D20-023 **DESIGN GROUP**

作品名稱

混合卷積神經網路硬體加速器系統設計與其模型訓練分析工具

Hybrid CNN Accelerator System Design and the Associated Model Training/Analyzing Tools

隊伍名稱

留言辦抽獎 Liu Yan Ban Chou Jiang

隊長

蔡家齊 交通大學電子研究所

隊員

曾建霖 交通大學電子研究所
孔　睿 交通大學電子研究所
張恩誌 交通大學電子研究所

作品摘要

深度學習近幾年來於 AI 應用中表現相當搶眼，透過學習物件特徵後再實際推論，在物件分類、物件偵測、語義分割等技術上有所突破，並且相關應用已實際應用於工廠產線檢測輔助、自駕車駕駛輔助系統技術、智能 3C 等產品中。而在終端裝置平台上為了進行高效率的處理 CNN 模型運算，通常會使用 DSP、GPU 或是 NPU 來加速，並且不同於 CNN 模型訓練時使用的 floating-point 資料格式，在 inference 時這些 CNN 模型運算都會是以 fixed-point 資料格式來進行，才能夠確實達到低功耗與高效能之設計目標。

本作品提出了一個能夠支援混合型卷積神經網路的 Hybrid CNN 硬體加速器系統設計與其模型訓練與分析工具，其整體系統如圖一所示，主要可以分為軟體端與硬體端。在硬體端本作品提出了一個可以動態支援 Hybrid fixed-point layer 運算的 CNN 硬體加速器以及其相對應的 Hybrid CNN Model Parser/Compiler。本作品之硬體加速器共使用了 638 KB 的 SRAM，高資料重複使用率的資料流以及 Ping-pong 資料緩衝器的預存機制讓本團隊所設計之 Hybrid fixed-point CNN 硬體加速器提供了優異的計算效能與硬體使用效率。在計算單元的配置上，本作品採用了 6x16 大小的 PE Array，並且提供了彈性的 PE Array 配置方式，讓常見的 CNN Layer 都能夠有效配置到本作品硬體加速器上執行運算，本作品時脈運作於 200MHz 時（以 TSMC40nm 製程實作時），可處理 8-bit feature map/8-bit weight 卷積運算之最高效能（Peak performance）可以達到 921.6 GOPS，而若是進行 1-bit feature map/1-bit weight 運算時，其 Peak performance 則可達到 7.37 TOPS。在 TSMC 40nm 的製程條件下本作品晶片面積約為 4,514 x 4,515 um^2，在 200MHz 的時脈頻率下，本作品的功率消耗為 559.89mW。

在 Hybrid fixed-point CNN 模型訓練上，本作品透過 Knowledge Transfer 與 Dynamic Quantization 的方法，在 MobileNet-SSD 模型實例下之模型大小壓縮率可以達到 91%，而在 VGG16 的模型實例下則可以達到 52% 的模型大小壓縮率。另外在軟、硬體整合驗證上，本作品提供所設計之 Hybrid CNN 硬體加速器相對應的 Hybrid CNN Model Parser/Compiler，目前支援原生的 BVLC caffe 框架以及本作品基於 caffe 所開發的 Hybrid CNN 模型訓練、分析工具，在 inference 時，本作品的軟、硬體運算結果可以達到 bit accurate 準確度（即軟體模擬結果與硬體執行結果完全相同），不會因為模型從 FP32 的運算平台移植到定點數運算平台而造成運算結果不同而導致其準確度下降。在整體軟、硬體系統整合上，本作品已具備完整性，目前透過 Xilinx ZCU-102 FPGA 平台的系統驗證，本作品在 150 MHz 的運算時脈下可以達到 Tiny Yolov2 模型 30FPS 的即時運算效能。

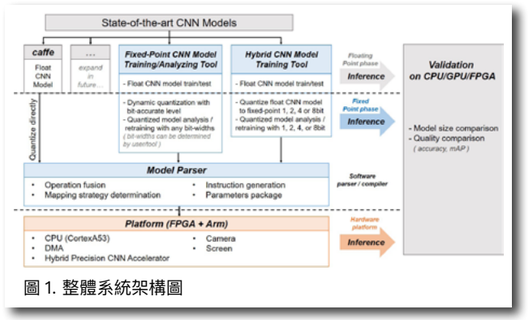

圖 1. 整體系統架構圖

指導教授

郭峻因 交通大學電子工程學系

- 交通大學電子博士，現為交通大學電子工程學系特聘教授。曾任聯合技術學院電子工程系主任、中正大學 SOC 研究中心主任 / 特聘教授兼系主任、交通大學電子研究所所長 / 晶片系統設計中心主任 / 交通大學電機學院副院長。
- **研究領域：**超大型積體電路設計、數位訊號處理、數位 IP 及 SoC 設計、智慧視覺處理

Abstract

In recent years, deep learning technology has been applied to many noticeable AI applications. Deep learning technology does well in object features, object detection, semantic segmentation and other emerging applications. It is also applied to factory production line inspection assistance, ADAS/Self-driving car applications, intelligent 3C products, etc. In order to efficiently inference CNN model on the edge device platform, DSP, GPU or NPU are usually used to accelerate these operations. CNN model inferencing is different to the model training. It usually adopts the fixed-point data type in CNN model inferencing to achieve both the low computation cost and high performance.

This work proposes a Hybrid CNN hardware accelerator system design and the related model training/analyzing tools that can support a hybrid fixed-point convolutional neural network. The Hybrid CNN Inferencing system is shown in Fig. 1, which can be divided into software and hardware parts. On the hardware side, this work proposes a CNN hardware accelerator and the corresponding Hybrid CNN Model Parser/Compiler that can dynamically support the Hybrid fixed-point layer operations. The proposed hardware accelerator uses 638KB of on-chip SRAM, having dataflow with high data reuse rate and the ping-pong buffer pre-storage mechanism to make the Hybrid fixed-point CNN hardware accelerator providing high computing performance and high hardware utilization rate. This work equips 6x16 PE array and a flexible PE Array configuration method, such that the common operations of CNN layer can be efficiently mapped to the hardware accelerator of this work. This design operates at 200MHz when implemented in TSMC 40nm Technology. It offers 921.6 GOPS peak performance when operating 8-bit feature map/8-bit weight convolution and 7.37 TOPS peak performance when operating 1-bit feature map/1-bit weight convolution. Under the TSMC 40nm technology, the area of the proposed design is 4,514 x 4,515 um^2 and the power consumption is 559.89mW.

In the software side, we train the Hybrid fixed-point CNN model through the Knowledge transfer and dynamic quantization according to our proposed tool, ezHybrid-M. It reduces 91% model size for MobileNet-SSD test case and 52% model size reduction for VGG16 test case. The highly model compression rate makes the model inferencing more low power and low cost. For the hardware and software integration and verifications, our CNN accelerator and the model parser/compiler support the native BVLC caffe framework and our CNN model training/analyzing tool, which is developed base on the caffe. By using our training tool, the calculation results of software framework and hardware accelerator will be exactly the same in bit accurate level. The accuracy will not be degraded when the CNN model is porting from the floating-point32 data type training framework to the fixed-point data type hardware accelerators. For the integration of hardware and software, the proposed design has been verified on the Xilinx ZCU102 FPGA platform. It can achieve the real-time computing performance 30fps of Tiny Yolov2 model when it is operated at 150MHz.