AI 搶灘陣式②：訓練、推論還不夠，自主學習是下一步

仿效人腦決策！Intel Nervana NNP 開先河

■文：任芷萍

如果說，物聯網是科技業界的「下一件大事」，那麼，人工智慧 (AI) 就是下一個產業巨浪——英特爾 (Intel) 如是說。不甘讓 GPU 獨領風騷，Intel 於 2016 年收購在 AI 圈小有名氣的 Nervana Systems 公司，順勢延攬該公司的聯合創辦人暨首席技術長 Amir Khosrowshahi 加入 Intel 人工智慧產品事業群 (AIPG)、擔任副總裁暨技術長一職。Khosrowshahi 在神經網路 (Neural Networks)、機器學習 (Machine Learning) 和深度學習 (Deep Learning) 的專業知能廣受業界認同。

照片人物：Intel 人工智慧產品事業群 (AIPG) 副總裁暨技術長 Amir Khosrowshahi

雲端服務大躍進，運算週期耗時長

Khosrowshahi 揭示：AI 正逐漸融入我們日常生活，在消費、保健、財務、零售、政府、能源、交通、產業……，全面激起千層浪；可預見的是，今後，「資料洪流即將爆發」！預估到 2020 年，每位網路使用者、平均每天將產生 1.5 GB 的資料流量；每家智慧醫院將超過 3,000 GB；每輛自駕車上看 4,000 GB；而智能工廠更驚人，將超過 1,000,000 GB！為應對漫溢的數據並尋求運算突破及創新，屆時，從資料中心的大型主機、邊緣 (Edge) 基礎伺服器到雲端平台的 AI 運算週期 (Compute Cycles，應用程式之處理、執行的時間總和) 將遽增 12 倍！

Khosrowshahi 表示，目前 AI 運用還停留在描述、診斷型的初級操作，但用於預測、指示及認知的進階分析正在興起。深度學習 (Deep Learning) 只是核心起點，向外擴展至神經網路和機器學習，才是最終應用取向。以人臉辨識為例，典型的機器學習只是以臉部 T 字部位為基準、定出若干重點函數，然後透過支援向量機 (Support Vector Machine, SVM)、隨機森林 (Random Forest)、原始貝氏機率

表：「Nervana」平台之產品組合

產品	說明
Xeon	可擴充之運算處理器。方便為 AI 日後更高的工作負載預留升級空間，並針對最密集的深度學習訓練 (training) 工作，推出名為「Lake Crest」的專屬晶片。
Mobileye	專為自駕車設計的視覺技術。
FPGA	用以執行深度學習推論 (inference) 的可編程加速器。
Movidius	低功耗視覺技術，讓機器學習得以在多樣化的終端裝置執行。

資料來源：Intel 提供，編輯部整理

(Naive Bayes) 演算，建立決策樹 (Decision Trees) 模型、進行邏輯迴歸 (Logistic Regression) 分析，最終加以組合。

然而，結合神經網路的「深度學習」可沒這麼簡單！它須建置好幾個運算層，至少包括 6,000 萬個參數，以擷取資料找出特徵、在抽象層萃取特性；期借助更多資料改善效能、提高表徵再現能力 (Representational Power)。放眼未來 AI 與其他關鍵商務作業負載並行運作需求，Intel 將多年研發與併購成果統整成為「Nervana」平台，企圖通吃 AI 從資料中心主機、邊緣 (edge) 裝置到雲端服務的運算大餅。Khosrowshahi 強調，「Nervana」平台擁有完整的軟體堆疊、友善的使用者介面及整套系統 (Turnkey System)，可縮短開發週期並做個別加速。

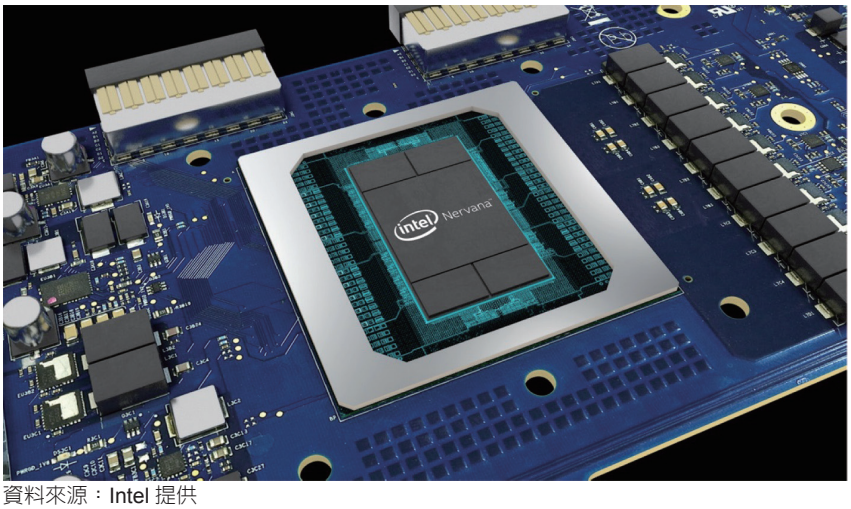
揮別 Cache！Nervana NNP 用軟體管理片上記憶

Intel 與 NASA Frontier Development Lab 合作，運用 Nervana 深度學習技術協助描繪月球地圖，以及包括太空天氣、太空資源和行星防禦等多項太空任務研究；而新近發佈的 Nervana 類神經網路處理器 (Neural Network Processor, NNP)，則是業界首款專為類神經網路 (Neural Network) 所設計的晶片，將於今年底前出貨。NNP 的設計靈感源自人腦，訓練電腦根據模式 (pattern) 與關聯性 (association) 做決策。Intel 已著手研發不同世代的 NNP 產品，期於 2020 年達成將 AI 效能提高百倍的目標。

Nervana NNP 之所以令人矚目，是因為它繞過標準快取記憶體 (Cache)，改用軟體管理片上記憶體，以極大化每個矽晶裸片的運算利用率！單一晶片上的神經網路運算主要受限於電源和記憶體頻寬，Nervana NNP 特有高速晶片互連設計，可提高吞吐量、大量雙向傳輸數據，讓多個晶片上的神經網路參數模型使出渾身解數做平行運算。此外，Intel 還獨家新創名為「Flexpoint」的數字格式，允許將標量計算 (scalar computations) 導入定點加乘，擴大共享指數的動態範圍；受惠於電路縮小，不僅大幅提高晶片的平行能力，也降低每次運算的必要功率。

與此同時，Intel 還展開自主學習 (self-learning) 的類神經型態測試晶片，不必經過「訓練」就能執行複雜的感知作業，例如：解讀心律、偵測異常網路。台灣分公司企業解決方案事業群協理鄭智成補充，「Nervana」平台的意義在於將 Intel 所有 AI 產品集結整合，包括先前推出的「Crest」系列特定應用積體電路 (ASIC) 晶片——可直接內建在伺服器的電路板或做成插槽板卡，但當下只供高階伺服器產品使用，預計明年才會普及至所有伺服器產品。他指出，雲端服務供應商可藉 ASIC 優勢加速 AI 運

圖 1：Intel Nervana 類神經網路處理器 (NNP)



資料來源：Intel 提供

照片人物：Intel 台灣分公司企業解決方案事業群協理鄭智成

作速度，神乎其技地縮減 AI 訓練或推論時間。

AI 最強研發能量，在學界！

鄭智成聲明，對比歷史悠久、已有多種應用、同樣可作為加速器的 FPGA，多以開發板形式供貨，兩者角色並不衝突。他另提到，AI 其實早在五、六十年前就已出現，近來的百花齊放若用「AI 復興」形容也不為過，幕後最大動力就是「類神經網路」的長足進展；以人臉、物件辨識的正確率為例，2012 年之前，要超過八成極其困難，但現已普遍來到 95% 以上的水準。遺憾的是，上一波 AI 熱潮並沒有太多實質產出，以致企業對相關投資興趣缺缺；倒是在學術界持續努力下，迄今已蘊藏可觀的研發人才及能量。

鄭智成透露，中國大陸曾對全球大學在 AI 的發展狀況做統計，發現全亞洲只有台灣、日本及新加坡三地的大學有參與 AI 盛宴。「台灣有很多優秀師生團隊在海外 AI 競賽皆奪冠，顯示台灣在 AI 有很好的實力；但若不能善用硬體資源、將系統最佳化，恐虛擲許多光陰在模型訓練及修改上。有時單是訓練一個模型就得十多天，調整參數又須耗上好幾十天」，他感嘆說。有鑑於此，Intel 特別為大學院校師生制訂一套方案，在日前 Intel AI Day 活動中，宣佈將在台灣引進專為學術研究而設立的「Intel Nervana AI 學院計畫」。

Intel 與 Coursera 教育機構合作開設 AI 線上課程，包括訓練、實作演練工作坊；除了獨家提供可遠端存取的工具及資源，並敦請專家指導。Intel 先前與子宮頸癌研究單位 MobileODT 及資料分析平台 Kaggle 合作，舉辦子宮頸癌篩檢研究競賽專案，亦是結合 Nervana AI 學院資源的具體展現。為強化支援開放 AI 產業體系，Intel 特推出一系列開發工具以增進易用性與跨平台相容性，支援多種開源深度學習框架。Intel Nervana DevCloud 計畫可讓 AI 開發者存取雲端資源。

縱向＋橫向擴展，讓運算發揮最大效益

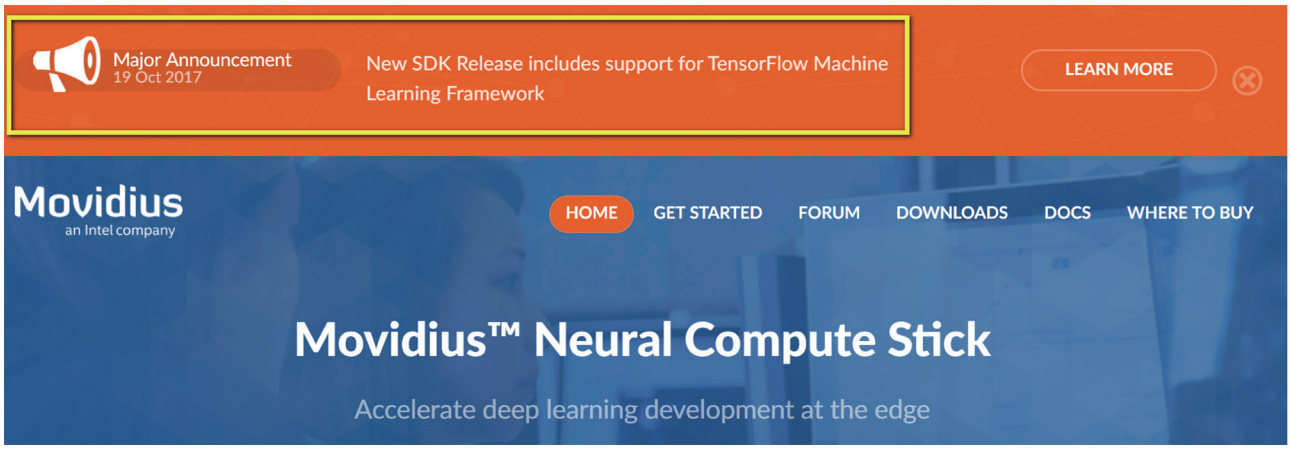
鄭智成以 GitHub 上最活躍的兩大主流框架——Caffe 和 Google TensorFlow 為例，由 Intel 發行、維護的 Caffe 版本的價值在於：透過系統優化手段，實現單一節點的縱向擴展 (Scale-up) 及多節點的橫向擴展 (Scale- out)，可提升數十到數百倍效能。例如，以 128 核伺服器完勝雙核、四核的 Notebook / PC，或善用電腦叢集 (Cluster) 技術擴增節點數，讓原需費時三天的運算工作壓縮至一小時內完成；「就算整個資料中心只有一台伺服器，也能運用大型叢集加速工作時程」。至於 Google TensorFlow，由於其授權模式類似 Android，故由 Google 全權維護。

即使如此，Intel 自 TensorFlow 1.1 版後仍積極參與

圖 2：Intel 支援多種 AI 開源深度學習框架

資料來源：Intel 提供

圖 3：Movidius 神經運算棒 (NCS) 已於今年 10 月發佈支援 TensorFlow 框架的 SDK，如黃色框線所示



資料來源：https://developer.movidius.com/

優化 (最新版本為 1.3)，並將程式碼貢獻出來，讓開發者能最大限度利用硬體資源。另在軟體開發套件 (SDK) 方面也不馬虎，例如，內建 Movidius Myriad 2 VPU 的神經運算棒 (NCS) 有兩個版本：一是搭載訓練完成的 Caffe 演算模型 (售價 79 美元)，一是新近發佈支援 TensorFlow 的版本 (售價 99 美元)。那麼，不同應用對於軟、硬體架構是否有不同要求？鄭智成的回答是，AI 實際應用層面還是源於使用者創意，主要區別在軟體演算法，硬體層級並無差異，而用於訓練或推論也僅差在資料量大小。

例如，用訓練晶片做推論、一個大晶片只做單一通道的辨識，似乎不太有效率。因此，Intel 現階段意在為開發者提供通用框架 (general- purpose framework)、讓設備／服務供應商得以基於自身需求最佳化，並未針對特定 AI 應用再細分不同架構。不過他亦沒把話說死，表示日後若有市場需要，Intel 也不排除為影像、語音等分眾應用推出專用晶片。鄭智成認為，以往軟體演算法需時數天的運算，經由 AI 硬體加速器可能只要數小時就搞定，預料 FPGA、終端裝置或閘道器 (Gateway) 將因而蓬勃發展，資料中心並非 AI 唯一市場。

例如，微軟就採用 Intel Stratix 10 FPGA 作為其深度學習加速平台 (Project Brainwave) 的硬體加速器，強化雲端環境的「AI 即時運算效能」，因應智慧手機人臉、語音識別或自駕車所需。另成本高昂的無人機，可運用 AI 物件辨識避免與樹木或障礙物碰撞、墜毀，甚至會認主人，只要比個手勢就能拍照或讓它自動降落在手上，都是有趣的 AI 終端應用。 CTA