AI on Chip 2024

LAB II Quantization

REPORT

Student name: \_\_胡家豪\_\_\_\_\_\_\_\_\_

Student ID: \_\_N26122246\_\_\_\_\_\_\_\_

目錄

[一Question List 3](#_Toc162542642)

[1. Question. 1 3](#_Toc162542643)

[2. Question. 2 3](#_Toc162542644)

[3. Question. 3 3](#_Toc162542645)

[二.Problem 4](#_Toc162542646)

[1.Problem 1 4](#_Toc162542647)

[2.Problem 2 4](#_Toc162542648)

[三.Paper reading questions 4](#_Toc162542649)

[四.My opinion 4](#_Toc162542650)

1. Question List
2. Question. 1

一張含有 文字, 螢幕擷取畫面, 陳列, 軟體 的圖片

自動產生的描述

1. Question. 2

一張含有 文字, 螢幕擷取畫面, 字型, 軟體 的圖片

自動產生的描述

1. Question. 3

一張含有 文字, 螢幕擷取畫面, 軟體, 多媒體軟體 的圖片

自動產生的描述

1. Problem
2. Problem 1 - What is the size of the model after int8 quantization if its original size is 50MB?
3. Problem 2- If M = 0.2, determine values for M0 and n such that the equation on page 11 is true

、

1. Paper reading questions

軟體：由於在推論時參數已經固定，所以可以將batch norm的參數fold進linear layer，如此一來可增加效率。具體的計算是利用下列公式將權重折疊

一張含有 文字, 字型, 行, 螢幕擷取畫面 的圖片

自動產生的描述

另外，作者也提到，一般的fuse layer中，他會選擇將bias使用32 uint進行運算。這樣做的原因是因為32uint可以降低模型的偏差，所以不會捨棄bias而且還使用32bits。但是如此一來，如果要使用int8的話，必須將其在進入activation layer之前再次轉換成uint8。

硬體：在硬體上，量化的式子因為可以寫成

一張含有 字型, 文字, 白色, 印刷術 的圖片

自動產生的描述

這條式子只有M的部分是浮點數(而且通常小於1大於0)，所以我們可以將其拆解成



這部分我覺得論文說明有些奇怪，但是我覺得他想表達的是將M使用整數(M0)做乘上2的負冪次，如此一來，上面的式子除了2的負冪次都為整數，而2的負冪次則使用位移進行運算。如此可以均使用整數的乘法器。

1. My opinion

本次作業讓我更加理解量化的過程：分成不須

額外進行訓練的 Post-training quantization (PTQ)與利用訓

練尋求更佳效果的 Qunatiztion-aware training (QAT)。

透過量化，可以降低硬體的功率與面積(因為不需要浮點運算)。但是在理解量化的過程還是令人有些吃力。因為量化不只分為上述兩種，而且還有針對不同layer而要進行不同的量化方式，而每一種量化方式又各有不同。

不過本次作業給的範例code算是清晰明瞭，只要好好研究還是可以一窺量化的奧妙。所以即使流程繁瑣，還是能夠從中學到不少東西。