設計組
Design Group

D18-040

應用於二進制深度類神經網路以記憶體內運算之靜態隨機存取記憶體

A High Speed, Energy Efficient Computing-In-Memory SRAM For Binary DNN Processors

隊伍名稱　嚇到吃手手 Tasty hand
隊　　長　陳嘉璟 / 清華大學電子工程研究所
隊　　員　李佳芳 / 清華大學電子工程研究所
　　涂詠甯 / 清華大學電機工程研究所
　　吳思妍 / 清華大學電機工程研究所

指導教授

張孟凡
清華大學電機工程學系

交通大學電子工程博士，現為清華大學電機工程學系教授。曾任美國Mentor Graphics工程師、台積電設計服務處主任工程師、積丞科技矽智產事業處處長。

研究領域

記憶體積體電路設計、非揮發性邏輯電路設計、人工智慧晶片之記憶體內運算電路設計。

作品摘要

隨著IoE裝置與無線傳輸系統發展，低功耗處理器成為熱門技術開發。然而，IoE裝置應用系統低電容量與使用環境能源不穩定問題，其系統需具長待機時間與低耗能的特性。在揮發性記憶體中，雖然SRAM的價格比DRAM貴上許多，面積也比DRAM大，但讀取速度更為快速、低功耗（特別是在待機狀態），只要記憶體是在通電的情況下，內存的值就會維持，使得SRAM更為容易控制，相較於DRAM需要週期性的更新資料。因此SRAM首選用於帶寬要求高，或者功耗要求低，通常拿來做為快取記憶體。

傳統處理器為范紐曼（Von Neumann）架構，運算資料於處理器與記憶體巨集間傳遞，此架構速度受限於資料傳輸介面的輸入/輸出端（I/O）數，能量大多消耗在搬動資料的過程，稱為「Von Neumann Bottleneck」。近幾年來，隨著大數據的時代來臨，類神經網絡架構（Neural Network）成為熱門的研究項目。有別於傳統范紐曼架構，類神經網絡可於單晶片中實現並行運算，達到加快速度且降低功耗的優勢。

利用高速且穩定的靜態隨機存取記憶體（SRAM），本研究成功設計應用於深度神經網路之記憶體內運算（Computing-In-Memory）電路與提出三大改善機制有效降低能量損耗、提升運算效率（Efficiency）和精準度（Accuracy）。下列為本研究之電路特色。

1. 首次結合靜態隨機存取記憶體（SRAM）於記憶體內運算（Computing-In-Memory）應用於實際微處理器

2. 記憶體內運算電路（CIM）開發，作為深度神經網路之加速器，可有效降低計算功耗與速度提升。

3. 利用類神經網絡架構可於記憶體巨集內實現平行運算，無須將資料傳遞至處理器再進行運算。

我們提出三大機制
（1）演算法相依非對稱控制機制（ADAC）
（2）動態輸入生成參考電壓機制（DIARG）
（3）低共模敏感與小偏移電壓之電壓式感測放大器（CMI-VSA）

有效利用靜態隨機存取記憶體（SRAM）特性來解決不必要的能量損耗、並提高深度神經網路之加速器於記憶體內運算（Computing-In-Memory）的精準度。

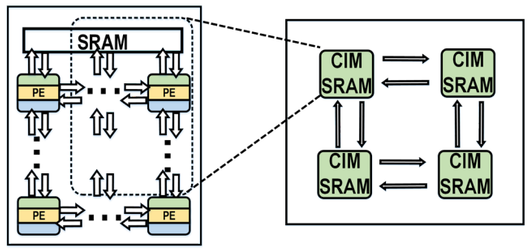

圖1. 我們提出了使用靜態隨機存取記憶體內計算（SRAM-CIM）巨集，使SRAM-CIM中的記憶體不僅可作為進行資料存儲存取亦可進行資料運算

Abstract

With the development of IoE devices and wireless transmission systems, low-power processors have become a popular technology development. However, the IoE device application system has issue of low energy consumption and unstable energy environment, its system needs long standby time and low power consumption. In volatile memory, although the price of SRAM is much more expensive than DRAM, the read speed is faster and the power consumption is lower (especially in standby state). As long as the memory is powered on, the value of the memory will be maintained, making the SRAM easier to control. Compared with DRAM, it requires periodic update of data and random access. Therefore, SRAM is preferred for high bandwidth requirements or low power requirements. It is usually used as a cache memory.

The traditional processor is a Von Neumann architecture. The computing data is transferred between the processor and the memory. The speed of the architecture is limited by the number of input/output (I/O) of the data transmission interface. The energy is mostly consumed in the process of moving data that is called "Von Neumann Bottleneck". In recent years, with the arrival of the era of big data, Neural Networks has become a popular research project. Different from the traditional Von Neumann architecture, the neural network can realize parallel operations in a single chip to achieve high speed and reduce power consumption.

Using high-speed and stable Static Random Access Memory (SRAM), this study successfully designed a computing-in-memory circuit for deep neural networks and proposed three major improvement mechanisms to reduce energy consumption effectively and improve operational efficiency and accuracy. The following are the circuit features of this study:

1. It is the first time combining Computing-In-Memory (CIM) with Static Random Access Memory (SRAM) use for microprocessor.
2. The development of Computing-In-Memory (CIM) circuit, as an accelerator for deep neural networks, it can reduce computational power consumption and speed effectively.
3. Using a neural network architecture to achieve parallel operations in memory without passing data to the processor.

Three major mechanisms are proposed
(1) Algorithm Dependent Asymmetric Control (ADAC)
(2) Dynamic Input Sensing VREF Generation (DIARG)
(3) Common Mode Insensitive Small Offset Voltage Mode Sense Amplifier (CMI-VSA)

Using Static Random Access Memory (SRAM) features to solve unnecessary energy losses and improve the accuracy of computing-in-memory operations for deep neural network accelerators.

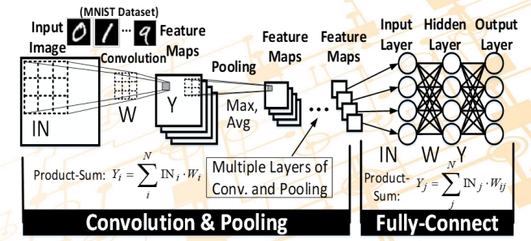

Fig.2 We proposed a Static Random Access Memory Computing-in Memory (SRAM-CIM) macro, which focus on speeding up the fully connected layer in the Convolutional Neural Network (CNN)