



D23-068

應用於擴增實境行動運算之異質卷積 神經網路深度學習系統晶片設計與實現

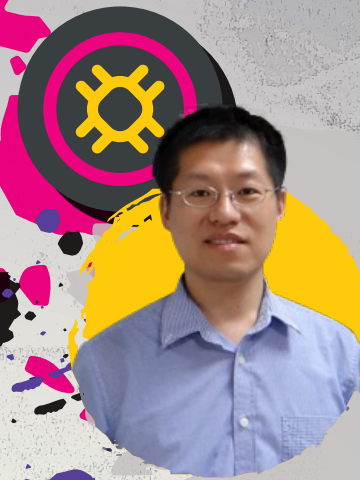
Design and Implementation of a CNN-GCN Deep
Learning SoC for Mobile Augmented Reality

隊伍名稱 | 龍山寺板橋一家親

The Longshan Temple Banqiao Family

隊長 | 林奕廷 / 臺灣大學電子工程學研究所

隊員 | 林瑩昇 / 臺灣大學電子工程學研究所



指導教授

楊家驥 | 臺灣大學電機工程學系

美國加州大學洛杉磯分校電機博士，現為臺灣大學電機工程學系教授。實驗室致力於開發低功耗之客製化晶片以提升資料處理速度與能量效率。

研究領域

AI晶片設計、基頻通訊積體電路、生醫訊號處理晶片設計

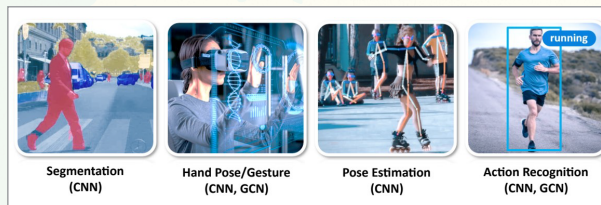
作品摘要

近年來隨著擴增實境的蓬勃發展，我們能在現實與虛擬世界間進行互動體驗。在擴增實境應用中，電腦視覺演算法被使用來從視覺影像中提取更高級別的資訊。其中，卷積神經網路被用於提取視覺特徵，如物體偵測、語義分割、身體/手勢姿態估計和動作識別。此外，圖卷積網路在基於人體骨架的動作識別模型中扮演重要角色。儘管使用繪圖處理器（GPU）可以實現即時的卷積和圖卷積網路運算，但高功耗使其在移動裝置上的應用受限。因此，本作品旨在為卷積和圖卷積網路設計一個高性能的擴增實境加速器，以滿足在行動設備上的計算需求。

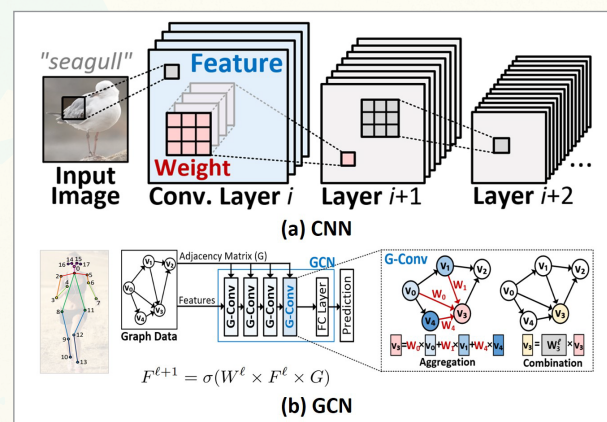
這次參賽的作品展示了一個專為卷積神經網路以及圖卷積網路設計的高能效加速器。在卷積神經網路方面，我們運用網路中特徵張量在通道方向上的稀疏性，設計了高效的稀疏張量編碼方法。不僅可以在模型準確度損失極小（<1%）的情況下配置不同的稀疏度，也能在不同稀疏性下都達到100%的硬體使用率。另外，本作品提出的資料流架構能夠有效重複使用特徵資料，減少高達75%之記憶體存取。在圖卷積網路方面，我們運用人體骨架圖的特性和矩陣轉換技巧，將模型的整體運算量和參數量分別降低71%和39%。本作品提出的圖卷積引擎能夠有效利用骨架圖中的稀疏特性減少最多84%的處理時間。而我們設計的快速矩陣處理單元則可以減少69%的矩陣乘法運算量和計算時間。

本設計使用28nm的製程，晶片核心面積為8.28 mm²。操作電壓在0.81V、200MHz的時序頻率下，此晶片對於稀疏卷積能達到3277 GOPS的峰值效能。此外本晶片最高能達到25.1 TOPS/W的能量使用效率。針對人體動作識別的演算法，此晶片能夠達到最快每秒判斷72個動作的吞吐量，並且達到0.89 actions/mJ的能量使用效率。相比於過去的最佳設計，在卷積加速器方面，本作品達到了2.3倍

的能量效率的提升。而在人體動作辨識方面，本設計在提升18倍的識別吞吐量的同時，也將能量使用效率提升了5.4倍。



圖一 擴增實境應用。



圖二 卷積神經網路與圖卷積網路模型。

Abstract

In recent years, with the flourishing development of augmented reality (AR), we are able to have interactive experiences between the real and virtual worlds. In AR applications, computer vision algorithms are used to extract higher-level information from camera images. Convolutional neural networks (CNNs) are used to extract visual features such as object detection, semantic segmentation, body/hand pose estimation, and action recognition. Additionally, graph convolutional networks (GCNs) play an important role in skeleton-based action recognition models. However, despite the ability to achieve real-time inference for CNNs and GCNs using graphics processing units (GPUs), their high power consumption limits their applicability in mobile devices. Therefore, this project aims to design an energy-efficient accelerator for CNNs and GCNs for AR applications.

This project presents an energy-efficient accelerator specifically designed for CNNs and GCNs. For CNNs, we utilize the channel sparsity in feature tensors and design an efficient sparse tensor encoding method. This method allows for different sparsity configurations with minimal model accuracy loss (<1%). The proposed dataflow architecture achieves 100% hardware utilization under different sparsity levels and efficiently reuses input features to reduce up to 75% memory access. For GCNs, we leverage the characteristics of skeleton graphs and matrix transformation technique to reduce the overall model computation and parameters by 71% and 39%, respectively. The proposed GCN engine leverages the sparsity in skeleton graphs to reduce up to 84% of the processing time, and the fast matrix multiplication core reduces the number of matrix multiplication operations and latency by 69%.

The design is implemented using a 28nm CMOS technology, with a core area of 8.28 mm². When operating at a voltage of 0.81V and a clock frequency of 200MHz, this chip achieves a peak performance of 3277 GOPS for sparse convolution. It also delivers an energy efficiency of up to 25.1 TOPS/W. For human action recognition tasks, the chip achieves a maximum throughput of 72 actions per second and an energy efficiency of 0.89 actions/mJ. Compared to previous

state-of-the-art designs, this work achieves a 2.3x improvement in energy efficiency for the CNN accelerator. As for human action recognition, this work achieves a 18x increase in recognition rate, while improving energy efficiency by 5.4x.

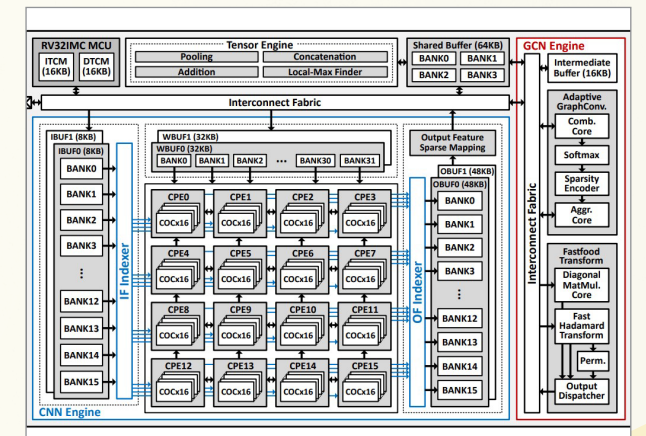


Fig.3 System architecture.

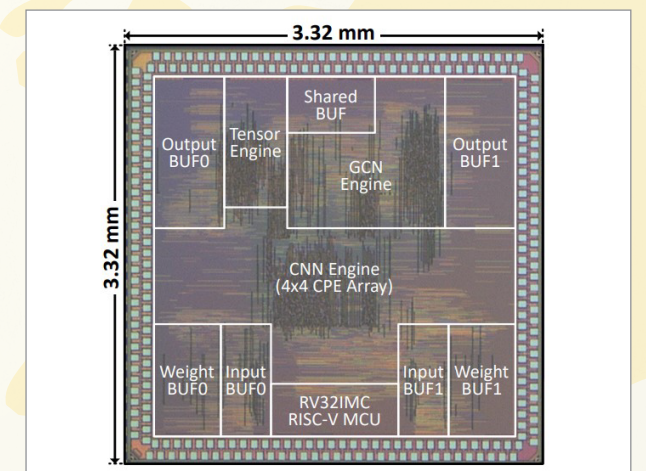


Fig.4 Die photo.