16 奈米製程實現之 MobileNetV2 專用雙引擎高效能 CNN 加速器

A High-Performance Dual-Engine CNN Accelerator for MobileNetV2 in 16-nm Technology

組別:晶片系統組　　指導教授:王進賢 特聘教授　　組員:翁尹希、王翔則、邱昱甄

摘要

卷積神經網路 (CNNs) 已廣泛應用於電腦視覺與邊緣裝置，但仍受限於龐大的運算量與記憶體需求。MobileNetV2 透過深度可分離卷積 (DSC) 有效降低參數量與計算複雜度，其中逐點卷積 (PWC) 與深度卷積 (DWC) 之間的運算負載與特性差異過大，現有多數加速器未能充分利用其分離結構特性，導致 DWC 階段的運算單元使用率偏低。為解決此問題，本研究提出一款專為 MobileNetV2 優化的雙引擎加速器，利用 PWC–DWC 並行處理結合權重預載機制，降低中間結果記憶體開銷，並提升整體效能。本架構可達 98.3% 的運算單元利用率，在 16 奈米 FinFET 製程下以 588 MHz 運作時可達 1712.3 GOPS 之吞吐量，相較單引擎基準架構吞吐量提升 139%。該設計在維持高能效的同時，大幅提升硬體利用率，適用於邊緣裝置上的 MobileNetV2 推論。

硬體架構

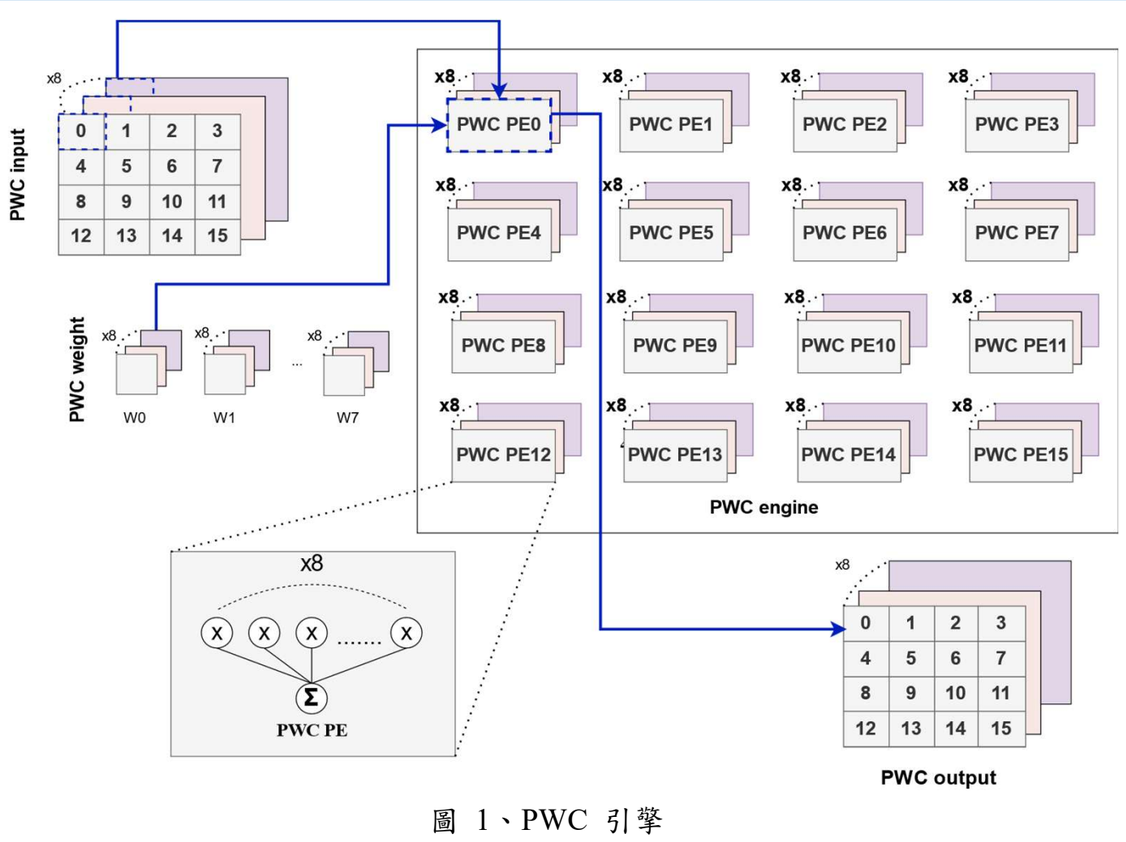


圖 1、PWC 引擎

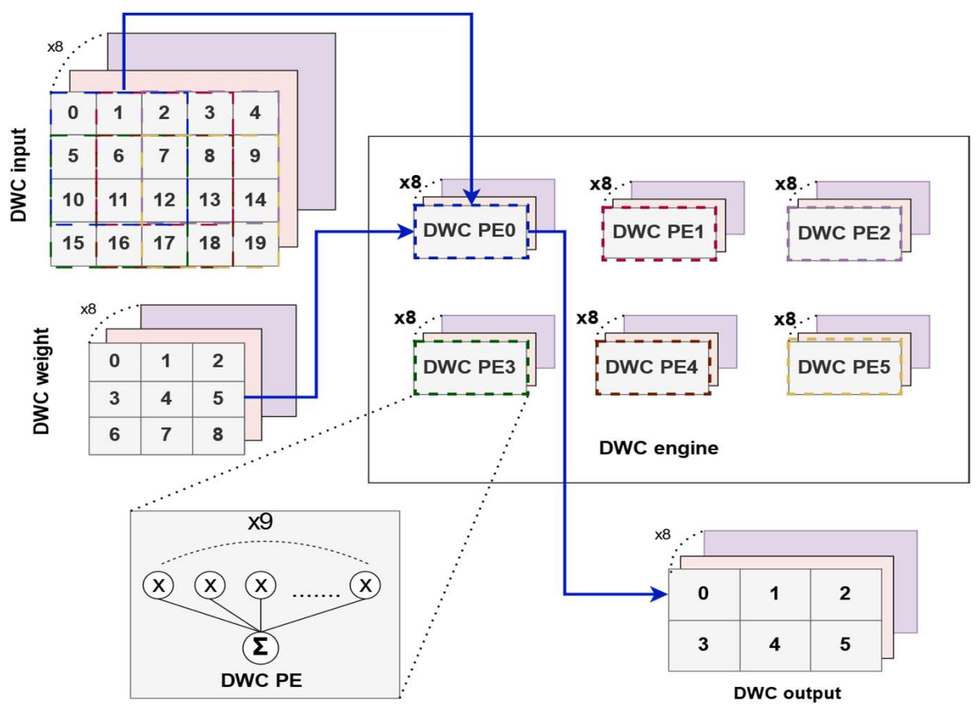


圖 2、DWC 引擎

雙引擎設計

PWC engine:

空間對映上，16 個 PE 對映到輸入特徵圖的 4×4 個空間位置；通道方向上，同時計算 8 個輸出通道。每個時脈可產生 4×4×8 的輸出。

DWC engine:

空間對映上，每個 PE 內含 9 個乘法器對映 3×3 的權重；通道方向上，同樣可同時計算 8 個輸出通道。每個時脈可產生 2×3×8 的輸出。

雙引擎設計可以減少 PWC engine 等待 DWC engine 計算的時間，維持良好硬體利用率。

權重預載

為緩解 SRAM 頻寬限制，在權重記憶體與計算引擎之間插入暫存器作為緩衝，其中儲存的是計算過程中會被重複使用的權重，藉此降低記憶體到運算引擎的延遲，並提供高頻寬的權重讀取通道，提升吞吐量。

資料流

為最大化權重重用，本設計採用 spatial-first 策略，PWC 先計算 RC 方向，再計算 K 方向。與 channel-first 相比，可將 DWC 的權重存取次數減少為原本的 $(N\times M)/(T_N\times T_M)$ 倍，顯著降低記憶體存取開銷並提升資料重用率。

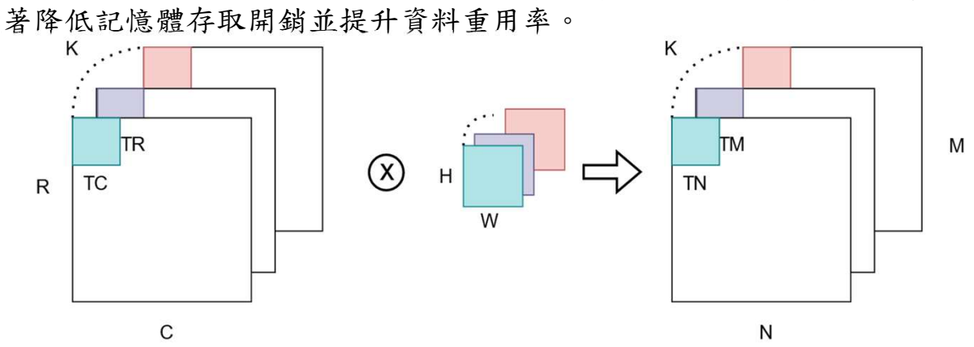


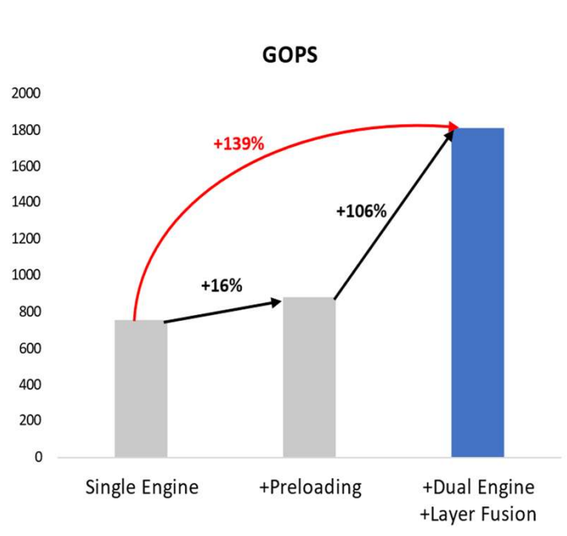
圖 3、DWC 層之分塊策略示意圖

表 1、不同 DWC 排程策略下的輸入與權重之記憶體存取次數比較

Engine	Scheduling	Input Access	Weight Access
DWC Engine	channel-first	$N\times M\times K$	$\frac{H\times W\times K\times N\times M}{T_N\times T_M}$
	spatial-first	$N\times M\times K$	$H\times W\times K$

實作結果

單引擎架構加入權重預載機制後，吞吐量提升 16%，顯現降低記憶體存取延遲的效益。在此基礎上，進一步採用並行化的雙引擎架構，相較於權重預載設計可再提升 106%。整體而言，與單引擎基準相比，本研究的雙引擎並行化設計可提升 139% 的吞吐量，展現了結合架構並行化與記憶體最佳化的設計優勢。



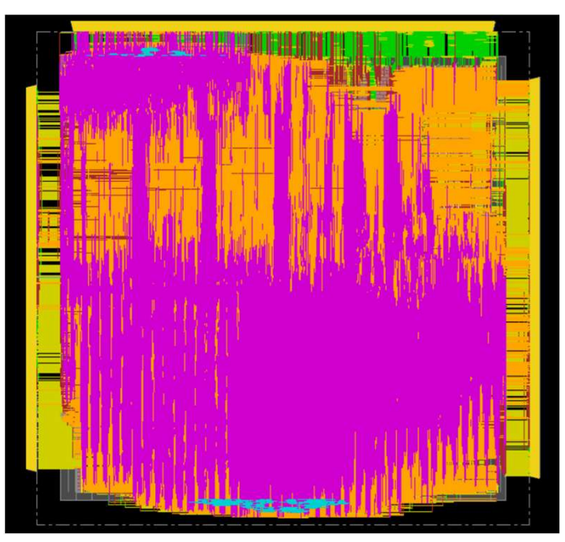


圖 4（左）、雙引擎、權重預載與層融合實現之吞吐量提升
圖 5（右）、本研究之晶片佈局

在 16 nm 製程下，本設計可達到 1712.3 GOPS 與 10.1 TOPS/W 的能效，驗證了雙引擎架構搭配記憶體最佳化在效能與能效上的優勢，展現專用加速器在行動與邊緣裝置推論應用上的實際價值。

表 2、實作結果與既有設計比較

		ISSCC'20 [1]	ISCAS'21 [2]	ISSCC'23 [3]	This Work
Engine Type		Dual-core shared array	Single-engine	Dual-engine	Dual-engine
Benchmark Network		MobileNetV1	MobileNetV2	MobileNetV2	MobileNetV2
Techonology (nm)		7	40	28	16
Core Area (mm^2)		3.04	1.03	7.81	2.16
Supply Voltage (V)		0.575 ~ 0.825	0.85	0.66 ~ 1.3	0.72 ~ 0.88
Frequency (MHz)		290 ~ 880	100	100 ~ 500	303 ~ 588
Number of MACs	PWC	1024	72	1024	1024
	DWC	1024	72	128	432
	Total	2048	72 (No dedicated PWC/DWC)	1152	1456
MAC Utilization		N/A	N/A	97.3%	98.3%
Throughput (GOPS)		3604	52.8	1024	1712.3
Power (mW)		174 ~ 1053	18.3	17 ~ 174	87.1 ~ 271.4
Normalized Energy Efficiency (TOPS/W)		2.8	7.2	13.1	10.1

參考文獻

[1] C.-H. Lin et al., "A 3.4-to-13.3TOPS/W 3.6TOPS Dual-Core Deep-Learning Accelerator for Versatile AI Applications in 7nm 5G Smartphone SoC," ISSCC, pp. 134-135, 2020.

[2] Y. S. Chong et al., "An Energy-Efficient Convolution Unit for Depthwise Separable Convolutional Neural Networks," 2021 IEEE International Symposium on Circuits and Systems (ISCAS), Daegu, Korea, 2021, pp. 1-5.

[3] C. -Y. Du et al., "A 28nm 11.2TOPS/W Hardware-Utilization-Aware Neural-Network Accelerator with Dynamic Dataflow," 2023 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 2023, pp. 1-3.