人工智慧晶片設計與應用

AI-ON-CHIP FOR MACHINE LEARNING AND INFERENCE

PRA 1

E14096724

鄭喆嚴

我選閱讀的論文是 Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks 這篇論文。

Eyeriss 這篇論文的動機是因為卷積神經網絡在現代進行深度學習中取得了很高的的準確性，但是目前最先進的 CNNs 在單次傳遞就需要非常大量的運算，這導致在計算過程中會有大量數據移動進出晶片，因為數據移動消耗的能量比計算本身還多，所以為了提高能源效率，必須設計數據的移動來處理 CNN 的高度並行性和高通量，同時還需要適應 CNN 中不同形狀的高維度卷積。

Eyeriss 這篇論文提出的解決方案是利用一種 CNN 加速器叫做 Eyeriss，Eyeriss 可以讓整個系統的能源效率提升許多，Eyeriss 包含 168 個處理單元 Pes 並且有四層記憶體層級來減少從 DRAM 拿取資料，Eyeriss 採用了一種稱為 Row Stationary 的資料流可以根據不同的 CNN 形狀進行重構，並且使用了一種 network-on-chip (NoC)架構使用多播和點對點單周期數據傳輸來支持 Row Stationary 資料流，Eyeriss 還利用 Run-length compression (RLC) 和 PE 數據閘控制技術，統計 zero 資料來進一步提高能源效率。論文中有進行實際測試和性能分析結果表明，Eyeriss 能夠以高效率處理卷積神經網絡，並且在處理速度、能源效率以及所需的 DRAM 訪問等方面取得了優異的表現。

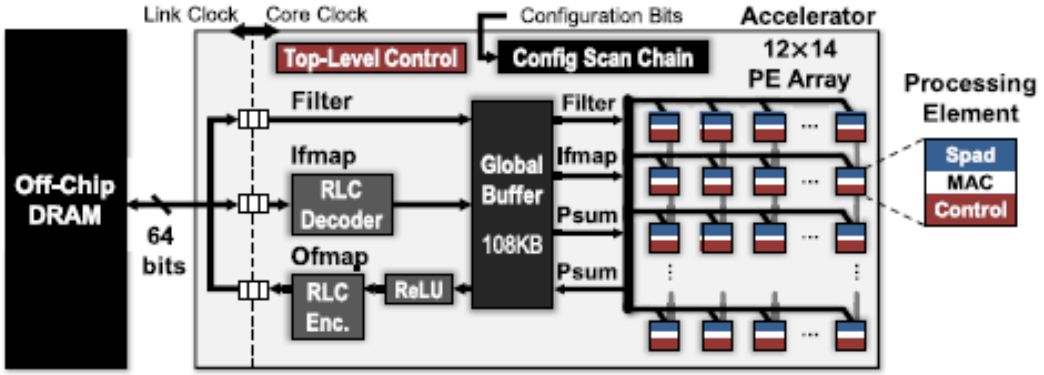

Fig. 2. Eyeriss system architecture.

我認為對於卷積神經網絡在傳遞過程中需要大量的數據移動和計算，導致了高能耗且能源效率低，使用特別設計過的加速器是非常合理的，像是 Row Stationary 的資料流、四層記憶體層級和 network-on-chip (NoC)架構都是設計來解決卷積神經網絡大量的數據移動導致的能源效率低的問題。我覺得這篇論文未來的方向可以是提升數據和權重的重用性或是加強數據和權重的壓縮處理來支持稀疏矩陣的運算，解決這兩個問題又可以在能源效率上有更大的進步，這兩個問題也是 eyeriss V2 也就是 eyeriss 的升級版本所解決的問題。