D23-033

適用於高品質且高解析度智能影像處理應用之高效節能CNN處理器

An Energy-efficient CNN Processor for High-quality and High-resolution Intelligent Image Processing

隊伍名稱｜心如止水
VCSLab
隊　　長｜丁友鈞 / 清華大學電機工程研究所
隊　　員｜林楷平 / 清華大學電機工程研究所
林俊曄 / 清華大學電機工程研究所
陳永泰 / 清華大學電機工程研究所

指導教授

黃朝宗｜清華大學電機工程學系

臺灣大學電子工程博士，現為清華大學電機工程學系副教授。曾服務於聯誌科技，亦曾於麻省理工學院進行博士後研究。曾獲傑出人才基金會年輕學者創新獎、未來科技獎、中國電機工程學會優秀青年電機工程師獎、清華大學傑出教學獎、旺宏金矽獎最佳指導教授獎等獎項。

研究領域

近來研究以實現高效能、高品質之腦視覺與計算攝影學應用為主，包含卷積神經網路處理器、立體 3D 光場顯示器、光場相機等相關研究，是國內極少數能同時發表頂尖論文至計算機架構（ISCA/MICRO）、晶片設計（ISSCC/VLSIC/ESSCIRC）、電腦視覺（CVPR/ICCV/TPAMI）此三大熱門研究領域之研究者。

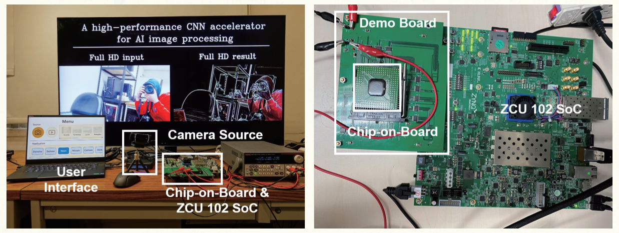
作品摘要

卷積神經網路（Convolutional Neural Network，CNN）為專用於智慧視覺的神經網路，在許多影像成像畫質修復的應用如影像去噪（Image Denoising）或超解析（Super-resolution）皆取得卓越成果，帶來更高品質視覺感受，甚至可提供傳統演算法難以完成的新穎應用，如影像風格轉換（Style Transfer）。因此若能有效地在手機或電視等嵌入式電子產品上實現智能網路的即時運算處理，將有機會於影像擷取及播放設備上帶來新一波品質革命。然而，運算影像處理應用的CNN模型需要龐大的DRAM頻寬以及極高的運算能力，這對於，這對要求低頻寬、低功耗的嵌入式電子產品是相當大的挑戰，使這項技術難以被廣泛應用。

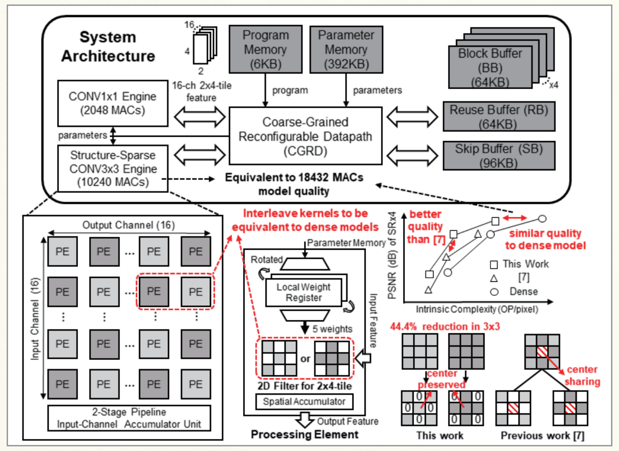
有鑑於此，本計畫開發了一個高效節能的智能影像處理CNN加速器晶片來克服這個困難。本晶片透過模型演算法及系統電路架構協同設計的方式進行功耗與頻寬的最佳化，在嵌入式裝置的資源限制下完成高品質高畫質的智能影像處理應用，加速智慧視覺技術應用落地提供下世代影像處理器設計更多可能；除此之外，如圖一，我們也將此加速器晶片設計與Xilinx ZCU102 SoC開發板整合完成即時展示系統，直接透過影像輸出多種高品質智慧視覺應用。

圖二為本晶片設計之系統架構。針對高品質深層模型與高畫質影像帶來的大量資料頻寬存取，本設計提出了一混合式層融合記憶體系統架構，有效降低記憶體使用成本；此外，有別於前作為了利用預訓練模型不規律的動態稀疏性降低功耗而引入複雜的控制電路，本設計透過軟硬體協同設計的方式提出了一個具結構性內核稀疏特徵的超高平行的運算引擎，在大幅降低運算能源消耗的同時不需額外的電路成本與複雜的控制，達到運算吞吐量與能源效率的最佳化。

本晶片設計透過40nm製程完成下線，晶片核心面積（core area）為3.2x3.2mm²，整合了8.0M的邏輯閘（logic gate count）及814KB的SRAM使用量。最高可輸出達8 TOPS之等效運算吞吐量，能在7.6 TOPS/W的能源效率下完成4K Ultra-HD 30fps規格的高品質的深層模型運算，並且所需之DRAM頻寬小於2GB/s。我們期許透過此晶片設計與配套展示系統，加速高品質智慧影像處理技術應用普及，提供下世代影像處理器設計更多可能。



圖一 本設計晶片與即時晶片展示系統。



圖二 本設計系統架構圖。

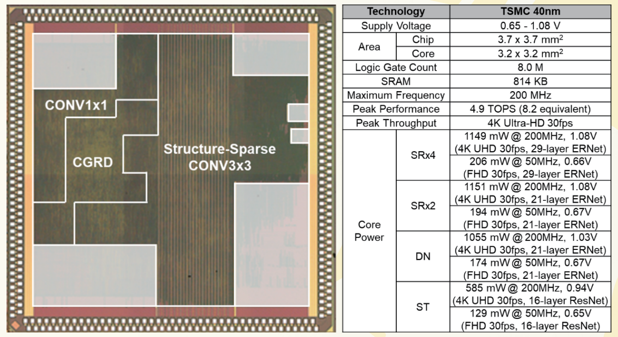
Abstract

Convolutional Neural Network (CNN) is a specialized type of neural network designed for computational imaging. It has achieved remarkable results in various applications related to image processing and quality enhancement, such as image denoising and image super-resolution. Moreover, CNNs can provide higher-quality visual experiences and can even enable novel applications that traditional algorithms hard to accomplish, such as image style transfer. Therefore, CNN has the potential to bring about a new wave of quality revolution in image capture and displays while implementing real-time computational processing of intelligent networks can be implemented on embedded system. However, CNN models for computational imaging applications require massive DRAM bandwidth and extremely high computation power. This poses a significant challenge for embedded devices that have low bandwidth and power consumption requirements. As a result, this technology is difficult to be widely adopted in such applications.

This work aims to develop a memory and energy efficient image processing CNN accelerator chip to overcome these challenges. We adopt collaborative design approach that combines model algorithms and system circuit architecture to optimize bandwidth and power. It successfully accomplishes high-quality and high-resolution intelligent image processing applications within the resource constraints of embedded devices. Furthermore, we have also integrated this accelerator chip with the Xilinx ZCU102 SoC development board to create a real-time demonstration system, and directly displayed various high-quality intelligent visual applications.

To address the significant bandwidth access required by high-quality deep models and high-resolution images, this design proposes a hybrid layer fusion inference flow that effectively reduces memory usage costs. Additionally, unlike previous works that introduced complex control circuits to leverage the irregular dynamic sparsity of pretrained models for power reduction, this design employs a structured sparse interleaved-kernel highly parallel convolution engine through algorithm-hardware co-optimization. This approach enables high computation throughput and energy efficiency at the same time, substantially reducing computing energy without additional circuit costs or complex control requirements.

This chip is fabricated in TSMC 40nm CMOS process. Fig. 3 shows the die photo and the specification of this chip. Its core area is about 3.2x3.2mm² with 814 KB of SRAM usage. This chip can achieve a maximum computational throughput of 8 TOPS (tera operations per second). It can efficiently perform high-quality deep model computations at 4K Ultra-HD 30 fps, achieving an energy efficiency of 7.6 TOPS/W. Additionally, the required DRAM bandwidth is less than 2GB/s. With this chip design and its accompanying demonstration system, we aim to accelerate the adoption of high-quality intelligent image processing technologies and provide more possibilities for the design of next-generation image processors.


CONV1x1
CGRD
Structure-Sparse CONV3x3

Technology		TSMC 40nm
Supply Voltage		0.65 - 1.08 V
Area	Chip	3.7 x 3.7 mm²
	Core	3.2 x 3.2 mm²
Logic Gate Count		8.0 M
SRAM		814 KB
Maximum Frequency		200 MHz
Peak Performance		4.9 TOPS (8.2 equivalent)
Peak Throughput		4K Ultra-HD 30fps
Core Power	SRx4	1149 mW @ 200MHz, 1.08V (4K UHD 30fps, 29-layer ERNet)
		206 mW @ 50MHz, 0.66V (FHD 30fps, 29-layer ERNet)
	SRx2	1151 mW @ 200MHz, 1.08V (4K UHD 30fps, 21-layer ERNet)
		194 mW @ 50MHz, 0.67V (FHD 30fps, 21-layer ERNet)
	DN	1055 mW @ 200MHz, 1.03V (4K UHD 30fps, 21-layer ERNet)
		174 mW @ 50MHz, 0.67V (FHD 30fps, 21-layer ERNet)
	ST	585 mW @ 200MHz, 0.94V (4K UHD 30fps, 16-layer ResNet)
		129 mW @ 50MHz, 0.65V (FHD 30fps, 16-layer ResNet)

Fig.3 Die photo and chip specification.