SAMSUNG

【專訪】先進技術助推雲端級運算智慧導入裝置端 AI

在經典科幻電影中，AI 的形象往往是巨大的電腦系統或大型伺服器；如今 AI 卻成為稀鬆平常的技術，一機在手即能取用。有鑑於此，三星電子正將裝置端 AI 擴及智慧型手機和家電等各種產品上，讓 AI 能在本機執行，不必仰賴外部伺服器或雲端，打造更快速、安全的體驗。

不同於伺服器型系統，裝置端環境必須在嚴苛的記憶體和運算限制下運作。因此，勢必得縮小 AI 模型尺寸並盡可能提升執行環境效率。面對上述挑戰，三星研究院 AI 中心現正引領核心技術的開發工作，包括模型壓縮、執行環境軟體最佳化，乃至於新式架構開發。

三星新聞中心專訪三星研究院 AI 中心技術專家 MyungJoo Ham 博士，探討裝置端 AI 的未來及其背後的最佳化技術。

▲ MyungJoo Ham 博士

邁向裝置端 AI 的第一步

生成式 AI 的核心是能解讀用戶語言並產生自然回覆的大型語言模型（LLMs）。而實現裝置端 AI 的第一步便是壓縮和優化這些大量且複雜的模型，以便在智慧型手機等裝置上順暢執行。

Ham 博士表示：「直接在智慧型手機或筆記型電腦上執行高階模型，約要進行數十億次運算，不僅快速耗電，致使裝置升溫、甚至延遲反應時間，導致用戶體驗受到顯著影響。模型壓縮技術之所以出現，就是為了解決這類問題。」

LLMs 使用極為複雜的數值表示來執行運算，而模型壓縮技術可透過量化程序，將這些數值簡化為更有效率的整數格式。Ham 博士說明：「正如壓縮高解析度相片，雖然檔案大小縮水了，但畫質幾乎維持不變。換言之，將 32 位元浮點計算轉換為 8 位元甚至 4 位元整數，可大幅縮減記憶體用量和運算負載，進而加速反應時間。」

▲ 模型壓縮技術透過量化模型權重以縮減大小、增加處理速度並維持效能。

量化期間，由於數字精準度下降，可能會使模型的整體精準度下滑。為了平衡速度和模型品質，三星研究院正在開發能在壓縮後嚴密測量與校正效能的演算法和工具。

Ham 博士指出：「模型壓縮的目的不僅是縮小模型尺寸，還要維持快速精準。我們使用最佳化演算法分析模型在壓縮過程中的損失函數並進行再次訓練，消除誤差較大的區域，直到輸出值接近原始模型。由於各模型權重的重要程度不一，因此我們將精確度提升，保留關鍵權重，同時更積極壓縮次要模型。此方法可在不犧牲精準度的情況下大幅增加效率。」

除了在原型階段開發模型壓縮技術，三星研究院亦因應智慧型手機和家電等實際產品，對模型進行調整和商業化。Ham 博士稱：「每個裝置模型皆有各自的記憶體架構和運算設定檔，因此單憑通用方法無法創造媲美雲端的 AI 效能。透過產品導向研究，三星自行設計的壓縮演算法能讓用戶切身感受到更強大的 AI 體驗。」

推動 AI 效能的幕後引擎

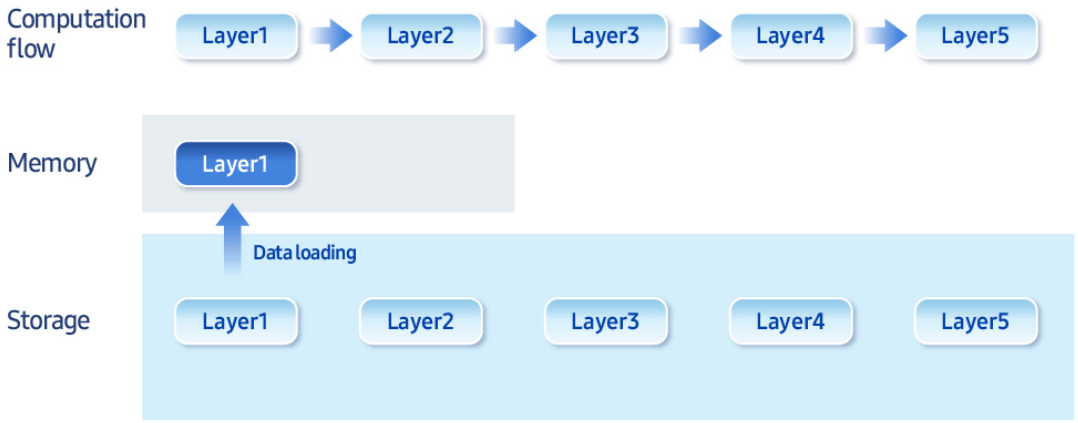
即使模型壓縮得宜，用戶體驗最終仍取決於其在裝置上的實際運作狀況。三星研究院現正開發 AI 執行環境引擎，用於改善裝置在執行期間的記憶體和運算資源使用方式。

Ham 博士解釋：「AI 執行時，本質上就是模型的引擎控制單元。當模型在中央處理器（CPU）、圖形處理器（GPU）和神經處理器（NPU）等多種處理器上運作時，執行環境會自動將各項作業指派給最適合的晶片，並盡可能減少記憶體存取次數，以提高整體 AI 效能。」

AI 執行環境也能讓較複雜的大型模型在同一裝置上以相同速度運作，如此不僅降低反應延遲，還可提升整體 AI 品質，呈現更精準的結果、更流暢的對話體驗，以及更細膩的圖像處理效果。

Ham 博士表示：「裝置端 AI 最大的瓶頸為記憶體頻寬和儲存空間存取速度，因此三星正在開發能智慧平衡記憶體和運算的最佳化技術。」舉例來說，只載入當下所需資料，而非將所有資料皆

儲存在記憶體中，藉此提高效率。博士補充：「三星研究院目前可在不到 3GB（通常需要 16GB 以上）的記憶體上執行 300 億參數的生成式模型。」



▲ AI 執行環境軟體可預測進行權重計算的時機，以減少記憶體用量並提高處理速度。

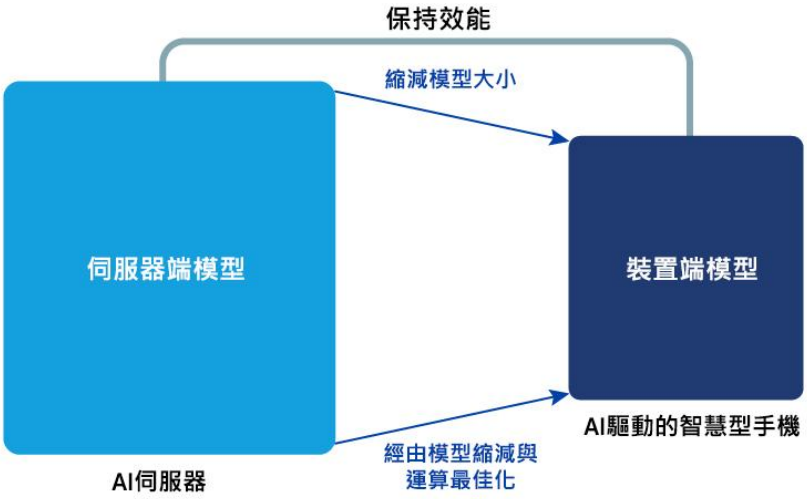
新世代 AI 模型架構

AI 模型架構是 AI 系統的基礎藍圖，相關研究工作也順利推展中。

Ham 博士指出：「由於裝置端環境的記憶體和運算資源有限，我們必須重新設計模型架構，以便在硬體上實現高效運作。三星的架構研究著重於建立能最大化硬體效能的模型。簡言之，我們的目標是從零打造裝置友善的架構，確保模型和裝置硬體從一開始就能完美整合。」

訓練 LLMs 需要投入大量時間與成本，倘若模型架構設計不佳，還可能大幅增加成本。為了減少效率低落的問題，三星研究院會在訓練開始前預估硬體效能並設計最佳化結構。Ham 博士表示：「三星的目標是以最小的晶片達成最高程度的智慧，這是我們致力追求的技術方向。」

如今，大多數 LLMs 皆仰賴轉換器（transformer）架構。轉換器能一次分析整個句子，以確定詞語之間的關係。雖然該方法在理解上下文方面表現出色，但重大限制在於只要句子拉長，計算需求便會急遽上升。Ham 博士說明：「三星目前在多方探索以克服這類限制，並根據實際裝置環境中的運作效率進行個別評估。我們不僅重點改善現有做法，也致力於以全新方法開發新一代架構。」



▲ 架構最佳化研究將大型模型的知識轉移至小型模型，在改善運算效率的同時兼顧效能。

裝置端 AI 的發展前景

裝置端 AI 未來面臨的最大難題為何？Ham 博士回應:「在裝置上直接締造雲端級別運算的效能。為了實現這一點，模型最佳化和硬體效率必須相輔相成，才能在不考慮網路連線的前提下，提供快速、準確的 AI。因此，同步改善速度、精準度和電源效率將成為更重要的課題。」

得益於裝置端 AI 的進展，用戶能隨時隨地享受快速、安全且更具個人化的 AI 體驗。Ham 博士認為：「AI 將變得更善於在裝置上進行即時學習，並適應每位使用者的環境。未來的重點在於提供自然且個別化的服務，同時保障資料安全。」

三星不斷突破極限，透過最佳化裝置端 AI 來打造更上層樓的體驗。憑藉上述行動，三星旨在賦予用戶更卓越、流暢的體驗。