節能系統設計之低功耗運算核心

捨棄加速器！ Tensilica 獨立自含式 DSP IP 更有效率

■文：任芷萍

網路直播的浪濤泉湧，數千人同時在線上閱聽影音內容已屬司空見慣，行動終端的運算能力亦須跟上時代，視覺串流的處理尤其備受關注；加上監控和穿戴式裝置以全年無休的「不斷線」(always-on)為發展職志，若無法壓低功耗，電力恐撐沒多久就玩完了。一般保全監控視訊、汽車光達 (LiDAR)/ 雷達、無人機和感測器融合等應用的視覺系統需要兩種優化運算：首先，運用傳統運算攝影 / 成像演算法對來自攝影機的輸入進行強化，其次，由神經網路的辨識演算法執行物體偵測和辨識。為達極致省電目的，從「IP 核心」根本革新有其必要性。

照片人物：Cadence Tensilica 處理器事業群資深總監 Steve Roddy

不只卷積層！突破 NN 引擎加速器極限，Vision C5 可加速所有運算架構

神經網路 (NN) 已成深度學習顯學，但運算極具挑戰。益華電腦 (Cadence) 旗下 Tensilica 新近發佈的 Vision C5，顛覆同業在影像數位訊號處理器 (DSP) 綑綁「NN 硬體加速器」(accelerator) 的作法，是業界首款真正專為 NN 獨立運作而生、被稱為「獨立自含式」的 DSP IP。Cadence Tensilica 處理器事業群資深總監 Steve Roddy 指出，早先 DSP+NN 引擎的方式乃將神經網路編碼分割處理，不斷在 DSP 的網路層與加速器的卷積層之間加載、卸載，而將其他層級的運算工作全數丟給主要 DSP/CPU/GPU 一肩獨攬。

「如此一來，不僅執行效率不佳、且會造成不必要的耗電」，Roddy 直戳 DSP+NN 引擎的痛點。他深入解說，如果 NN 架構的神經元 (Neurons) 數量增加，其間鍵結也會隨之平方增加；若利用硬體加速 NN 的運算速度，所需硬體結構的複雜度將大幅增加而變得不容易實現。相較之下，新款 Vision C5 所建構的「通用型」神經網路 DSP，可加速所有神經網路運算架構，包括：卷積 (Convolutional)、全連接 (Fully connected)、池化 / 取樣 (Pooling) 及標準化 (Normalization)，以精算「型態辨識」(Pattern recognition) 與相鄰

資料間的關係。

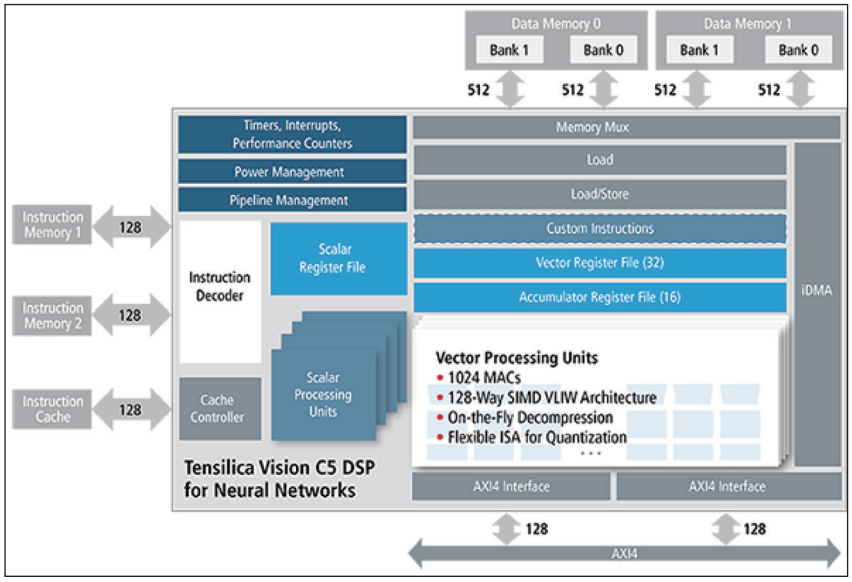
Roddy 觀察到 CNN 演算法有三大發展趨勢：1. 近來不到四年的時間，運算需求狂增十六倍；2. 網路架構趨於規律化，層次分明——例如，AlexNet 適用於規模較大的卷積運算、ResNet 適用於規模較小者，以及線性 (Linear) 或分支 (branch) 運算；3. 新應用層出不窮，遍及汽車、伺服器、家庭語音助理、手機及監控等，並強調：「非卷積演算」因鏈結關係相對簡單，運算次序無傷大雅，硬體加速器尚可應付；但若是具有綿密而複雜的對應關係、須步步為營的 CNN，邏輯一旦錯位，這些次序不明確且無法判讀意義的資料會讓網路混淆。

兼顧 Always-On 低功耗與 Heavy-Duty 高運算需求

他進一步表示，「除了低功耗和高速運算，嵌入式 always-on 系統的神經網路處理器還需具備靈活性和因應未來需求的能力；而 Vision C5 藉由消除神經網路 DSP 與主要視覺 / 影像 DSP 之間的外來資料移動，提供較 NN 加速器、GPU 和 CPU 更低功耗的解決方案及簡單的 NN 編程模型」。至於近年坊間出現的「視覺處理器」(VPU) 變種產品，Tensilica 的看法是：VPU 須用更多硬體才能實現同等的效能，將會導致整體功耗變高，亦非理想方案。簡言之，神經網路獨立運算可降低 DSP 負載，同時免去與主要 DSP 頻繁往返的疲於奔命。

於是，負責操持大局的視覺 / 圖像 DSP 便能騰出更多資源、專注執行影像應用程式，將所有神經網路運算交由同樣擁有完整戰力的 Vision C5 DSP 全權處置，分工合作。Tensilica 透露他們另一個實現低功耗的秘訣是：應對「重量級任務」(Heavy-Duty) 是以很大的運算單元一次搞定，之後便完全停工；經實驗結果證明，這將會比用大量平行運算、以低時脈分批運行的方式省電。「更重要的是，Vision C5 DSP 具備可編程和可擴展特性，易於整合。此時此刻所選定的介接硬體平台，將決定兩、三年後的產品銷售格局！」Roddy 重申。

圖 1：Tensilica Vision C5 DSP 框圖



資料來源：Cadence 官網

表：可在嵌入式系統中執行神經網路的方案比較

	CPU	GPU	NN 硬體加速器	視覺 / 影像 DS	★ Vision C5 DSP
開發容易程度	■純軟體 ■ IP 易獲取	■純軟體 ■ IP 易獲	硬體在試產瞬間已定，CPU/GPU/DSP 軟體必須在不同的可編程與加速器之間切割	■純軟體 ■ IP 易獲取	■純軟體 ■ IP 易獲取
功耗效率	最差	較 CPU 佳，但仍偏差	個別層級最佳，但全部加總後則不然	效率是 GPU 的 5~10 倍	較 DSP+NN 硬體加速器組合更佳
未來發展	可重新編程	可重新編程	難以重新編程，高風險	可重新編程	可重新編程
單一核心最大 NN 效能 (/sec)	<< 200 GFLOP (每秒浮點運算次數)	< 200 GFLOP (每秒浮點運算次數)	最高至 1 TMAC	200~250 GMAC	最高至 1 TMAC，但可擴充

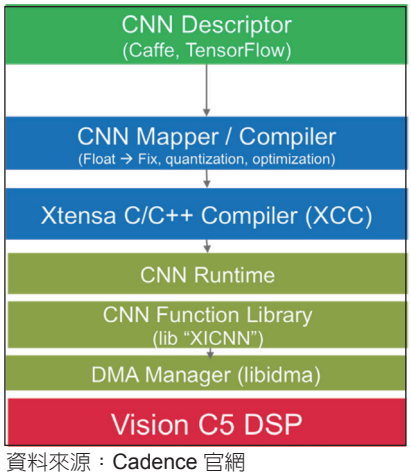
資料來源：Cadence (Tensilica)；筆者整理

他並提到，神經網路的工作量會因終端市場大相逕庭，例如，手機每秒的吞吐量多在200 GMAC (Giga Mac)以下，但保全監控和汽車半自動駕駛由於4K高清畫質的帶動、以便易於辨識，就上看1 TMAC (Tera MAC)左右，若是全自動駕駛的無人車，則至少10 TMAC起跳！「因此，效能指標不是越高越好，擴充的靈活度更應列入優先考慮，只有一種規格是無法跟上市場變化腳步的」，Roddy解釋。這多少也揭示為何Tensilica首發產品，是選擇從每秒1 TMAC的運算能力著手(以16nm製程、在不到1 mm^2的晶片面積實現)，或許正是抓取中間值而來。

Cadence對映器工具組加持，編程及擴展皆唾手可及

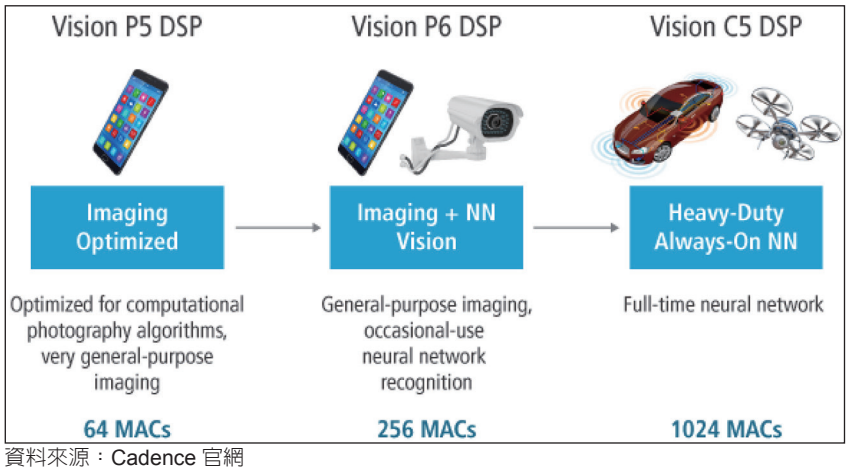
根據Tensilica發佈的資料顯示，Vision C5 DSP支援1024個8位元MAC或512個16位元MAC，兩種位元解析度均能實現優異效能，與GPU相比並不遜色；要比知名AlexNet CNN效能基準快六倍、更是Inception V3 CNN效能基準的九倍！若仍力有未逮，其「平台式」的多處理器設計支援可變核心大小、深度和輸入尺寸，亦能提供數個TMAC的高效能。它還包含多種係數壓縮／解壓技術，可隨時加入最新開發的層體，為日後所需預留空間；反觀硬體加速器因重新編程的能力有限，將來若想「平滑過渡」，最壞的局面恐須全部從頭來過！

圖2：Cadence神經網路對映器工具組(Mapper Toolset)提供標準的開源CNN框架，將資訊流導入CNN映射器直抵Vision C5 DSP



資料來源：Cadence官網

圖3：Cadence Tensilica Vision系列DSP所側重的應用市場各有不同



資料來源：Cadence官網

在指令集方面，Vision C5 DSP擁有128路8位元SIMD或64路16位元SIMD的VLIW SIMD架構；另整合128位元的iDMA及AXI4記憶體介面；其附帶的Cadence神經網路對映器工具組可運用神經網路程式庫功能，將所有Caffe和TensorFlow等主流框架生成的神經網路，對映成可執行且高度優化的Vision C5 DSP編碼。Roddy及隨機受訪的愛用者皆不諱言，這正是Tensilica與Cadence合併的最大優勢：縮短學習曲線並簡化認證作業，讓程式碼更容易移植、編程更容易上手。

特別是新推出的Vision C5 DSP與Tensilica自身的Vision P5/P6 DSP採用相同的實用軟體工具組，更有助於產品及時上市。在DSP授權市場連年奪冠的Tensilica，授權用戶約250家；全球前二十大半導體廠、就有多達十七家皆是其用戶，每年全球IC出貨總量約40億顆，在音訊DSP市場更是堪稱獨領風騷。事實上，除了視覺的高清辨識需要，聽覺的娛樂饗宴亦從未缺席；搭配人工智慧(AI)演算法的語音控制系統以及具有指向性的3D立體聲，有朝一日甚至能有更多功能性的應用。

隨著影音浪潮的無孔不入，借助神經網路架構的機會預料將有增無減；在特定應用DSP IP佈局甚深、掌握專業知識的Tensilica有了Cadence集團資源加持，對於開發高效率的指令集將可獲得最實質的挹注，繼續發光發熱。CTA