Bit Fusion: Bit-Level Dynamically Composable Architecture for Accelerating Deep Neural Networks

比特融合：对加速DNN的比特级动态可组合架构

ISCA'2018

汇报人：王铎【中国科学院大学】

May 22, 2019

目录

背景

- 量化方法，在不损失精度下减少操作数位宽
- 对不同的DNNs的比特宽变化不同,甚至需要单独调整每一层

问题

- 固定位宽度加速器要么提供有限的好处,以适应最坏的情况下的比特宽要求,或不可避免地导致最终精度的下降

目录

Bit Fusion加速器

- 探索动态位级融合/分解新维度
- 构成了一组位级处理元素BitBrick的数组 Fusion-PE
- 动态地融合到单个DNN层的比特宽
- 最优粒度下使计算和通信最小化

目录

动态融合

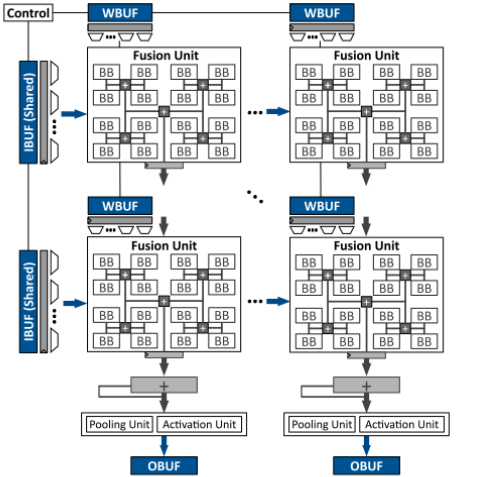

Figure 2: Bit Fusion systolic architecture comprising a collection of BitBricks (BBs) that can fuse to form FPEs.

BitBrick（BB）

- 均可执行单个二值(0、+ 1)和三值(-1、0、+ 1)的乘加操作

动态融合

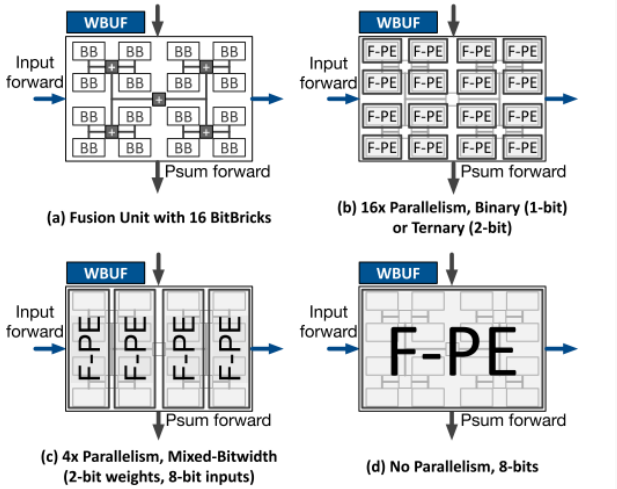

Figure 3: Dynamic composition of BitBricks (BBs) to construct Fused Processing Engines (FPEs).

Fusion Unit

- 一个具有16个BitBrick
- 可以提供1、2、4、8和16个有不同操作数位宽的Fused-PE

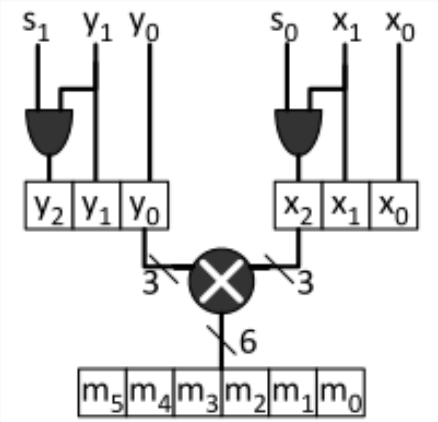

Figure 5: A single BitBrick.

BitBrick

- 两个2-bit操作数x1,x0 和y1,y0，两个符号位s1, s0
- 符号位定义了2-bit操作数x1,x0带符号（-2到1）或者无符号（0到3）

Bitwidth Mode	Operation
2n-bit × 2n-bit	$2^{2n} \times (A_{2n})_{hi} \otimes (B_{2n})_{hi} + 2^{n} \times (A_{2n})_{hi} \otimes (B_{2n})_{lo}$ $+ 2^{n} \times (A_{2n})_{lo} \otimes (B_{2n})_{hi} + 2^{0} \times (A_{2n})_{lo} \otimes (B_{2n})_{lo}$
2n-bit × n-bit	$2^{n} \times (A_{2n})_{hi} \otimes B_n + 2^{0} \times (A_{2n})_{lo} \otimes B_n$ $+ 2^{n} \times (C_{2n})_{hi} \otimes D_n + 2^{0} \times (C_{2n})_{lo} \otimes D_n$
n-bit × 2n-bit	$2^{n} \times A_n \otimes (B_{2n})_{hi} + 2^{n} \times C_n \otimes (D_{2n})_{hi}$ $+ 2^{0} \times A_n \otimes (B_{2n})_{lo} + 2^{0} \times C_n \otimes (D_{2n})_{lo}$
n-bit × n-bit	$2^{0} \times A_n \otimes B_n + 2^{0} \times C_n \otimes D_n$ $+ 2^{0} \times E_n \otimes F_n + 2^{0} \times G_n \otimes H_n$

Figure 6: Supported modes for a fuse_multiply$_{n \to 2n}$

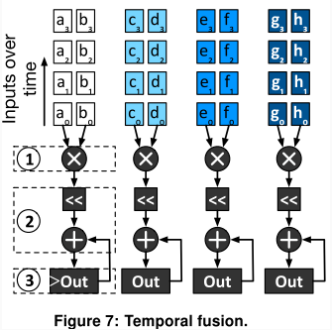

Figure 7: Temporal fusion.

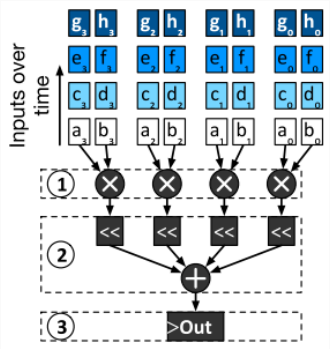

Figure 8: Spatial fusion.

空间融合比时间融合需要更多面积但是性能较好

混合模式：空间融合组合16个BitBrick以实现$\text{fuse_multiply}_{2->8}$算子，再利用时间融合组合每四个周期发射$\text{fuse_multiply}_{2->16}$

目录

Table 1: Bit Fusion Instruction Set.

OpCode (5-bits)	Operand Specification (6-bits)		Loop Identifier (5-bits)	Immediate (16-bits)
setup	op0.bitwidth	op1.bitwidth	X	X
ld-mem	scratchpad-type	mem.bitwidth	loop-id	num-words
st-mem				
rd-buf		X		X
wr-buf				
gen-addr		ld/st		stride
compute	fn			X
loop	X	loop-level		num-iterations
block-end	Address of next instruction			

代码优化

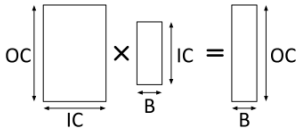

Figure 10: A single Fully-Connected Layer. The × symbol represents matrix-matrix multiplication.

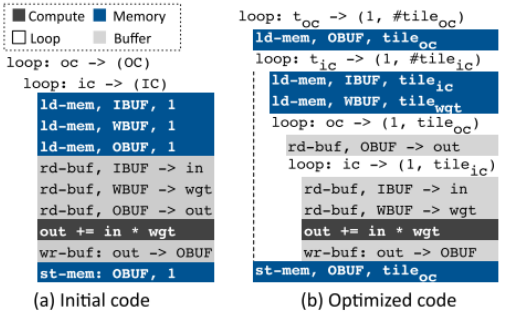

Figure 11: (a) The equivalent code for the Fully Connected Layer. (b) Optimized code using loop tiling and ordering. `setup` and `gen-addr` instructions omitted for clarity.

循环排序:优化外层循环和存储指令顺序以降低片外访问，在输入稳定、输出稳定和权重稳定之间切换。

循环切片:使得循环操作所需的数据适合于片上scratchpad。

目录

实验平台

Table 3: The evaluated ASIC and GPU platforms. *Stripes entries are per-tile.

	ASIC				GPU	
Chip	BitFusion	Eyeriss	Stripes*	**Chip**	Titan X	Tegra X2
Cores (1.1 mm^2)	512 FU	168 Pes	4096 SIPs	**Cores**	3,584	256
On-chip Memory	112 KB	108 KB	2 MB eDRAM 16 KB SRAM	**Memory**	12 GB	8 GB
Chip Area (mm^2)	5.87	5.87	3.62	**TDP**	250 W	7.5 W
Frequency	500 MHz	500 MHz	980 MHz	**Frequency**	1,531 MHz	875 MHz
Technology	45	45	45	**Technology**	16 nm	16 nm

实验结果

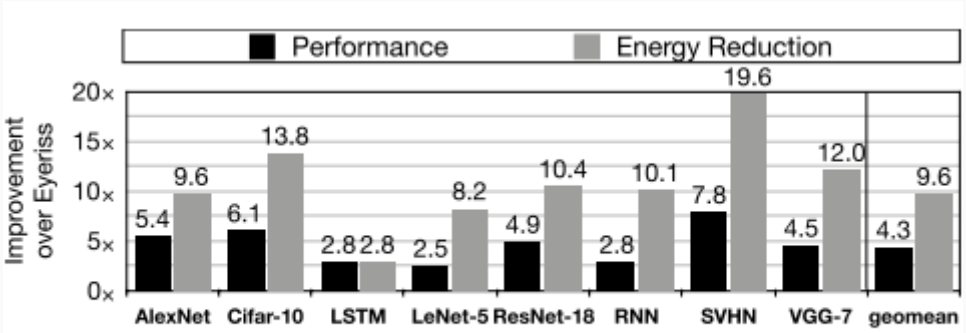

Figure 12: Bit Fusion performance and energy improvements over Eyeriss.

实验结果

Eyeriss

4.3×加速和9.6×能耗节省

Stripes[a]

2.4×加速和4.1×能耗节省

GPU

比Tegra X2有4.3×加速

[a] P. Judd, J. Albericio, T. Hetherington, T. M. Aamodt, and A. Moshovos, "Stripes: Bit-serial deep neural network computing," in MICRO, 2016.

总结

动态位级融合和分解

- 将比特级可组合的PE与DNN层的变化位宽相匹配
- 最优粒度下使计算和通信最小化

对位级组合能力的微结构设计

- 一种二维的BitBricks阵列
- 构造了一个可融合的处理引擎
- 在不同的位宽下执行DNN计算

硬件-软件抽象以比特灵活的加速 **Fusion-ISA**

Thanks

Bit Fusion: Bit-Level Dynamically Composable Architecture for Accelerating Deep Neural Networks

比特融合：对加速**DNN**的比特级动态可组合架构