在雲端、網路和邊緣部署高效的人工智慧深度學習推論

■作者：Daniel Eaton
賽靈思策略行銷發展資深經理

深度學習人工智慧 (AI) 應用是新時代生產力的關鍵，機器學習能提高與增強人類的創造力。人們致力於將數兆位元 (terabyte) 的訓練資料和海量的數學運算用於全面訓練每條神經網路。訓練可在離線情況下，以數天的時間進行大規模批次處理，然而經過訓練的網路要投入部署，就得符合嚴格的時限要求。

資料中心為了納入 AI 處理功能，需部署更新的伺服器，但有限的空間與功耗問題是阻礙 AI 部署的關鍵考量；同時，客戶又期望獲得即時回應，因此極低的延遲極為重要。

若以汽車駕駛輔助系統或自動駕駛車為例，當攸關生命安全時，最大限度地降低延遲與可靠的即時回應至關重要。此時，規模與功耗要求將比資料中心更為嚴格，且重量和散熱問題也是必須要考慮的問題。特斯拉 (Tesla) 近期在自動駕駛日 (Autonomy Day) 上分享為何他們選擇自行開發晶片，因為相對於 GPU 而言，其自有晶片能夠結合低

功耗 (小於 100 瓦) 和低延遲 (一個批次處理大小) 的優勢。

隨著 AI 持續被使用在越來越多的領域以更快解決高挑戰性問題的情況下，對於已部署神經網路的效能要求也隨之提高。

歷久彌新的高效推論

無論我們討論的是雲端 AI 或是 AI 在汽車等領域的嵌入式應用，這些驅動新興應用需求的推論引擎都必須確保低延遲、低功耗和小尺寸。

為了成功將經過訓練的神經網路實際應用於推論，不僅需要適當的剪枝 (pruning) 與最佳化，還要周延地考慮處理平台，以利在功耗、尺寸和散熱等常見的限制條件下達到所需的效能 (通常指反應時間或延遲)。隨著 AI 商業部署的成長及終端使用者的需求提升，處理器晶片廠商都在加強高階元件架構的開發以滿足相關需求。

部分最新型的晶片鎖定自動駕駛等應用領域，採用結合 CPU 和應用處理器核心的混合架構，並搭配大量 GPU 進行數學運算。儘管能運用晶片上的資源，但固定的架構使開發者受限於不靈活的記憶體介面寬度和資料解析度。8 位元整數通常是最小寬度，但深度學習演算法能以更低的解析度有效操作資料，有時甚至低至 2 位元或 1 位元。不靈活的 CPU 或 GPU 運算架構越來越難以滿足神經網路的效能需求，人們需要更靈活的架構來調適解析度及核心數量，進而達到最佳運算效能與功耗。

將受訓的神經網路進行剪枝與最佳化，並在目標處理器中達到高效實施已經非常困難，新型且更高效的神經網路開發速度卻還超越晶片的發展速度。因此即使在最初開發時採用最新技術的項目，在要部署時就已經過時了。不過，當部署最先進的神經網路技術時，幾乎無法在現有處理器架構中發揮全力。

可配置的 AI 加速

為解決這些效能、功耗和未來靈活應變能力的挑戰，開發團隊正運用 FPGA 所提供的靈活性優勢來建構 AI 加速器。

FPGA 能與成千上百甚至成千上萬個高度平行化的運算單元進行配置，最低能支援單位解析度，且還能為了消除瓶頸客製化記憶體介面。此外，FPGA 能夠輕鬆地進行再編程，讓開發者在不同世代的晶片之間更靈活地升級神經網路結構，以跟上技術發展的步伐。

賽靈思在 2017 年收購 AI 專業公司深鑒科技 (DeePhi Tech) 後，已強化先進的剪枝以及最佳化工具與 IP 的開發力度，因此在 FPGA 上能更佳地部署神經網路。剪枝技術能透過刪除不具備影響力的且近零訓練的權重，並可在最大限度減少運算操作數量及降低功耗的地方重組網路，進而精簡神經網路。深鑒科技的神經網路剪枝技術已被最佳化以運行在 FPGA 上，能減少高達九成的權重，同時達到可接受的圖形辨識精度，讓效能大幅提升 10 倍且能源效率也明顯提升。

從雲端到邊緣

自動駕駛對於說明對低運算延遲、小尺寸、低重量和低功耗的迫切需求來說是個簡單明瞭的例子。雷達或攝影機所偵測到的物體 (如車輛、自行車騎士或行人等) 需在偵測到的瞬間進行辨識。眾所皆知，人類在四分之一秒的時間內就能對視覺輸入做出反應，因此自動駕駛系統所需的視覺辨識功能也必須達到相同的速度，甚至更快。為符合人類駕駛的水準，系統在整個偵測、辨識與回應的過程中，應當在 1.5 秒之內做出緊急剎車的決定。

賽靈思近期宣佈與賓士開展研發合作，透過採用高效能 FPGA 的深度學習處理器來分析從攝影機、雷達和光達獲得的資料，以達到駕駛監控、車輛導引及防撞等功能。兩家公司的專家都在高度自行調適的汽車平台上運行 AI 演算法，並將為賓士的神經網路優化深度學習處理器技術。該技術在維持高能源效率的同時還能達到極低的延遲，確保系統在汽車有限的散熱條件下能夠可靠的運作。

此外，在資料中心領域，FPGA 也能支援深度學習加速器，其效能功耗比的水準大大超越典型的 GPU 處理器。SK Telecom (SKT) 採用 Kintex UltraScale FPGA 作為其資料中心的 AI 加速器，成功提升其語音助理 NUGU 的效能；此為韓國電信產業首次部署 AI。與傳統採用 GPU 的處理器相比，SKT 的自動語音辨識 (ASR) 應用速度提升高達 500%，且效能功耗比也高出 16 倍。此外，在既有僅使用 CPU 的伺服器上採用加速器來處理多個語音通道 (voice channel) 時，能顯著降低 SKT 的整體擁有成本 (TCO)。

另一個例子就是最新型的 AI 家庭安全功能。賽靈思與該新興市場其中之一的領導廠商 Tend Insights 共同開發雲端平台，FPGA 加速的低延遲推論作為該雲端平台的一部分，能夠支援更加智慧的監控功能和居家護理等創新服務。安裝在家裡各個位置的攝影機能夠辨識屋主可能有興趣的畫面，並將其上傳至運行在雲端並採用 FPGA 的 AI 加速器。這些加速器可通過一系列 API 存取 (本例中是作為賽靈思機器學習套件 (ML Suite) 硬體編譯器的一部分)，將家庭成員與寵物從陌生人與其他動物間區分出來，進而偵測威脅並發出警報。在主人的同意下，內部攝影機的畫面能進行分析，以判斷家庭成員是否面臨困難，例如像老人摔倒需要幫助時，系統能夠撥打電話給指定的家庭成員或專業護理人員以尋求幫助。

AI 在其他眾多場合也被用於進行複雜的模式匹配和影像辨識，以快速獲得結果，其中包含基因體分析來加速醫療診斷與治療。使用 FPGA 加速 AI 推論以將定序病人基因體和辨識異常因果所需的時間，從 24 小時縮短至大約 30 分鐘，且加速的速度也在推進之中。再來是原子研究，捕捉核融合的實驗畫面涉及極高清影像，每幀動不動就達到 1 億像素，且需在 25 毫秒 (ms) 之內處理完成。這類挑戰需要使用低延遲神經網路探測器，但傳統的 CPU 推論根本無法達成，甚至相差甚遠。然而，FPGA 已經在協助科學家獲得他們所需的答案了。

結論

AI 近期才成為確實可行的解決方案，但已在人們的日常生活中成為驅動各種服務的重要樞紐。降低營運成本、縮短客戶等待時間及創造全新增值服務的機會等前景，讓商業組織們感到興奮，但同時也加大提升效能的壓力，增加降低延遲、功耗和成本的需求。

創建 AI 有兩大要素，一是訓練最符合現有任務需求的神經網路類型；二是對經過訓練的網路進行剪枝與最佳化，成為在適當的處理器上運行的推論引擎。

FPGA 架構的靈活性和高效能、結合高效能最佳化工具來建置與最佳化 (如 ML 套件硬體編譯器和 Deephi 最佳化器)，及不用等待晶片重製即可建置最新神經網路架構所帶來的可重組能力，這三者成為在雲端、網路或網路邊緣落實加速 AI 推論的關鍵因素。CTA