DNN Accelerator Design and Implementation
深度神經網路加速器設計與實作

林柏杰、古皓丞

指導教授：張孟凡教授

Abstract

With artificial intelligence getting huge progress and widespread recognition, machine learning, a major subfield of artificial intelligence, has proven to be the technical basis. Among different fields in machine learning, deep neural network has the most applications and research centered around it. In the past, graphic processing units are used as the hardware for computing machine learning model trainings, they reach a far better performance compare to that of central processing units. However, in recent years, more deep learning developments in software and hardware have been conducted to make artificial intelligence processing unit, called AI chip or simply accelerators. The main purpose for these kinds of processors are that they serves an overall better power efficiency and greater computational performances.

摘要

人工智慧在近年來獲得龐大的關注以及進步，而人工智慧當中又以機器學習為主要的技術基礎。其中深度神經網路是機器學習裡頭應用範圍最廣與使用最多的分支。過去有圖形處理器作為機器學習的硬體根基，計算能效遠優於中央處理器，但近年來有更多專為深度學習開發的人工智慧處理器，或稱加速器，其目的就是為了在做深度學習時能夠獲得更高的能效表現。

Introduction

With the application of artificial intelligence (AI) starting to prosper in recent years, scientific research on related field has grown by a significant amount. The foundation of such growth, which is a major subfield of AI, is machine learning (ML). While researchers in ML continues to innovate structures and models, deep neural networks (DNNs) remain in the most crucial part of the neural network subfield. As the name DNN indicates, DNN is deep as in the number of layers. To a general consensus, a model with more than or equal to three layers can be called DNN. There are two main types of layers, convolution layer (CONV or CL) and fully connected layer (FC or FCL). Within each layer, some numerical calculations would be appended at the end for quantization, simplification and nonlinearity, these include pooling and activation functions.

The above mentioned DNN models are originally trained and tested on graphic processing units (GPUs) and achieved high efficiency compared with central processing units (CPU). However, GPUs are still not specifically designed to compute DNNs, therefore a number of application specific integrated circuits (ASIC) arise. Some famous examples include but not limited to Google’s tensor processing unit (TPU), MIT’s Eyeriss series and KAIST neural processing unit (NPU) series. Among these processors, some aim to achieve high computational speed while some aim to achieve low power and energy efficiency.

The result is organized in the following sequence. We first go through some background analysis on how an accelerator is constructed and the overall hardware architecture that is essential to process a DNN model. Next we explain the reasoning behind the model chosen to be realized and the dataset we intended to test on. After that, we dig into details on the description and implementation of each layer and hardware structure by our design. These include the fundamental of an accelerator, processing element (PE), which is used to do most of the multiply-and-accumulate (MAC), pooling and activation function in DNN models, memory segments for weight/filter loader, input activation/feature map loader and buffer to the off chip DRAM. Following is the integration on different parts within the accelerators. An hardware implemented result and result from software is compared for an individual layer, overall output may be included if outcome came out in time as desire. Last but not least is the lesson learned throughout the course year in special topic implementation I & II following up with future work roadmap and improvement as in how we could have done and modification on our work.

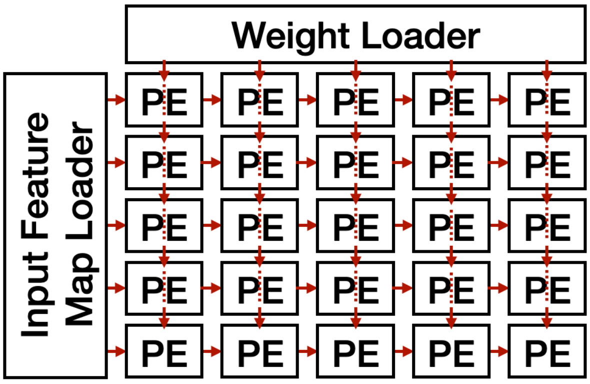
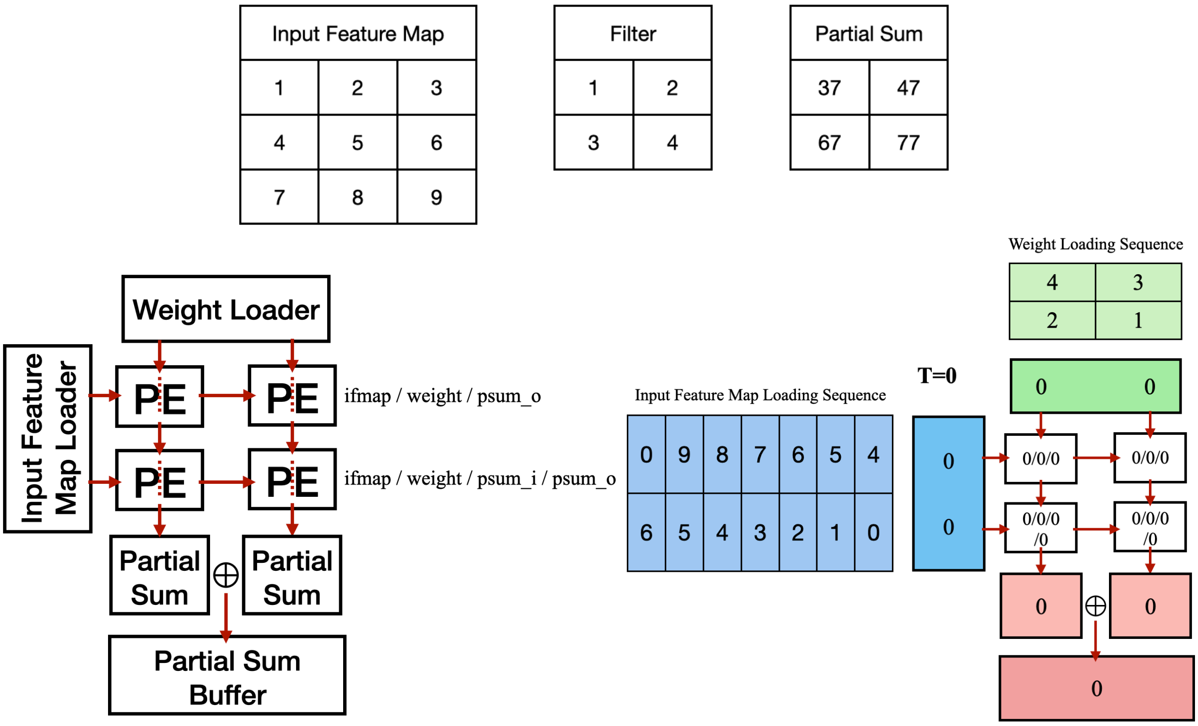


Figure 1. Systolic array of the accelerator



Input Feature Map		
1	2	3
4	5	6
7	8	9

Filter	
1	2
3	4

Partial Sum	
37	47
67	77

Weight Loading Sequence

4	3
2	1

Input Feature Map Loading Sequence

0	9	8	7	6	5	4
6	5	4	3	2	1	0

Figure 2. Example of dataflow for convolution layer (3x3 input feature map and 2x2 filter)

心得感想

經過一年下來的專題訓練，從最初的讀論文開始，發現在類神經網路的硬體設計上，原來已經有那麼多的架構。原本連裡面的分層以及計算都毫無頭緒，到過了一學期以後，這些專有名詞都深深烙印在腦中，也大致上有一點概念與可以突破的地方。一開始並不知道從何下手，但有賴於張孟凡指導教授與實驗室的學長姐提供意見與想法以後，才知道如何正確地看一篇論文，知道他的優勢以及劣勢，而後朝是否有改進的方向下手。在加速器方面，目前能夠完整過出一顆可以通用的晶片並且下線是相對困難的，但我們也在這次經驗中了解如何初淺的製作裡面該有的硬體架構，整體的設計計算方面反而不是最大的問題來源，真正困難的點反而是如何排序資料，並且收取有用的輸出。這些都是在這一年專題中團隊合作以及溝通中而來的。除了在閱讀論文時遇到諸多困難以外，發想自己的想法時同樣也遇到很多挫折，可能是不知道如何下手，也可能是以為很好的想法其實已經有人探討過並且發表了，所以只算是套用別人的研究成果，不過加速器的設計方面有許多點可以切入，所以我們大致上算獨立製作了一個半成品，未來也希望可以繼續改進並且能夠實現完整的運算。