22nm 製程下全數位 6T SRAM 記憶體內運算巨集於 Resnet - 20 中的邊緣運算

All-digital 6T SRAM-based In-Memory Computing (IMC) macro in 22nm technology for edge computing in ResNet-20 machine learning applications

指導教授: 王進賢 特聘教授　專題生: 林宇恩、吳明和、陳宥炘、翁嘉櫻、潘弘恩

研究摘要

傳統的計算架構（如 CPU、GPU）採用**馮紐曼架構**，需要在記憶體和運算單元之間傳輸數據。再者，深度學習涉及大量數據處理，使得資料傳輸速度比計算速度慢許多，導致時間和功耗的浪費。為了解決這個瓶頸，提出了「**記憶體內運算**」（**IMC**）的新架構。在 IMC 中，SRAM 同時作為記憶體和運算單元，讓計算和記憶體操作可以同時進行。相比傳統的馮紐曼架構，IMC 在性能和功耗上有明顯的改進，本次專題在操作電壓為 0.72V 且未加入動態電壓調節技術的情況達到 **56.6 Tops/W** 與 **0.68 Tops/mm²**之功耗表現。

專題中選擇 **ResNet - 20** 網路，將數據大小相似的卷積層替換為 IMC 架構。本研究參考了台積電團隊在 2021 年發表於 ISSCC 期刊的 IMC 電路設計，使用 6T SRAM 和 NOR 閘建立新的記憶體內運算電路，以及 2018 年 IEEE/CVF 計算機視覺與模式識別會議中的量化論文。

卷積層選擇介紹

由於在考量 IMC macro 硬體大小時以論文上提供的架構為優先，因此選定 ResNet-20 中**資料數量**與**步伐(stride)**最適合之卷積層：

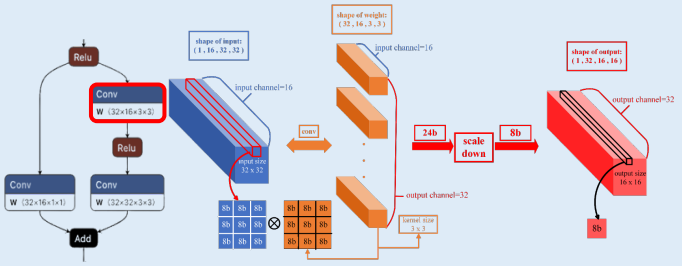

圖 1、卷積層架構與對應資料之大小

研究局部設計

Input / Weight SRAM

我們依次從上到下提取輸入激活位元，使我們能在 8 個周期內提取並計算與 kernel 卷積的 144 個輸入激活，並將一個 kernel 中 144 個 8 位元權重分為前 4 位和後 4 位，分別存放在兩個 IMC 陣列中：

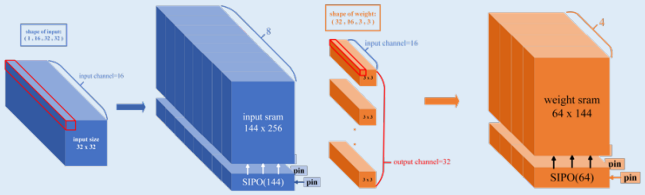

圖 2、SRAM 對應資料數量之大小設計

Sub IMC macro

IMC array

將 **6T SRAM** 中儲存的數值與 **NOR gate** 在**迪摩根定律**的應用下與 Input 相乘。並將此單元擴展成 144 列，每列四個計算單元的 IMC array。

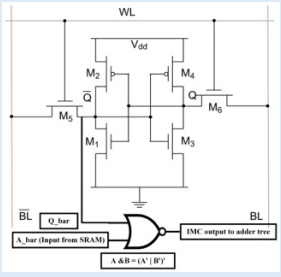

圖 3、6T SRAM 與 NOR 之電路

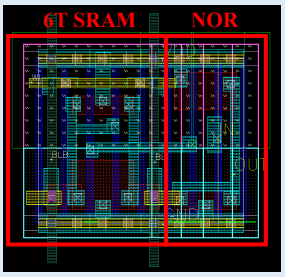

圖 4、6T SRAM 與 NOR 之 layout

Adder tree

將 IMC array 的輸出使用 **28T 全加器**進行樹狀相加，並判斷是進行有號或無號的運算：

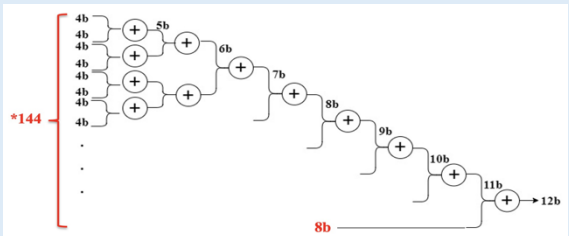

圖 5、144 inputs adder tree 架構

Accumulator

運用 **Robertson algorithm** 進行逐位累加，過程中運用 **counter** 來將 accumulator 中累積的數值清空：

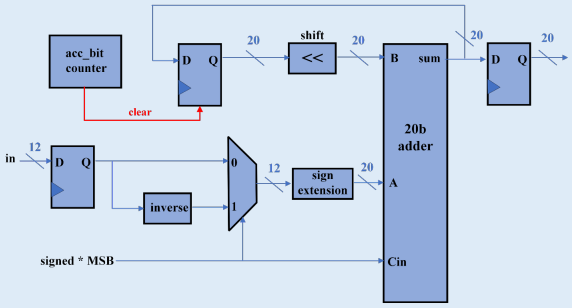

圖 6、accumulator 架構圖

PTSQ

透過調整模型的輸入激活值和權重值的表示範圍，將浮點運算轉換為整數運算，以實現模型壓縮和推論加速。透過 **zero point(Z)**與 **scaling factor** $(M_0 \cdot 2^{(-N)})$，將運算式寫為：

$$\textbf{Data after PTSQ} = \textbf{Data before PTSQ} \cdot (M_0 \cdot 2^{(-N)}) + Z$$

將權重值與輸入激活值提前量化，並把量化後的結果輸出至 output SRAM 中。

實作結果

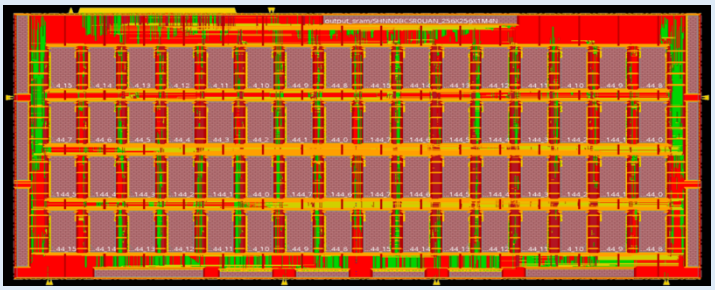

圖 7、IMC macro layout

規格與功能	文獻[1]	實際實現
Input activation	1~8bits	8bits
Weight bits	4,8,12,16	8
Signed and unsigned	Signed and unsigned	signed
Dynamic Voltage Scaling	Yes	No
MAC unit	256*64	144*64
Adder tree	28T and 14T	28T
Concurrent weight update function & MAC operation	Yes	No
F_{max}	100MHz@0.8/0.72V	50MHz@0.72V
Tops/W of single macro	unknown	66.5
Tops/ mm² of single macro	unknown	2.42
Total Tops/W	89	56.6
Total Tops/mm²	16.3	0.68

圖 8、實作結果與文獻[1]之數據比較圖

結論

分析了 IMC 的功耗，結果顯示在 0.72V、40°C 下，我們的設計達到 **56.6 Tops/W** 和 **0.68 Tops/mm²**。雖然低於文獻中提到的 89 Tops/W，但我們找出了功耗優化不足和架構差異等原因。未來計劃引入**動態電壓調節**、**優化加法樹結構**和**佈局設計**，以提升系統性能，期望接近或超越現有的研究成果。

參考文獻

[1] Y. -D. Chih et al., "16.4 An 89TOPS/W and. 16.3TOPS/mm2 All-Digital。SRAM-Based Full-Precision Compute-In Memory Macro in 22nm for Machine-Learning Edge Applications," *2021 IEEE International Solid-State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2021, pp. 252-254

[2] Jacob, Benoit et al. “Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference.” *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2017)*: 2704-2713.