D22-026

適用於邊際端運算支援稀疏度調整之高能效類神經網路訓練處理器晶片

An Energy-Efficient DNN Training Processor Supporting Sparsity Scaling for Edge Computing

隊伍名稱　帕青哥症候群
Bazinga Syndrome
隊　　長　傅子興 / 臺灣大學電子工程學研究所
隊　　員　李諭奇 / 臺灣大學電子工程學研究所
張峻瑋 / 臺灣大學電子工程學研究所

指導教授

楊家驥　臺灣大學電機工程學系

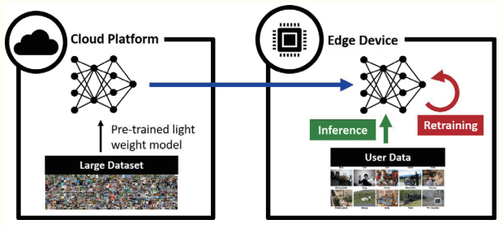
美國加州大學洛杉磯分校電機博士，現為臺灣大學電機工程學系教授。實驗室致力於開發低功耗之客製化晶片以提升資料處理速度與能量效率。

研究領域

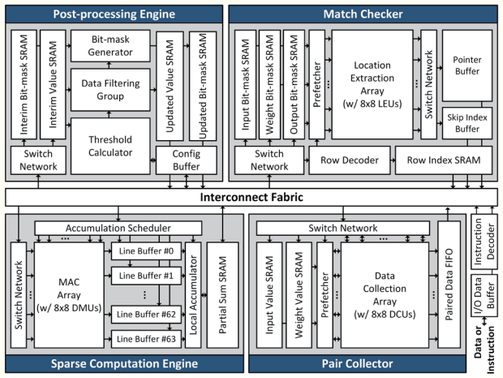
AI 晶片設計、基頻通訊積體電路、生醫訊號處理晶片設計

作品摘要

隨著人工智慧應用的個人化以及隱私權保護需求，邊際端訓練的技術日趨重要。於邊際端進行深層類神經網路的訓練可以在不暴露個人隱私的情況下，協助模型使用邊際端的資料以達到更好的準確度。為了要應對模型訓練的高運算複雜度，在過往的文獻中已經討論了如何利用資料的稀疏度來降低需要處理的運算量。然而，除了原本就存在於模型訓練中的資料稀疏度之外，更高的稀疏度可以由濾除掉較小的中間值來達到。在提升稀疏度的同時，亦可以維持與原本訓練結果相當的模型準確度。利用這種方式，訓練中的資料稀疏度可以提升30%至55%，而最終準確度僅會差異不到2%。伴隨提高稀疏度而來的資料值域縮減使得更進一步的資料量化，例如：使用8位元塊浮點數，成為可能。這類低精度的類定點數運算相比於高精度浮點數（例如16位元浮點數）所需使用的運算單元更小更省能，有助於提升整個設計的能源效率。此外，提高稀疏度的方法也進一步使得資料的搬移更加有效率。外部資料搬移被視為是類神經網路運算加速的重大瓶頸。資料的稀疏度提高意味著在進行資料壓縮後整體的資料量將會減少，再加上資料量化促使個別資料所需的位元數降低，訓練所需的外部記憶體存取將可以大幅度的縮減。本研究採用了改良的提升稀疏度的演算法，並利用了上述提及之所有稀疏度提升所帶來的特性進行設計。我們所設計的晶片為文獻上首個實現可以進行訓練中資料稀疏度提升的晶片，可以將稀疏度提升至超過90%並同時維持最終精確度差異在2%以內；相比於過往文獻之其他設計亦支援最多維度之資料稀疏度的利用。此外，使用我們的訓練方法所需之最高精度為8位元塊浮點數，相比過往之量化方法所需之運算單元面積為最小。我們的晶片達到文獻上最高之646.6TOPS/W之能源效率。相比於過去最佳設計，本研究之晶片之能源效率達到其之3.7倍，面積效率則達到其之4.9倍，為下個世代更複雜的人工智慧應用提供了一個能源與面積使用上更有效率的解決方案。



圖一 邊際端類神經網路訓練



圖二 本設計之系統架構

Abstract

On-device training has become critical for personalized, secure edge AI applications. It enables deep neural network adaptation to achieve a better accuracy by utilizing local information while protecting users'privacy. To tackle the high training complexity, data sparsity has been leveraged and several sparsity-aware DNN training processors have been presented in previous works. Although the sparsity comes from the neural network itself, an even higher sparsity can be achieved by filtering out the smaller interim data while maintaining acceptable training performance. A narrower data range that stems from sparsity-scaling training also enables aggressive quantization, like 8-bit block floating point, for data computation. Compared to long-bitwidth floating point arithmetic, sparsity-scaling training allows for a shorter bitwidth for MAC operations, which is associated with smaller computation unit and higher energy-efficiency. In addition, sparsity-scaling training also makes the data movement more efficient. The high sparsity caused by sparsity-scaling training enables compression on training data, resulting in a smaller amount of data, leading to less energy dissipation and cycle count for accessing the external memory. In this work, these characteristics are fully utilized to enhance both the energy efficiency and the area efficiency. This work presents the first DNN processor that supports sparsity-scaling training. The proposed processor enhances the sparsity of interim data during training, and achieves an over 90% sparsity with less than 2% accuracy loss. Besides, 8-bit block floating point arithmetic and hierarchical bit-mask encoding is utilized for better performance. Our chip achieves an energy efficiency of 646.6TOPS/W. Compared to the state-of-the-art DNN training processors, this work achieves 3.7x [4.9x] improvement in energy [area] efficiency, providing a better hardware solution for next-generation AI applications.

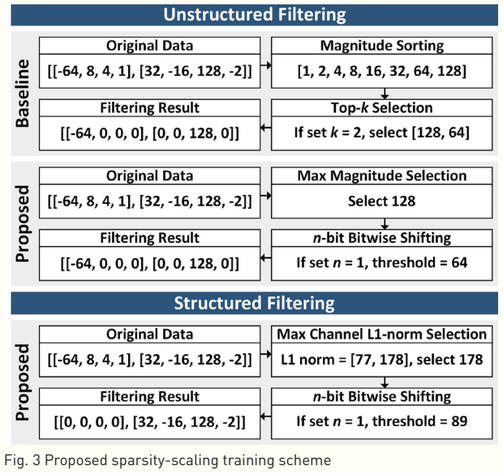


Fig. 3 Proposed sparsity-scaling training scheme

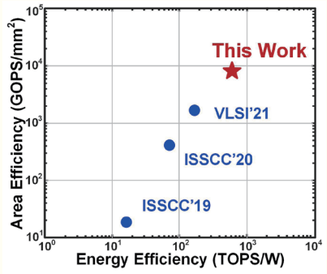


Fig. 4 Hardware performance comparison