**AOC Paper Reading and Review 3**

**Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep**

**N26122246 胡家豪**

1. **Motivation**

本篇paper提出了一種CNN的加速器，稱為「Eyeriss」，其由168個PE組成，並且作者提倡使用「Row stationary」的data flow，並且使用了「RLE」的方式將數據進行壓縮。在這個架構之下，作者表示「Eyeriss」能夠達到high throughput與 energy-efficiency

1. **Proposed solution**

一.Eyeriss Architecture

一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述

在本篇的Eyeriss 架構中，分別有Core Clock 與 Link Clock兩種 Clock，其透過 Asynchronous FIFO將DRAM內的ifmap、Filter傳入。

1.Top-Level Control

Top Control 主要負責 DRAM、Global Buffer以及PE Array之間的溝通。其控制DRAM取出 Filter、Ifmap至 chip上的 Buffer、控制Buffer取出Filter、Ifmap至PE array與Psum的累加。

2.lower -Level Control

一張含有 文字, 圖表, 螢幕擷取畫面, 字型 的圖片

自動產生的描述

在每一個PE單元內均有一個control，並且彼此之間互相獨立。

Lower Control 控制一般PE的MAC運算，值得一提的是作者在ifmap中有額外進行一個判斷是否為0的邏輯，如果成立便會跳過後面的dataflow直接執行下一個非0的運算，如此一來可以減少45%的功耗。

二. RLE(Run Length Encoding)

一張含有 文字, 字型, 螢幕擷取畫面, 行 的圖片

自動產生的描述

由於作者發現較知名且有效的模型，通常會進行ReLU運算。如此一來會導致ofmap的結果有很多0的出現，所以作者在ofmap進行 RLE，如此一來可以用較少的bits表示0的數量，達到節省存儲空間的目的。

三. Row stationary

一張含有 文字, 圖表, 字型, 行 的圖片

自動產生的描述

Row stationary 是本篇首次提出的data flow，主要作法是將Filter與ifmap以row為單位進行運算。

如此是為了最大化減少各種資料的data movement(不論是ifmap, filter 或 psum)，以達到energy efficienct的效果。

1. **Evaluation**

1.Performance

作者使用「Eyeriss」執行VGG-16的model，Batch size 設為3，Supply voltage 為 1V，on-Chip clock 為200MHz ：外部的clock為60MHz。

一張含有 文字, 數字, 字型, 螢幕擷取畫面 的圖片

自動產生的描述

Power(mW)：每一層所花的功率，(total 為平均功率)

Total Latency (ms)：完成此層所花費的時間

Processing Latency(ms)：單純進行MAC所花費的時間

Num of MACs：此層所執行的乘加運算數目

Num of Active PE：在執行此層運算，work的PE比例

Zero in Ifmaps：此層運算時，input feature 有 0的比率

Global Buff Accesses：執行此層運算時，on Chip的Buffer存取了多少data

DRAM Accesses：執行此層運算時，DRAM 的Buffer 存取了多少data

1. **My analysis**

我認為Eyeriss 提出的「Row stationary」算是一個蠻大的創舉。在他之前多數的架構都是用「weight stationary」或是「output stationary」等等比較傳統的方式將一種資料固定，對其進行重複利用。

而Eyeriss 提出的「Row stationary」則是首次將每一個ROW一次計算完畢以後再進行下一個ROW的計算，如此可以最大化weight和activation 在PE array的使用次數。如此一來就可以降低功耗。

但是Eyeriss的PE相較於一般的systolic MAC，Eyeriss的PE控制更加複雜、而且由於PE 內部需要SPAD，所以面積會比較大。

而且由於PE需要等待ifmap與filter的procast，比起systolic的計算還是較為小一些，所以Eyeriss還是在能耗方面比較有優勢。

不過在讀這篇paper之前，過去我只知道systolic array它的核心精神是將資料盡可能的併行處理與讓資料在PE單元內多「流動」，以此減少記憶體的存取，達到加速與節能。但是對於各種stationary的方法雖然聽過，但是都不太知道具體是如何操作，這篇論文讓我更加了解了加速器內部的結構與操作。