研究計畫

應用於邊緣裝置之低功耗高精度 AI 加速晶片設計

摘要—

設計一款應用於應用於邊緣裝置之低功耗高精度 AI 加速晶片，無須連網就能進行 AI 訓練，特別適合有低延遲或無線需求的裝置，如自駕車等。通過輕量化模型設計，搭配 CASCADE 加速器架構、重疊資料線、超參數調整等技術，強化 AI 加速器在速度、功耗上的表現。期望能拓展 AI 產品的應用範疇。

Key Word: AI 加速晶片、邊緣裝置、低功耗、高精度、輕量化模型

一、前言

在邊緣裝置上實現 AI 模型已成為趨勢，從牙刷、手錶、手機到掃地機器人和汽車，AI 應用無處不在。然而，AI 需要大量的計算資源，通常需依賴網路來提供算力。傳統的雲端方案無法滿足像自駕車這類應用對低延遲和高傳輸的要求。此外，在某些無網地區或特定邊緣裝置上，我們期望能無線且獨立運作，這進一步凸顯對低功耗、高精度 AI 加速晶片在邊緣裝置中應用的需求。

二、研究動機與目的

AI 產品是現今火燙的題材，題材小到牙刷大到汽車，企業紛紛以「AI」來行銷其產品，並展示眾多便捷功能。然而，許多產品的 AI 並非在邊緣裝置上訓練，而是依賴連網處理，這增加了產品應用的侷限性。特別是像自駕車這類對低延遲有嚴格要求的應用，連網方案無法滿足需求，最終還是必須在邊緣裝置上自行運算。因此，若能開發出一款適用於邊緣裝置的通用、低功耗且高精度的 AI 加速器，不僅能解決這些問題，還能進一步擴展 AI 產品的應用範疇。

三、文獻探討

[1]中提到速度、吞吐量、面積與功耗之間存在著典型的取捨關係（trade-off）。若要提升速度，通常需要增加面積，同時也會伴隨功耗的提升。在處理不同任務時，CPU、GPU 與 FPGA 各有其優勢。GPU 擅長處理平行運算，但散熱問題是一大挑戰。相較之下，FPGA 在相同功耗下可以表現得更佳，尤其當設計者精確掌握時脈訊息時，能夠進一步最佳化設計，提升效能，並有效解決 GPU 的

散熱問題。此外，已有超過六成的文獻選用 FPGA 作為硬體加速器的設計平台。

[2] 一文探討了深度學習模型在無線通訊自動調變辨識 (AMR) 中的應用，旨在透過超參數調整和模型壓縮提升卷積神經網路 (CNN) 的效率及硬體相容性。首先，研究了量化與剪枝技術對兩個現有 CNN 模型準確性及計算成本的影響，接著引入一個新的 CNN 模型。與先前的工作相比，該模型以更低的計算複雜度實現了更高的精度。實驗結果顯示，該模型有效降低了模型大小與計算複雜度且能保持精度，使其適合邊緣設備上的即時應用。

[3] 提出了一個名為 CASCADE 的加速器綜合框架。CASCADE 採用新穎的資料流 CARD，有效管理卷積運算中的不規則記憶體存取，並透過設計空間探索自動優化層級並行性和 FIFO 深度。

[4] 一文針對 YOLOv5 目標辨識網路，提出了一個基於重疊資料流的可重複使用 FPGA 硬體加速器架構，以解決卷積神經網路在行動裝置上部署時面臨的高計算資源消耗、能源消耗過多和模型尺寸過大的挑戰。

[5] 指出隨著物聯網（IoT）時代的來臨，邊緣設備越來越需要獨立的推理能力。過去依賴將數據傳送至雲端處理的方式已不再適用，因為網絡連接不一定隨時穩定。例如在自駕車行駛過程中，車輛需要即時處理大量數據，如果依賴傳統的雲端數據傳輸方式，無法滿足自駕車對低延遲和高傳輸率的需求。因此，終端計算（Edge Computing）成為解決這類問題的關鍵技術，讓設備在邊緣即時處理數據，避免傳輸延遲。

四、研究設計與架構

大多 AI 模型相當龐大且複雜，不適合在邊緣裝置上使用。故本研究將分析不同深度學習模型架構，並輕量化 AI 模型使其適用於邊緣裝置。使用 FPGA 設計 AI 加速晶片，應用如 CASCADE 的加速器框架、重疊資料流設計、參數調整壓縮，開發具有彈性高、通用性佳、低功耗且高吞吐率的加速器架構。搭配國研院半導體中心所開發的「人工智慧系統晶片設計與驗證平台」進行設計，進而製造出即時運算與低功耗的 AI 晶片，並驗證其功能。

以下為研究步驟:

1.分析現今各 AI 所使用的深度學習演算法

2.了解現今 AI 加速晶片的設計流程與架構

3.搭配「人工智慧系統晶片設計與驗證平台」進行設計與驗證

4.燒錄到 FPGA 上進行 AI 進行訓練，觀察其功耗、訓練成效，了解此研究在邊緣裝置上的應用成效。

五、預期實驗結果與分析

不同 AI 所使用的演算法不盡相同，故需要個別進行客製化設計，先設計出適用於邊緣裝置的輕量化 AI 模型。相較於 GPU 訓練的結果，本研究在 FPGA 上運行時，能在相同功耗下運算取得運算結果較優的成果，同時實現能在邊緣裝置訓練的 AI 加速晶片，但功耗處理上可能會面臨一定的困難，特別是在資料存取的過程中會產生大量功耗，需要而外進行研究處理。

六、參考文獻

[1] T. Mohaidat and K. Khalil, "A Survey on Neural Network Hardware Accelerators," in *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 8, pp. 3801-3822, Aug. 2024, doi: 10.1109/TAI.2024.3377147.

[2] M. Vaziri, S. Vakili and J. M. P. Langlois, "Accuracy-Aware Low-Complexity Deep Learning Models for Automatic Modulation Recognition," 2024 International Conference on Computing, Internet of Things and Microwave Systems (ICCIMS), Gatineau, QC, Canada, 2024, pp. 1-5, doi: 10.1109/ICCIMS61672.2024.10690306.

[3] Q. Guo, H. Luo, M. Li, X. Tang and Y. Wang, "CASCADE: A Framework for CNN Accelerator Synthesis With Concatenation and Refreshing Dataflow," in IEEE Transactions on Circuits and Systems I: Regular Papers, doi: 10.1109/TCSI.2024.3452954.

[4] J. Yang, X. Lei, D. Zhu and D. Lei, "Design of FPGA-based Accelerator for Cattle Posture Recognition," *2024 7th International Conference on Computer Information Science and Application Technology (CISAT)*, Hangzhou, China, 2024, pp. 1208-1212, doi: 10.1109/CISAT62382.2024.10695417.

[5] 許雅音。AI & Big Data 的演變趨勢(中)—運算能力篇(西元 2021 年 8 月 16 日)。TAcc+。西元 2022 年 10 月 11 日，取自:https://taccplus.com/technews-2021-08-16/