用於邊緣運算的神經網路加速器

Neural network accelerator for edge computing

指導教授：林英超
專題成員：何明堡、陳裕霖、程采婕
開發工具：Python3.8 、Tensorflow2.12、ncverilog、Synopsys Design Compiler、Synopsys IC Compiler
測試環境：Ubuntu 20.04.6

一、簡介：

邊緣運算（Edge Computing）在物聯網時代扮演非常重要的角色，又隨著人工智慧的普及，越來越多的神經網路需要在邊緣裝置上運行。因為能源科技的限制，邊緣裝置需要在有限的功率內完成計算，故本研究將探索如何從無到有，使用軟硬體整合之技術設計出高效能的硬體神經網路加速器。

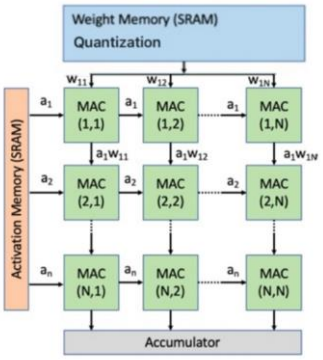
本研究首先運用開源軟體庫（Tensorflow）建立神經網路，並實作神經網路權重量化（Weight Quantization）的演算法，在保持模型精度的情況下，減少加速器所需要讀取的權重個數。接著我們利用 MNIST 手寫資料集以測試量化過後神經網路模型精確度，並將卷基層的權重、輸入、輸出，轉換成十六進位的格式輸出，再以此為測試資料，完成基於脈動陣列的神經網路的硬體加速設計。最後，本研究遵照完整的硬體設計流程以聯電 U18製程進行電路合成及 APR（Automatic Placement & Routing），完整的測試我們所設計的電路，驗證我們提出的脈動陣列架構所需要花費之功耗與面積。

甲、權重量化

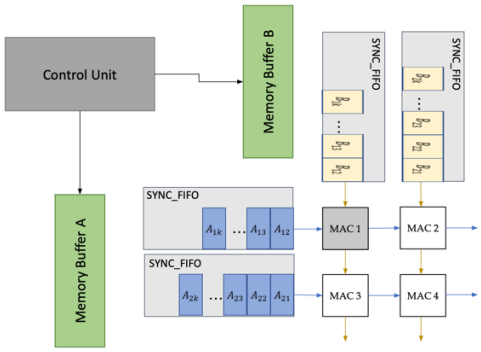
權重量化藉由減少每筆參數大小來節省所需的儲存空間。根據現有研究，在 45nm CMOS 製程中，從 DRAM 存取 32 位浮點數的權重會需要 640pJ 的能量而從 SRAM 讀取僅需要 5pJ。從 DRAM 上存取一筆權重的功耗相當於 SRAM 上的128倍。因此我們需要藉由權重量化，在有限的 SRAM 空間下完成神經網路的計算。

乙、基於脈動陣列的加速設計

脈動陣列（Systolic Array）是一種並行計算架構，常用於加速矩陣和向量運算。本研究利用 Verilog 硬體描述語言，實作一個基於脈動陣列的神經網路加速器。我們的設計如（圖一）所示：神經網路前一層的輸出（Activation）將從左側的記憶體（Activation Memory）送進脈動陣列。量化過後的神經網路權重（Weight Memory）從上方送進脈動陣。如（圖二）所示，本架構利用 Synchronous FIFO 連接記憶體與運算單元（MAC units），此設計方法能夠大幅簡化脈動陣列周邊的電路設計。以下是加速器的系統架構圖：



（圖一）



（圖二）

二、測試結果：

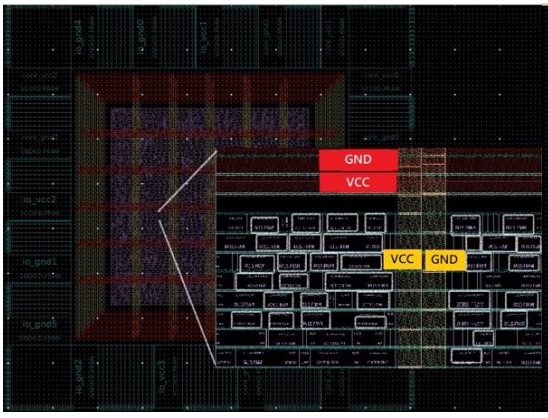
本研究最終目標為探索如何從無到有，使用軟硬體整合之技術，設計出高效能的神經網路加速器，此設計包含了軟體端權重量化以及硬體端的脈動陣列加速器設計實作，其結果如下。（圖四）表示合成後晶片的功耗。（圖五）表示其合成面積為 137,828.68 um^2。（圖六）為包含 I/O Pad 合成的結果，圖中紅色與黃色的線條為提供 std cell 電源的 power stripe，白色方框處為加速器中的 MAC 單元。（圖七）為 Power Netlist Analysis (PNA) Voltage Drop 壓降分析圖，圖中白色接點為連結 power ring 的地方，顏色愈接近紅色代表壓降愈大。本研究使用聯電 U18製程，標準電壓為 1.8V。由（圖七）中可見晶片的壓降都控制在 10% 的標準電壓以下。

```
Global Operating Voltage = 1.62
Power-specific unit information :
    Voltage Units = 1V
    Capacitance Units = 1.000000pf
    Time Units = 1ns
    Dynamic Power Units = 1mW    (derived from V,C,T units)
    Leakage Power Units = 1pW


Attributes
----------
i -- Including register clock pin internal power


  Cell Internal Power  =   2.6415 mW   (79%)
  Net Switching Power  = 690.1647 uW   (21%)
                         ---------
Total Dynamic Power    =   3.3316 mW   (100%)

Cell Leakage Power     = 230.1636 nW


                 Internal         Switching         Leakage           Total
Power Group      Power            Power             Power             Power   (   %    ) Attrs
--------------------------------------------------------------------------------------------
io_pad           0.0000           0.0000            0.0000            0.0000  (  0.00%)
memory           0.0000           0.0000            0.0000            0.0000  (  0.00%)
black_box        0.0000           0.0000            0.0000            0.0000  (  0.00%)
clock_network    2.2421           0.0000            0.0000            0.0000  (  0.00%)  i
register         0.2180           0.1404        1.1866e+05            2.6006  (  78.05%)
sequential       0.0000           0.0000            0.0000            0.0000  (  0.00%)
combinational    0.1813           0.5498        1.1150e+05            0.7312  (  21.95%)
--------------------------------------------------------------------------------------------
Total            2.6415 mW        0.6902 mW     2.3016e+05 pW         3.3318 mW
```

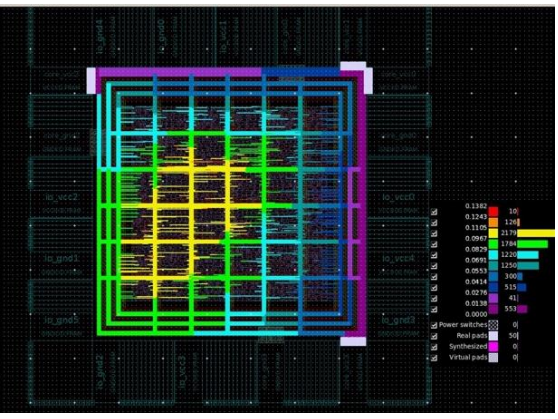
（圖四）功耗分析

```
****************************************
Report : area
Design : tpu
Version: T-2022.03
Date   : Mon May 22 16:57:19 2023
****************************************

Information: Updating design information... (UID-85)
Library(s) Used:

    fsa0m_a_generic_core_ff1p98vm40c (File: /home/nfs_cad/CBDK/CBDK018_UMC_Faraday_v1.1/ \
                                      CIC/SynopsysDC/db/fsa0m_a_generic_core_ff1p98vm40c.db)

Number of ports:                         1899
Number of nets:                          6690
Number of cells:                         4644
Number of combinational cells:           3482
Number of sequential cells:              1103
Number of macros/black boxes:               0
Number of buf/inv:                        624
Number of references:                      62

Combinational area:              72017.265837
Buf/Inv area:                     5127.796732
Noncombinational area:           65811.413322
Macro/Black Box area:                0.000000
Net Interconnect area:      undefined  (Wire load has zero net area)

Total cell area:                137828.679160
Total area:                 undefined
1
```

（圖五） 合成面積



（圖六）包含 I/O Pad 的合成結果



（圖七）PNA Voltage Drop