General Deep Neural Network Accelerator

通用型深度神經網絡加速器

組別 : B265　　指導教授 : 鄭桂忠　　組員 : 蔡博智、薛耕典、江振源

Abstract

深度卷積神經網絡中，輕量化神經網絡改變了卷積運算的形狀和尺寸，使一般DNN加速器無法有效率地運算輕量化網絡模型。本專題透過分析四種卷積運算的特性，結合數據流與硬體架構的設計，我們設計出一種可以根據不同卷積運算，對不同方向展開做平行化運算的硬體架構，使加速器在計算四種卷積運算時都可以維持高運算單元使用效率。

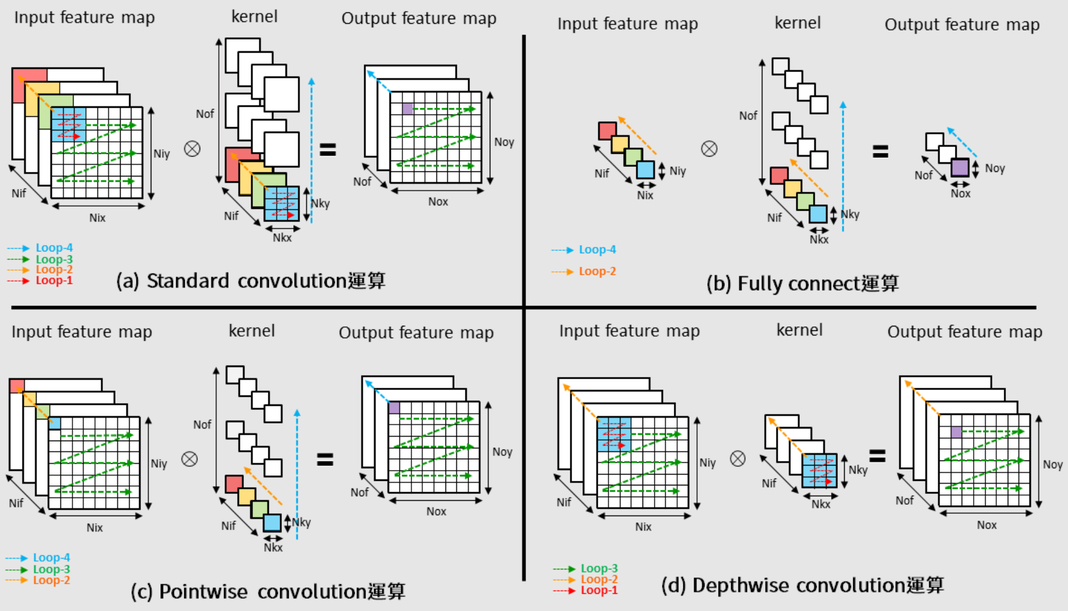
Method

我們引入四循環卷積運算設計技巧，用於分析DWC、PWC、FC運算，並發現沒有共通的迴圈可以做平行化的展開，於是我們結合數據流與硬體架構做設計，跳脫分成weight與pixel SRAM的框架，分成高、低bandwidth的SRAM，做不同類型卷積運算時將weight與pixel存在不同的SRAM，以達成在相同硬體架構下對不同迴圈的做展開的結果，對於不同類型卷積運算都可以有高運算單元使用效率，以提高運算速度、降低功耗。

Implementation

1. 分析 :

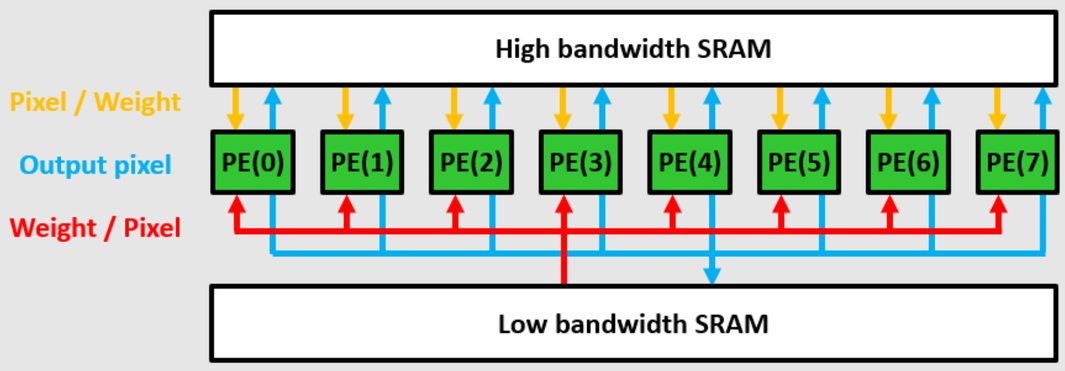
分析不同類型卷積運算，發現並沒有共通的迴圈，若是展開特定的方向，必定會在某些卷積運算有很低的運算單元使用效率。



	loop-1	loop-2	loop-3	loop-4
Standard Convolution	V	V	V	V
Fully Connect		V		V
Pointwise Convolution		V	V	V
Depthwise Convolution	V	O (無累加)	V	

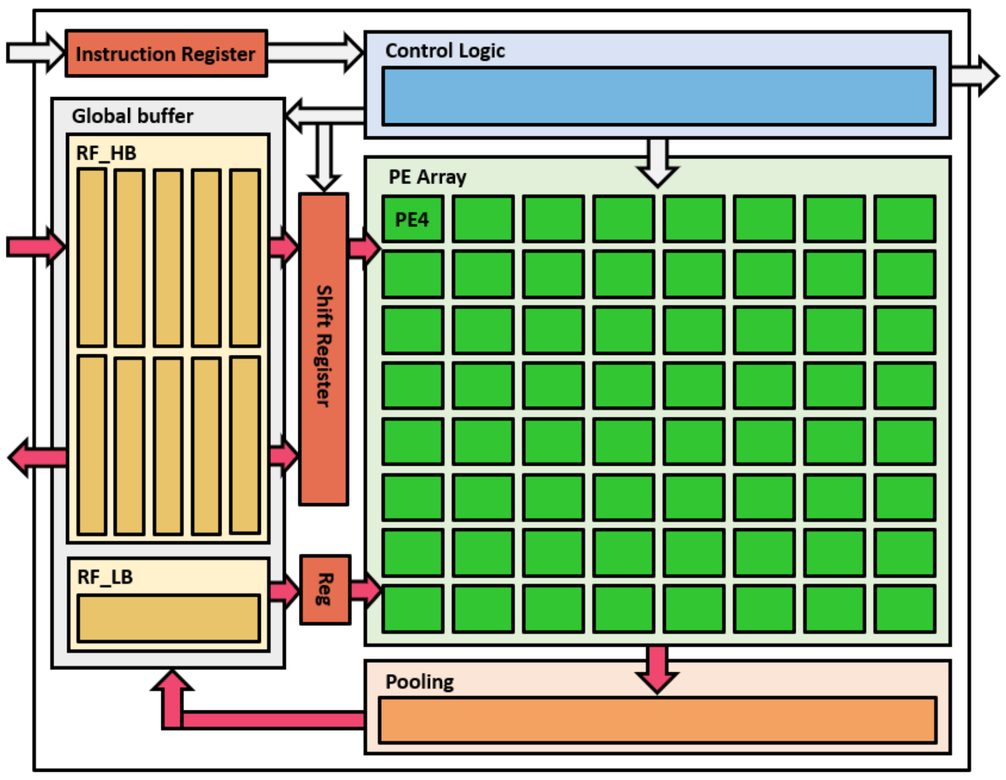
2. 數據流設計 :

下一層為FC時，將Output存入Low bandwidth SRAM，使加速器做Loop-4方向展開，下一層為其他運算則將Output存入High bandwidth SRAM，使加速器做Loop-3方向展開，並使用Output Stationary的設計，避免partial sum的讀寫，與增加weight的複用，來降低SRAM的讀寫量，降低功耗。



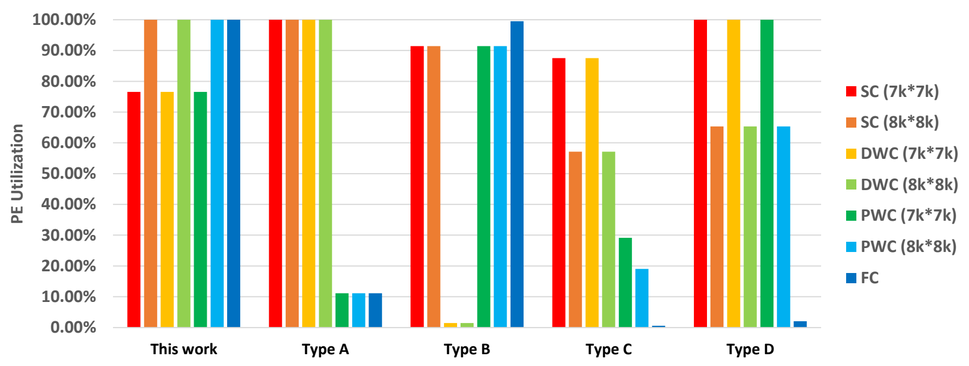
3. 硬體架構設計 :

a. Instruction Register : 暫存器，儲存外部輸入Instruction指令
b. Control Logic : 根據Instruction，產生address與PE的控制訊號
c. Global Buffer : 為Register File儲存pixel, weight資料
d. PE Array : 256個運算單元(PE)做乘累加運算，並將結果做ReLU
e. Pooling : 累加與除法器，支援不同大小池化層做Average pooling



Results

我們的設計相較於其他常見展開方式 ，在不同類型卷積運算下都維持著高運算單元使用效率。



	This work	Eyeriss	Eyeriss v2
Technology	40nm	65nm	65nm
Gate Count (NAND-2)	1553.4k gate	1176k gate	2695k gate
On-Chip SRAM	90.1 KB	108 KB	246 KB
Number of PEs	256	168	384
Clock Rate	200 MHz	200 MHz	200 MHz
Bit Precision	8b	16b	8b
Peak Throughput	51.2 GMACS	33.6 GMACS	76.8 GMACS
Area efficiency	32960 (MACS/gate)	28571.4 (MACS/gate)	28497.2 (MACS/gate)
Power	134.2 mW	278 mW (AlexNet)	584 mW (AlexNet)
Power efficiency	381.55 (GMACS/W)	120.9 (GMACS/W)	131.5 (GMACS/W)

Conclusion

我們提出通用於四種卷積運算的數據流與硬體架構，並驗證在不同類型卷積運算、不同卷積核大小下，若輸入特徵圖為8的倍數，有著100%的PE使用效率，若輸入特徵圖為7的倍數也有76.5%的PE使用效率，相較於其他數據流設計，有著最高的通用性，且實際合成電路，達到面積效率32960(MACS/gate)與能源效率381.55(GMACS/W)。