# 人工智慧晶片設計與應用

AI-ON-CHIP FOR MACHINE LEARNING AND INFERENCE

PRA 3

E14096724

鄭喆嚴

我選擇閱讀的論文是 SCNN: An Accelerator for Compressed-sparse Convolutional Neural Networks 這篇論文。這篇論文的動機是對於稀疏的 CNN 利用 activations 和 pruned weights 的稀疏性，來盡可能地減少的計算週期、數據移動和記憶體的存取。

在論文中有提到多個解決的辦法，像是 SCNN 有對稀疏資料進行壓縮編碼，將稀疏的 activations 和 pruned weights 進行壓縮編碼成向量，這種方法不僅減少了不必要的零的資料傳輸，也降低了零所佔的儲存空間，提高了整體的計算效率；SCNN 也有較高效的資料傳輸，SCNN 的架構有助於將稀疏的 activations 和 pruned weights 更有效傳輸到 PE 陣列中，並在 PE 陣列中進行重複使用，這更好利用了計算資源並減少多餘記憶體存取的操作；SCNN 還使用笛卡爾積資料流來計算非零的 activations 和 pruned weights 的向量相乘，這種資料流特別設計來避免零的運算，減少許多不必要的算術運算，SCNN 用笛卡爾積資料流實現高效的計算。

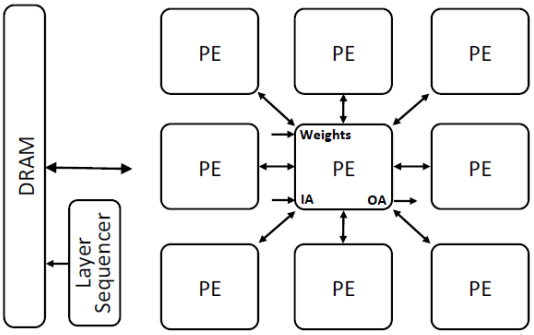



Figure 5: Complete SCNN architecture.

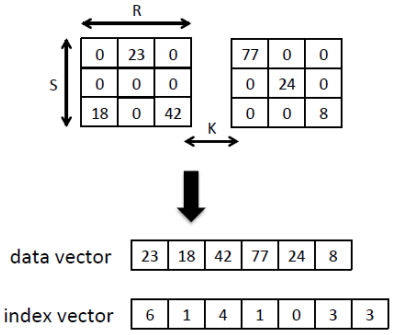
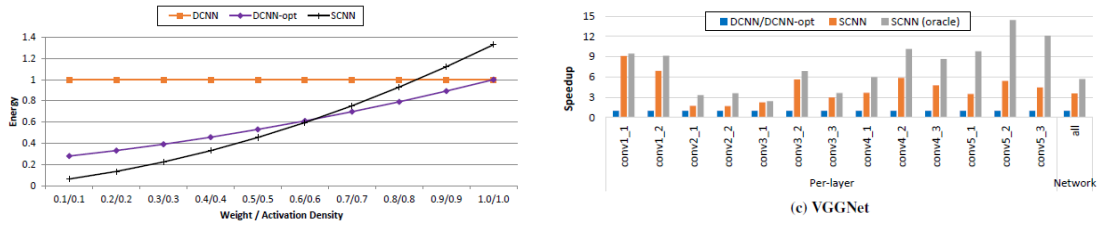



Figure 7: Weight compression.

SCNN 在對於稀疏性敏感度的評估上先進行了 SCNN、DCNN 和 DCNN-opt 三種架構能源效率的比較，發現 SCNN 的架構對於稀疏性的表現比 DCNN 和 DCNN-opt 還要更好，但是在比較不稀疏的情況下 SCNN 的能源效率是最差的；在性能的評估上進行和 DCNN 架構在 AlexNet、GoogLeNet 和 VGGNet 上表現的比較，SCNN 架構的性能表現始終優於 DCNN 架構，分別實現了平均 2.37 倍、2.19 倍和 3.52 倍的性能改進。




(c) VGGNet

我認為在稀疏性較高的情況下 SCNN 確實能夠達到很好的性能和能源效率的表現，在非常稀疏時 SCNN 的表現是非常好的，但在稀疏性沒那麼高的時候或是不稀疏的時候 SCNN 架構的表現其實是比其他的架構還要差的，並且在這篇論文中蠻多的評估都是用很理想的情況下去進行計算，所以不知道在現實中不是像論文中非常稀疏的情況下會比其他架構優化多少，但可以確定的是 SCNN 架構在稀疏性較高的情況下的性能和能源效率表現是非常好的。

我覺得這篇論文的未來發展方向可以進行再更進一步的性能優化，論文中的架構是為了對稀疏性進行優化，雖然 SCNN 已經在性能和能源效率方面有不錯的成果，但未來的研究仍可以尋找更有效的稀疏性優化算法，可以犧牲更多在稀疏性較低時的表現，再進一步提性能和能源效率；或是可以往其他應用領域發展，論文中比較專注在圖像處理中的 CNN 加速器，但是稀疏性優化方法可能在其他領域也適用，未來可以探索在其他應用領域中對稀疏性進行優化的潛力。