

引用格式:程智,杨靓,王硕,等. 一种 2D 权值固定数据流架构的研究[J]. 微电子学与计算机, 2021, 38(2): 30-33. [CHENG Z, YANG L, WANG S, et al. Research on a 2D weight stationary dataflow architecture[J]. Microelectronics & Computer, 2021, 38(2): 30-33.]

一种 2D 权值固定数据流架构的研究

程 智, 杨 靓, 王 硕, 娄 冕
(西安微电子技术研究所, 陕西 西安 710054)

摘 要: 随着人工智能算法的发展, 卷积神经网络(CNN)在图像、音频等方面的应用越来越广泛, CNN 算法的计算量也越来越大. 权值固定数据流(WS)将权值固定在寄存器中, 是一种最大化利用卷积重用和 filter 重用的数据流. 不过当前的权值固定数据流结构存在建立流水线时间过长的问题. 本文研究了一种去除 PE(Process Element)行之间的 FIFO, 用加法器连接 PE 行的 2D 权值固定数据流结构. 这种 2D 权值固定的数据流结构计算 AlexNet 时减少了近 2.7 倍建立流水线时间, 并且能够灵活地调整卷积步长.

关键词: 卷积神经网络; 权值固定; 数据流; CNN 加速器; 流水线

中图分类号: TP302.1 **文献标识码:** A **文章编号:** 1000—7180(2021)02—0030—04

Research on a 2D Weight Stationary Dataflow Architecture

CHENG Zhi, YANG Liang, WANG Shuo, LOU Mian

(Xi'an Microelectronics Technology Institute, Xi'an 710054, Shaanxi, China)

Abstract: With the development of artificial intelligence algorithms, convolutional neural networks (CNN) is more and more widely used in image, audio and other aspects, and the amount of calculation of CNN algorithms is also increasing. Weight stationary (WS) is a dataflow that maximizes the use of convolutional reuse and filter reuse by fixing weights in registers. However, the current WS dataflow structure has the problem that the pipeline filling time is lager. This paper studies a 2D-WS dataflow structure that uses PE adders to remove the FIFO between PE lines. This kind of 2D-WS dataflow structure reduces the pipeline filling time by nearly 2.7 times when calculating AlexNet, and can flexibly adjust the stride size.

Key words: convolutional neural network; weight stationary; dataflow; CNN acceleration; pipeline

1 引言

近年来, 卷积神经网络算法在语音识别^[1]、图像识别^[2]等应用上性能表现显著. 随着卷积神经网络朝着更深层方向的发展, 算法的精度得到了很大的提高, 但是计算量也随之增大. 例如 ResNet 网络最多达 152 层, 需要 11.3G 次 MACs(乘累加)操作和 60M 的权值参数^[3]. 目前, 大部分神经网络算法的训练和推断都是在通用处理器上完成的, 如 GPU、CPU 等. 但是由于功耗、速度以及成本的问题, 专用加速器被认为是加速神经网络算法的未来^[3]. 卷积

神经网络的计算量主要集中在卷积操作的部分, 因此对卷积操作的加速就成了研究 CNN 加速器中的关键之一. 大部分加速器主要包括存储单元和运算单元. 由于硬件资源的限制, 卷积计算所需要的 Filter(卷积核)、Input feature map(输入特征图)以及 Psum(部分和)不可能全部存放在片上缓存中. 不过卷积神经网络在计算过程中存在大量的数据重用^[4], 因此选择合适的数据流能够有效减少特定数据在片上缓存中的读写开销, 从而提高神经网络的性能. 权值固定数据流是一种将权值固定在寄存器中, 可以有效减少权值在片上缓存中的读写开销的

数据流. 对比其他数据流, 权值固定数据流具有较好的可扩展性^[5-7]. 随着计算规模的增大, 运算阵列的面积和耗能方面优势明显. 但是长的建立流水时间影响了它的性能表现. 本文通过对现有的权值固定数据流的分析, 研究了一种能够减少建立流水线时间的2D权值固定数据流结构.

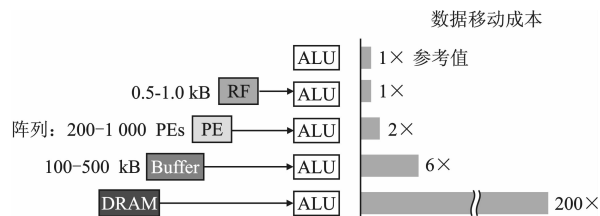


图1 数据在各层次移动的成本比较^[3]

2 数据流

卷积的每一次 MACs 操作需要进行三次存储器的读取 (filter, 输入 feature map 与 Psum) 和一次存储器的写入 (更新后的 Psum). 有可能所有存储器的访问都必须经过片外 DRAM, 这将会严重影响加速器的吞吐量和能效. 如今, 大部分加速器采用存储器分层的方法来存储上边提到的三种数据. 比如说, 由 DRAM, 几百 KB 的全局缓冲存储器, 内部 PE 网络和每个 PE 中几 KB 的寄存器组成的四层存储器层次. 如图 1 这四层存储器层次的能耗成本由高到低, 价格成本由低到高. 例如, 从寄存器层次读取数据比从 DRAM 中读取数据的能耗低 1 到 2 个数量级. 因此, 对于从高能耗级别存储器中读取到低能耗级别存储器的数据, 都希望可以尽可能多地重用这些数据. 这样就可以最大程度上减少对高能耗级别存储器的访问次数, 从而降低了数据移动的成本.

数据流决定了将哪些数据读入存储器层次结构的哪个级别以及何时处理它们. 由于卷积神经网络在处理过程中不具有随机性, 因此可以设计一种支持特定数据流的结构来处理它, 这个数据流能够适应不同形状和大小的卷积神经网络.

在 CNN 的卷积计算过程中存在多种数据重用, 其中就包括 filter 的重用. 如果把被重用的 filter 固定存放在寄存器中一段时间, 不断地更换卷积操作需要的其他参数, 直到与这些 filter 相关的所有输入数据全部与之完成计算再去替换它们. 那么整个卷积计算将会大大地减少对全局缓冲存储器中 filter 的访问次数, 这就是权值固定数据流. 权值固定数据流将多个相邻输出映射到不同的流水线阶

段, 以共享输入和重用存储在本地的权值.

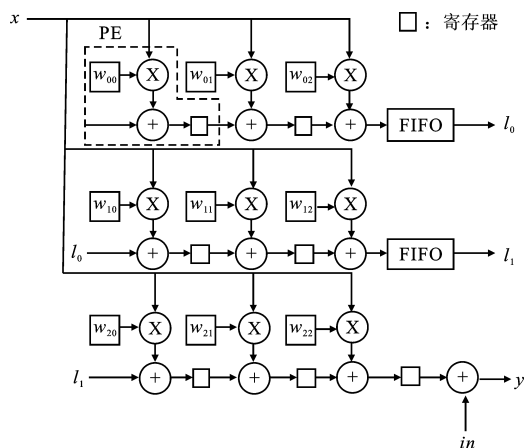


图2 CNP 架构的权值固定数据流阵列

Farabet 等人提出的加速器 CNP 正是采用了权值固定的数据流结构, 加速器的卷积计算部分是用 PE 阵列完成的. 每一个 PE 由一个乘法器, 一个加法器, 存放 filter 的寄存器和存放 Psum 的寄存器组成. 加速器是一个 $R \times S$ ($R \times S$ 是 filter 的大小) 的 PE 阵列, 每一行中的 PE 从左到右级联, 相邻 PE 行通过一个片上 FIFO 相连, 见图 2. 在 T_0 时钟周期, 向所有 PE 广播输入 feature map (i, j) 位置; 在 T_1 时钟周期, 向所有 PE 广播输入 feature map $(i, j+1)$ 位置 (或者下一行). 在计算过程中, 输入 x 与寄存器中存储的权值相乘并累加, 得到一个输出的 Psum, 然后将这个 Psum 传递给相邻的 PE. 最后一个加法器中输入 in, 完成不同通道计算结果的累加, 最终得到一个输出. 它在每个时钟周期完成下面计算:

$$y_{ij} = in_{ij} + \sum_{m=0}^{S-1} \sum_{n=0}^{R-1} x_{i+m, j+n} w_{mn} \quad (1)$$

式中, x_{ij} 是输入 feature map 的 (i, j) 位置; w_{mn} 是 filter 的权值; in_{ij} 是不同通道的卷积结果; y_{ij} 是输出 feature map 的 (i, j) 位置或一个 Psum.

整个 PE 阵列其实是一个深度流水线结构, 行间 FIFO 的大小等于输入 feature map 的宽度减去 filter 的宽度, 流水线的深度大概等于输出 feature map 的宽度乘以 filter 的高度. 因此它的流水线深度是动态变化的, 是随着输入 feature map 的宽度和 filter 的高度的增加而增大的. 在卷积层的步长不为 1 时, 每计算输出 feature map 新的一行, 都需要重新建立流水. 图 3 给出了 CNP 权值固定数据流在计算 AlexNet 各个卷积层时, 输出结果时间和建立流水时间占比情况, 可以看出建立流水消耗的时间占

卷积操作的近 50%。

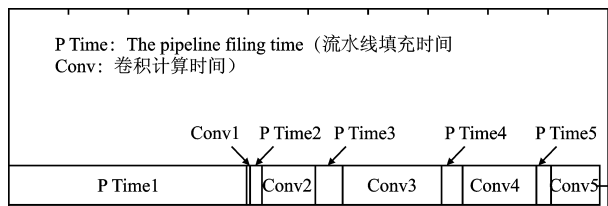


图 3 AlexNet 不同层建立流水线时间和卷积输出时间占比

3 2D 权值固定数据流

如图 4 所示,2D 权值固定数据流采用了 3×3 大小的 PE 阵列完成卷积操作,通过填充 0 或分解 filter 也可以计算非 3×3 大小的 filter. 每一个 PE 由一个乘法器,一个加法器,存放权值的寄存器和存放 Psum 的寄存器组成. 与 CNP 权值固定数据流结构相比,连接两个 PE 行的 FIFO 被替换成了一个加法器和一个寄存器. 每个 PE 行有独立的数据广播通道,他们分别输入 feature map 的相邻三行. PE 行之间的加法器能够将这三行输入 feature map 的计算结果聚合起来,使得 Psum 的传递方向从一维扩展到了二维. 如图 5 所示,在 T_0 时钟周期,开始向第一行 PE 广播输入 feature map 的第 0 行,在 T_1 时钟周期,开始向第二行 PE 广播输入 feature map 的第 1 行,在 T_2 时钟周期,开始向第三行 PE 广播输入 feature map 的第 2 行. 输出 y 在第 T_5 时钟周期已经能够得到结果 y_{00} ,所以这个 2D 权值固定数据流的流水线深度为 6,并且流水线的深度是与输入 feature map 的宽度和 filter 的高度无关.

$$y_{00} = x_{00} \cdot w_{00} + x_{01} \cdot w_{01} + x_{02} \cdot w_{02} + x_{10} \cdot w_{10} + x_{11} \cdot w_{11} + x_{12} \cdot w_{12} + x_{20} \cdot w_{20} + x_{21} \cdot w_{21} + x_{22} \cdot w_{22} \quad (2)$$

并行数据输入的方式使得加速器调整步长 U 变得非常容易,仅仅需要将合适的数据存放在输入 feature map 的缓冲区中.

4 对比分析

通过对 AlexNet^[3]、CaffeNet^[8]、NinNet^[9]、ResNet-18^[10]四个算法卷积层的分析,对比了权值固定数据流结构与 2D 权值固定数据流结构总消耗的建立流水时间. 发现 2D 权值固定数据流结构处理卷积神经网络时,能够有效的减少建立流水线的时间.

根据结构分析中提到的两种权值固定数据流流水线深度的不同,可以分别计算上述四种算法总共

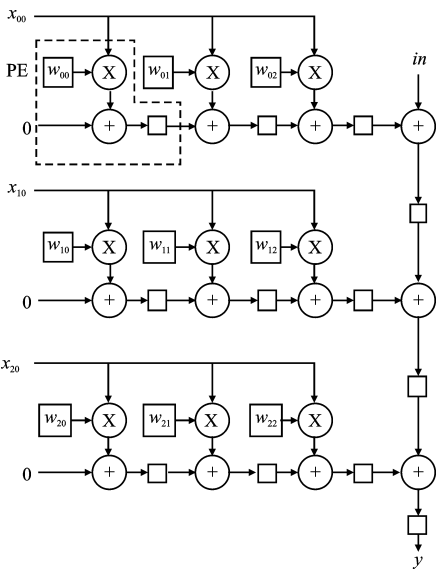


图 4 2D 权值固定数据流阵列示意图

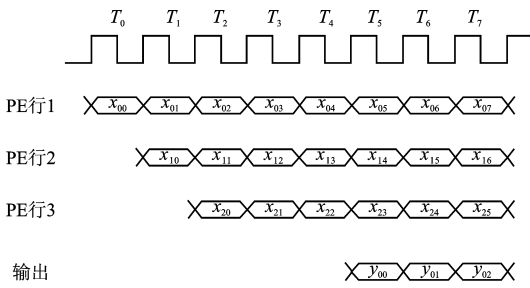


图 5 2D 权值固定计算时空图

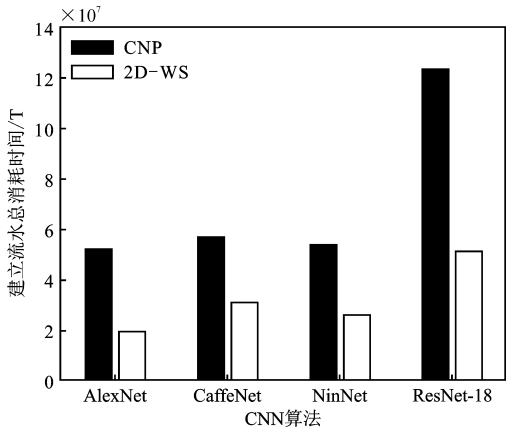


图 6 四种 CNN 算法建立流水时间消耗对比

消耗的建立流水线时间,结果如图 6 所示. ResNet-18 的建立流水时间明显更多的原因是它有更深的卷积层.

在步长不为 1 时,CNP 架构的建立流水时间会因为频繁建立流水而迅速增多,并且在处理大尺寸图像时,这个时间会更多. 2D 权值固定数据流流水线深度则与步长和输入 feature map 的宽度均无关.

对于 AlexNet 网络 CNP 架构的建立流水时间是 2D 权值固定数据流结构的 2.7 倍。

5 结束语

权值固定数据流实现神经网络加速器是一种简单、可行的,并且具有良好的可扩展性的数据流结构,但是这种结构存在建立流水时间过长的问题。当处理大尺寸图像和深层卷积神经网络时,这个问题将会尤其突出。本文通过介绍现有权值固定数据流的组成结构和计算过程,研究了一种能够改善建立流水时间过长问题的 2D 权值固定数据流结构。这种权值固定数据流结构是将 Psum 的数据流动从一维变换到二维,三行输入 feature map 数据同时计算。因此一定程度上缩短了建立流水线的时间,加速了卷积的计算过程。

参考文献:

- [1] DENG L, LI J Y, HUANG J T, et al. Recent advances in deeplearning for speech research at Microsoft [C]//Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, BC, Canada: IEEE, 2013: 8604-8608. DOI: 10.1109/ICASSP.2013.6639345.
- [2] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks [C]//Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, Nevada: ACM, 2012: 1097-1105.
- [3] SZE V, CHEN Y H, YANG T J, et al. Efficient processing of deep neural networks: a tutorial and survey [J]. Proceedings of the IEEE, 2017, 105(12): 2295-2329. DOI: 10.1109/JPROC.2017.2761740.
- [4] CHEN Y H, EMER J, SZE V. Eyeriss: a spatial architecture for energy-efficient dataflow for convolutional neural networks [C]//Proceedings of 2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture. Seoul, South Korea: IEEE,

2016: 367-379. DOI: 10.1109/ISCA.2016.40.

- [5] FARABET C, POULET C, HAN J Y, et al. CNP: an FPGA-based processor for convolutional networks [C]//Proceedings of 2009 International Conference on Field Programmable Logic and Applications. Prague, Czech Republic: IEEE, 2009: 32-37. DOI: 10.1109/FPL.2009.5272559.
- [6] CHEN H Y, KRISHNA T, EMER J S, et al. Eyeriss: an energy-efficient reconfigurable accelerator for deep convolutional neural networks [J]. IEEE Journal of Solid-State Circuits, 2017, 52(1): 127-138. DOI: 10.1109/JSSC.2016.2616357.
- [7] LU W Y, YANG G H, LI J J, et al. FlexFlow: a flexible dataflow accelerator architecture for convolutional neural networks [C]//Proceedings of 2017 IEEE International Symposium on High Performance Computer Architecture. Austin, TX, USA: IEEE, 2017: 553-564. DOI: 10.1109/HPCA.2017.29.
- [8] DONAHUE J. CaffeNet [EB/OL]. https://github.com/BVLC/caffe/tree/master/models/bvlc_reference_caffenet.
- [9] LIN M, CHEN Q, YAN S C. Network in network [J]. arXiv;1312.4400, 2013.
- [10] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition [C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016: 770-778. DOI: 10.1109/CVPR.2016.90.

作者简介:

程智男,(1993-),硕士研究生.研究方向为嵌入式系统设计. E-mail: xga_Cz@163.com

杨靓男,(1975-),博士,研究员.研究方向为嵌入式计算机体系结构

王硕男,(1988-),学士,工程师.研究方向为计算机体系结构.

娄冕男,(1987-),博士,高级工程师.研究方向为嵌入式处理器体系结构.