

2025 年【科學探究競賽-這樣教我就懂】

大專/社會組 科學文章格式

文章題目： NPU:專為 AI 而生的處理器

摘要：隨著 AI 時代的快速發展，CPU 與 GPU 在 AI 運算的能力有限，為了因應此困境，NPU 應運而生。NPU(神經處理單元)是專門設計來加速神經運算處理速度，可提供更高效能以及更低的功耗，推動 AI 技術的發展。

文章內容： (限 500 字~1,500 字)

一、前言:

NPU 與傳統的 CPU、GPU 差在哪裡?我們可以把 CPU 比喻為「博士生」，可以解決複雜且多樣的任務；

而 GPU 是一群「數學系大學生」，擅長加減乘除，可以同時執行大量的運算工作；NPU 則是專攻 AI 技術的「技職生」，用在各種 AI 相關的裝置中。

二、NPU 是什麼?

NPU 具有兩大技術特點:第一個是模擬人類神經網路的運作方式，擅長平行運算處理，適當分配晶片內的「任務流」，減少閒置的運算資源；第二個是經由「近記憶體運算(將處理器盡量靠近 DRAM，減少傳輸延遲以及功耗)」或「記憶體內運算(將簡單的邏輯運算移至記憶體陣列中)」，從而提升 AI 運算的執行速度以及效率，並且降低功耗。

三、NPU 的主要架構:

1. 張量計算單元(TPU):

負責進行矩陣和張量操作。使用平行運算處理，處理神經網路中的卷積、矩陣乘法、加權求和等操作。

2. 神經網路引擎(NNE):

用來加速神經網路模型推理的引擎，支持深度學習模型的不同層級的快速計算。

3. 專用的內存和數據路徑:

NPU 通常配備高帶寬的內存系統，以減少數據傳輸的延遲。此外，NPU 會有專門的數據路徑來優化張量運算的處理流程，避免 CPU 或 GPU 的內存延遲問題。

4. 數據重用和壓縮技術:

NPU 支持數據重用和權重壓縮技術，可以減少內存的讀取次數以及數據冗餘，提高效率並且降低功耗。

四、NPU 的計算流程:

1. 數據輸入:

原始數據(圖片、語音、文字等)，傳感器、記憶體或其他處理器。

2. 數據預處理:

格式轉換、歸一化或降維等。將數據轉換為適合神經網路運算的形式。

3. 儲存與緩存:

將預處理後的數據存入 NPU 內部的 SRAM，使後續運算更快速的存取。

4. 計算與激活:

數據送入 NPU 運算單元，執行神經網路運算的各種數學運算和激活函式。

5. 中間結果儲存:

在運算過程中的結果儲存於 NPU 中的 SRAM 內，以便後續存取。中間結果包含每層的輸出、梯度等。

6. 數據後處理:

計算完成後，將數據轉換為所對應輸出的形式。

7. 結果輸出:

將最終處理後的結果輸出。

五、NPU 的發展前景:

隨著 AI 技術的高速發展，對高效能、低功耗的計算加速器的需求只會不斷增長。NPU 的發展也將不斷推進，並融入更多先進的 AI 加速技術。NPU 將持續優化，以供更複雜的 AI 運算，以及更快速地完成任務。

六、結論:

隨 AI 技術的發展，運算愈來愈複雜，需要更強大甚至專為 AI 而生的處理器來運算，而 NPU 因此誕生，改善 CPU 以及 GPU 在 AI 運算上的不足，提供高效能以及低功耗的優點，提升 AI 技術的發展速度，因此 NPU 是 AI 時代的關鍵技術。

參考資料

白話科技 | NPU 是什麼？跟 CPU、GPU 差別是？圖解 AI PC 最重要關鍵字

<https://www.bnext.com.tw/article/79456/aipc-npu-explanation?>

Intel NPU (Neural Processing Unit) 的技術細節、架構和應用

<https://corp.sinya.com.tw/post/1039/>

NPU 的數據路徑

<https://blog.csdn.net/u012276729/article/details/136700421>

註：

1. 未使用本競賽官網提供「科學文章表單」格式投稿，將不予審查。

2. 字數沒按照本競賽官網規定之限 500 字~1,500 字，將不予審查。

PS.摘要、參考資料與圖表說明文字不計入。

3. 建議格式如下：

- 中文字型：微軟正黑體；英文、阿拉伯數字字型：Times New Roman

- 字體：12pt 為原則，若有需要，圖、表及附錄內的文字、數字得略小於 12pt，不得低於 10pt
- 字體行距，以固定行高 20 點為原則
- 表標題的排列方式為向表上方置中、對齊該表。圖標題的排列方式為向圖下方置中、對齊該圖