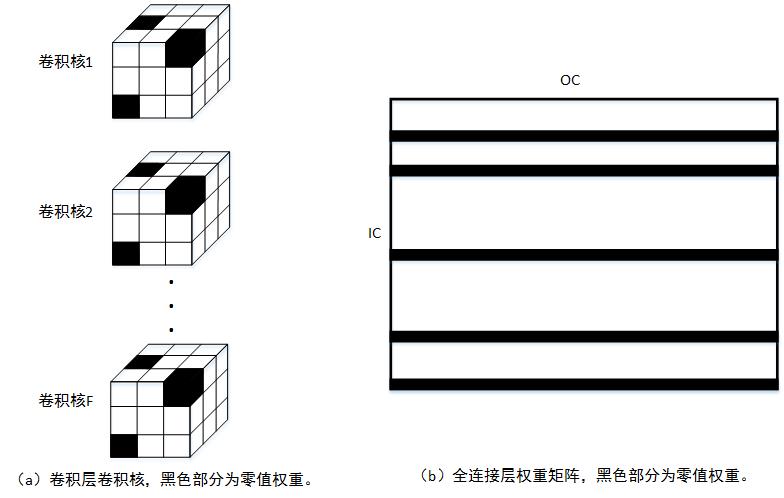


图 2

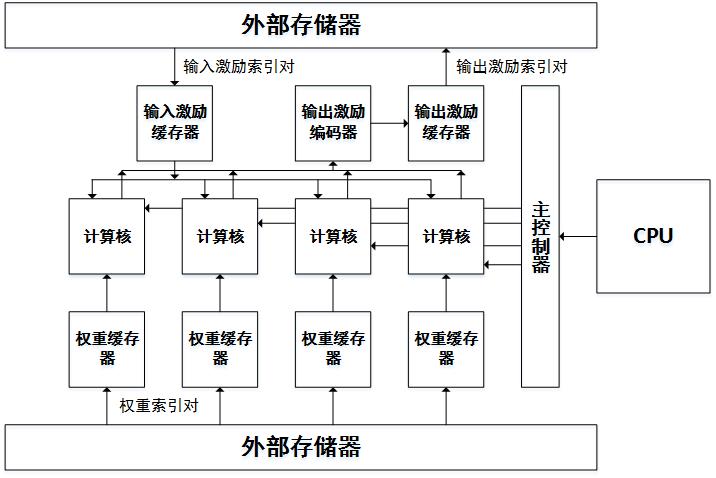


图 3

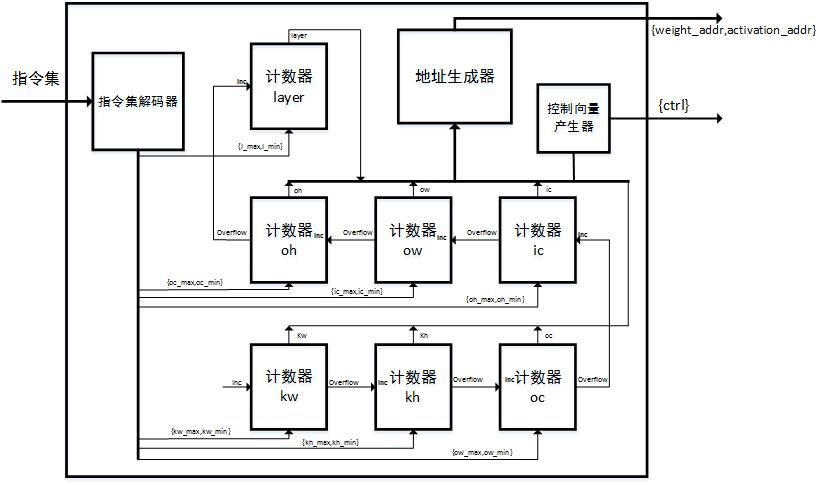


图 4

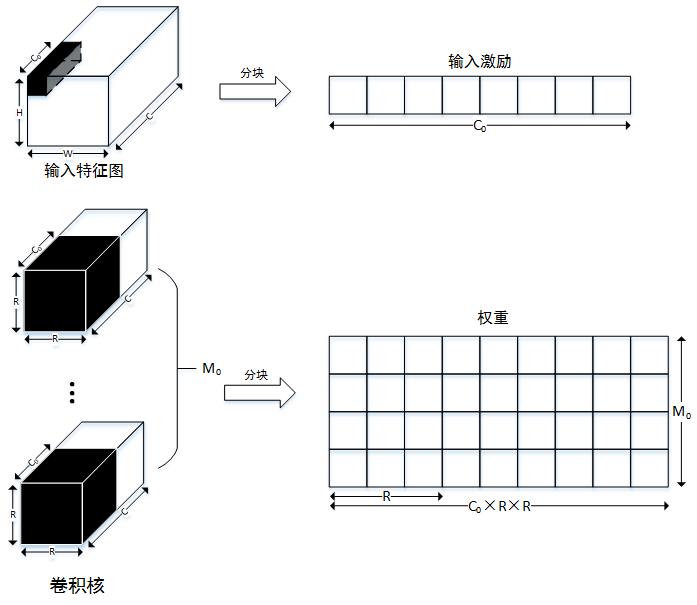


图 5

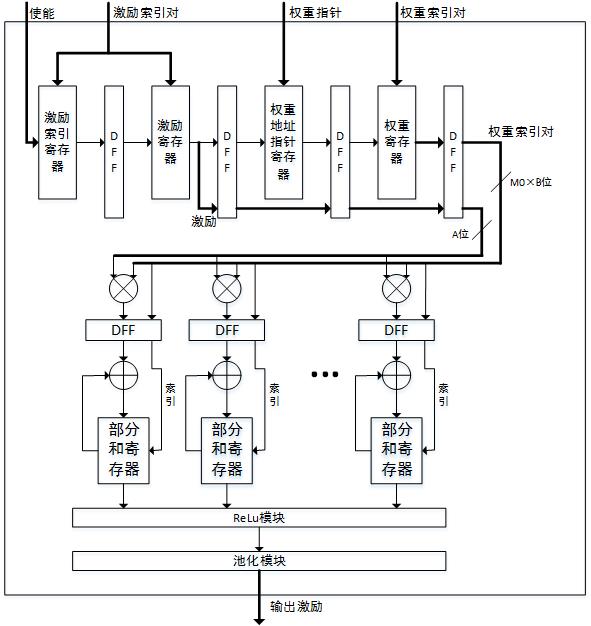


图 6

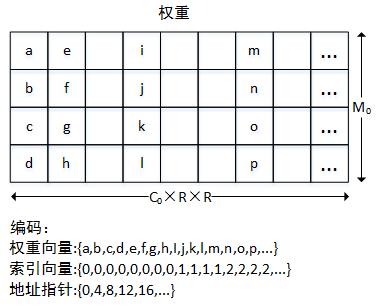


图 7

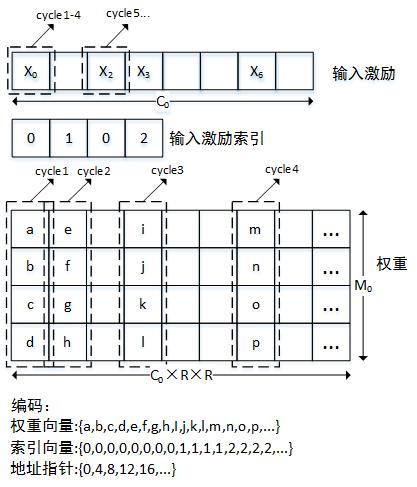


图 8