圖形處理單元與人工智慧

賴昭正
前清大化學系教授、系主任、所長；合創科學月刊

> 我擔心人工智慧可能會完全取代人類。如果人們能設計電腦病毒，那麼就會有人設計出能夠自我改進和複製的人工智慧。這將是一種超越人類的新生命形式。
>
> -史蒂芬・霍金(Stephen Hawking)
> 英國理論物理學家

大約在八十年前,當第一台數位計算機出現時，一些電腦科學家便一直致力於讓機器具有像人類一樣的智慧；但七十年後，還是沒有機器能夠可靠地提供人類程度的語言或影像辨識功能。誰又想到「人工智慧」(Artificial Intelligent,簡稱 AI)的能力最近十年突然起飛，在許多(所有?)領域的測試中擊敗了人類,正在改變各個領域—包括假新聞的製造與散佈—的生態。

圖形處理單元（graphic process unit,簡稱 GPU）是這場「人工智慧」革命中的最大助手。它的興起使得九年前還是個小公司的 Nvidia（英偉達）股票從每股不到$5，上升到今天（5 月 24 日）每股超過$1000（註一）的全世界第三大公司，其創辦人（之一）兼首席執行官、出生於台南的黃仁勳（Jenson Huang）也一躍成為全世界排名 20 內的大富豪、台灣家喻戶曉的名人！可是多少人了解圖形處理單元是什麼嗎? 到底是時勢造英雄，還是英雄造時勢?

在回答這問題之前，筆者得先聲明筆者不是學電腦的，因此在這裡所能談的只是與電腦設計細節無關的基本原理。筆者認為將原理轉成實用工具是專家的事，不是我們外行人需要了解的；但作為一位現在的知識分子或公民，了解基本原理則是必備的條件：例如了解「能量不滅定律」就可以不用仔細分析,即可判斷永動機是騙人的；又如現在可攜帶型冷氣機充斥市面上，它們不用往室外排廢熱氣，就可以提供屋內冷氣，可能嗎?

CPU 與 GPU

不管是大型電腦或個人電腦都需具有「中央處理單元」(central process unit, 簡稱 CPU)。CPU 是電腦的「腦」,其電子電路負責處理所有軟體正確運作所需的所有任務，如算術、邏輯、控制、輸入和輸出操作等等。雖然早期的設計即可以讓一個指令同時做兩、三件不同的工作；但為了簡單化，我們在這裡所談的工作將只是執行算術和邏輯運算的工作(arithmetic and logic unit, 簡稱 ALU)，如將兩個數加在一起。在這一簡化的定義下，CPU 在任何一個時刻均只能執行一件工作而已。

在個人電腦剛出現只能用於一般事物的處理時,CPU 均能非常勝任地完成任務。但電腦圖形和動畫的出現帶來了第一批運算密集型工作負載後，CPU 開始顯示心有餘而力不足：例如電玩動畫需要應用程式處理數以萬計的像素(pixel)，每個像素都有自己的顏色、

光強度、和運動等，使得 CPU 根本沒辦法在短時間內完成這些工作。於是出現了主機板上之「顯示插卡」來支援補助 CPU。

1999 年，英偉達將其一「具有集成變換、照明、三角形設定/裁剪、和透過應用程式從模型產生二維或三維影像的單晶片處理器」（註二）定位為「世界上第一款 GPU」，「GPU」這一名詞於焉誕生。不像 CPU，GPU 可以在同一個時刻執行許多算術和邏輯運算的工作，快速地完成圖形和動畫的變化。

依序計算和平行計算

一部電腦 CPU 如何計算 $7 \times 5 + 6/3$ 呢？因每一時刻只能做一件事，所以其步驟為:

(1) 計算 7×5;
(2) 計算 6/3;
(3) 將結果相加。

總共需要 3 個運算時間。但如果我們有兩個 CPU 呢？很多工作便可以同時（平行）進行:

(1)同時計算 7×5 及 6/3;
(2)將結果相加。

只需要 2 個運算時間,比單獨的 CPU 減少了一個。這看起來好像沒節省多少時間，但如果我們有 16 對$a \times b$要相加呢? 單獨的 CPU 需要 31 個運算的時間(16 個×的運算時間及 15 個+的運算時間)，而有 16 個小 CPU 的 GPU 則只需要 5 個運算的時間(1 個×的運算時間及 4 個+的運算時間)!

現在就讓我們來看看為什麼稱 GPU 為「圖形」處理單元。圖一左圖《我愛科學》一書擺斜了,如何將它擺正成右圖呢？一句話：「將整個圖逆時針方向旋轉 θ 即可」。但因為左圖是由上百萬個像素點(座標 x,y)組成的,所以這句簡單的話可讓 CPU 忙得不亦樂乎了: 每一點的座標都必須做如下的轉換

圖一：圖形處理的例子

$$x' = x\,cos\theta + y\,sin\theta$$

$$y' = -x\,sin\theta + y\,cos\theta$$

即每一點均需要做四個×及兩個+的運算! 如果每一運算需要10^{-6}秒，那麼讓《我愛科學》一書做個簡單的角度旋轉便需要 6 秒，這豈是電動玩具畫面變化所能接受的?

人類的許多發明都是基於需要的關係，因此電腦硬件設計家便開始思考:這些點轉換都是獨立的,為什麼我們不讓它們同時進行(平行運算,parallel processing)呢？於是

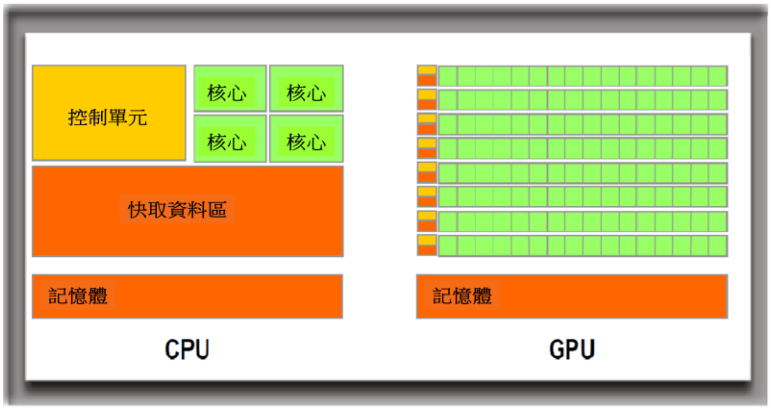
專門用來處理「圖形」的處理單元出現了—就是我們現在所知的 GPU。如果一個 GPU 可以同時處理10^6運算，那上圖的轉換只需10^{-6}秒鐘！

GPU 的興起

GPU 可分成兩種：［1］整合式圖形「卡」（integrated graphics）是內建於 CPU 中的 GPU，所以不是插卡，它與 CPU 共享系統記憶體，沒有單獨的記憶體組來儲存圖形/視訊，主要用於大部分的個人電腦及筆記型電腦上；早期英特爾（Intel）因為不讓插卡 GPU 侵蝕主機的地盤，在這方面的研發佔領先的地位，約佔 68%的市場。［2］獨立顯示卡（discrete graphics）有不與 CPU 共享的自己專用內存；由於與處理器晶片分離，它會消耗更多電量並產生大量熱量；然而，也正是因為有自己的記憶體來源和電源，它可以比整合式顯示卡提供更高的效能。

2007 年，英偉達發布了可以在獨立 GPU 上進行平行處理的軟體層後，科學家發現獨立 GPU 不但能夠快速處理圖形變化，在需要大量計算才能實現特定結果的任務上也非常有效，因此開啟了為計算密集型的實用題目編寫 GPU 程式的領域。如今獨立 GPU 的應用範圍已遠遠超出當初圖形處理，不但擴大到醫學影像和地震成像等之複雜圖像和影片編輯及視覺化，也應用於駕駛、導航、天氣預報、大資料庫分析、機器學習、人工智慧、加密貨幣挖礦、及分子動力學模擬（註三）等其它領域。獨立 GPU 已成為人工智慧生態系統中不可或缺的一部分，正在改變我們的生活方式及許多行業的遊戲規則。英特爾在這方面發展較遲，遠遠落在英偉達（80%）及超微半導體公司（Advance Micro Devices Inc.，19%，註四）之後，大約只有 1%的市場。

事實上現在的中央處理單元也不再是真正的「單元」，而是如圖二可含有多個可以同時處理運算的核心(core)單元。GPU 犧牲大量快取和控制單元以獲得更多的處理核心，因此其核心功能不如 CPU 核心強大，但它們能同時高速執行大量相同的指令，在平行運算中發揮強大作用。現在電腦通常具有 2 到 64 個核心；GPU 則具有上千、甚至上萬的核心。



圖二：典型的CPU與GPU架構

結論

我們一看到《我愛科學》這本書，不需要一點一點地從左上到右下慢慢掃描，即可瞬間知道它上面有書名、出版社等，也知道它擺斜了。這種「平行運作」的能力不僅限於視覺，它也延伸到其它感官和認知功能。例如筆者在清華大學授課時常犯的一個毛病是：嘴巴在講，腦筋思考已經不知往前跑了多少公里，常常為了追趕而越講越快，將不少學生

拋到腦後！這不表示筆者聰明，因為研究人員發現我們的大腦具有同時處理和解釋大量感官輸入的能力。

人工智慧是一種讓電腦或機器能夠模擬人類智慧和解決問題能力的科技，因此必須如人腦一樣能同時並行地處理許多資料。學過矩陣（matrix）的讀者應該知道,如果用矩陣和向量(vector)表達,上面所談到之座標轉換將是非常簡潔的（註五）。而矩陣和向量計算正是機器學習（machine learning）演算法的基礎!也正是獨立圖形處理單元最強大的功能所在!因此我們可以了解為什麼 GPU 會成為人工智慧開發的基石：它們的架構就是充分利用並行處理，來快速執行多個操作，進行訓練電腦或機器以人腦之思考與學習的方式處理資料━稱為「深度學習」（deep learning）。

黃仁勳在 5 月 22 日的發布業績新聞上謂：『下一次工業革命已經開始了：企業界和各國正與英偉達合作，將價值數萬億美元的傳統資料中心轉變為加速運算及新型資料中心━人工智慧工廠━以生產新商品「人工智慧」。人工智慧將為每個產業帶來顯著的生產力提升，幫助企業降低成本和提高能源效率，同時擴大收入機會。』

附錄

人工智慧的實用例子：下面一段是微軟的「copilot」代書、谷歌的「translate」代譯之「one paragraph summary of GPU and AI」。讀完後，讀者是不是認為筆者該退休了?

GPU（圖形處理單元）和 AI（人工智慧）之間的協同作用徹底改變了高效能運算領域。GPU 具有平行處理能力，特別適合人工智慧和機器學習所需的複雜資料密集運算。這導致了影像和視訊處理等領域的重大進步，使自動駕駛和臉部辨識等技術變得更加高效和可靠。NVIDIA 開發的平行運算平台 CUDA 進一步提高了 GPU 的效率，使開發人員能夠透過將人工智慧問題分解為更小的、可管理的、可同時處理的任務來解決這些問題。這不僅加快了人工智慧研究的步伐，而且使其更具成本效益，因為 GPU 可以在很短的時間內執行與多個 CPU 相同的任務。隨著人工智慧的不斷發展，GPU 的角色可能會變得更加不可或缺，推動各產業的創新和新的可能性。大腦透過神經元網路實現這一目標，這些神經元網路可以獨立但有凝聚力地工作，使我們能夠執行複雜的任務，例如駕駛、導航、觀察交通信號、聽音樂並同時規劃我們的路線。此外，研究表明，與非人類動物相比，人類大腦具有更多平行通路，這表明我們的神經處理具有更高的複雜性。這個複雜的系統證明了我們認知功能的卓越適應性和效率。我們可以一邊和朋友聊天一邊走在街上，一邊聽音樂一邊做飯，或一邊聽講座一邊做筆記。人工智慧是模擬人類腦神經網路的科技，因此必須能同時並行地來處理許多資料。研究人員發現了人腦通訊網路具有一個在獼猴或小鼠中未觀察獨特特徵：透過多個並行路徑傳輸訊息,因此具有令人難以置信的多任務處理能力。

******* 註 *******

（註一）當讀者看到此篇文章時，其股票已一股換十股，現在每一股約在$100 左右。

（註二）組裝或升級過個人電腦的讀者或許還記得「英偉達精視 256」(GeForce 256) 插卡吧?

（註三）筆者於 1984 年離開清華大學到 IBM 時,就是參加了被認為全世界使用電腦時間最多的量子化學家、IBM「院士(fellow)」Enrico Clementi 的團隊：因為當時英偉達還未有可以在 GPU 上進行平行處理的軟體層，我們只能自己寫軟體將 8 台中型電腦(非 IBM 品牌!)與一大型電腦連接來做平行運算，進行分子動力學模擬等的科學研究。如果晚生 30 年或許就不會那麼辛苦了?

（註四）補助個人電腦用的 GPU 品牌到 2000 年時只剩下兩大主導廠商：英偉達及 ATI(Array Technology Inc.)。後者是出生於香港之四位中國人於 1985 年在加拿大安大略省成立,2006 年被超微半導體公司收購，品牌於 2010 年被淘汰。超微半導體公司於 2014 年 10 月提升台南出生之蘇姿豐（Lisa Tzwu-Fang Su）博士為執行長後，股票從每股$4 左右，上升到今天每股超過$160，其市值已經是英特爾的兩倍，完全擺脫了在後者陰影下求生存的小眾玩家角色，正在挑戰英偉達的 GPU 市場。順便一題:超微半導體公司現任總裁(兼 AI 策略負責人)為出生於台北的彭明博(Victor Peng)；與黃仁勳及蘇姿豐一樣，也是小時候就隨父母親移居到美國。

（註五）$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} cos\theta & sin\theta \\ -sin\theta & cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$ 或 $x' = Mx$。

******** 延伸閱讀 ********

- 「熱力學與能源利用」，《科學月刊》，1982 年 3 月號；收集於《我愛科學》（華騰文化有限公司，2017 年 12 月出版），轉載於「嘉義市政府全球資訊網」。
- 「網路安全技術與比特幣」，《科學月刊》，2020 年 11 月號；轉載於「善科教育基金會」的《科技大補帖》專欄。