Roll No: 2023102062

Invigilator's Signature: _____

Instructions:

1. Write your answers clearly and concisely on the answer sheet provided.
2. Duration of the exam is 75 minutes.
3. Return the question paper with your answers plus any additional sheets you have taken before you leave the examination hall.
4. **Make suitable assumptions. No questions will be answered during the exam.**

Question	Points	Score
1	20	
2	20	
Total:	40	

1. **TinyCNN on ESP32**

Consider a simple convolutional neural network, called *TinyCNN*, designed for digit recognition. The network takes a color image of size 32×32 with 3 input channels (RGB) as input and processes it through the following layers:

- **Input:** $3 \times 32 \times 32$ image
- **Convolutional Layer:** 16 filters of size 3×3, stride 1, padding 1
- **ReLU Activation**
- **MaxPooling Layer:** 2×2 pooling
- **Flatten Layer:** converts the feature map into a vector of length 4096
- **Fully Connected Layer:** $4096 \rightarrow 10$ outputs (digit classes 0–9)

Answer the following:

(a) Compute the following quantities in **bytes**, assuming each activation and weight is stored as a 32-bit (4-byte) floating point number. [5 pts]

(1) Size of the **input activations** for the CNN layer.
(2) Size of the **CNN filter parameters** (weights + biases).
(3) Size of the **output activations** for the CNN layer.
(4) Sizes of the **input and output activations** for the ReLU layer.
(5) Sizes of the **input and output activations** for the MaxPooling layer.
(6) Sizes of the **input and output activations** for the Fully Connected (FC) layer.
(7) Size of the **FC layer parameters** (weights + biases).

(b) Compute the total number of **multiply–accumulate (MAC) operations** required in: [3 pts]

(1) The CNN layer.
(2) The Fully Connected (FC) layer.

(c) Consider the ESP32 microcontroller, which has a dual-core Tensilica processor running at 240 MHz. Assume each core can issue one 32-bit floating-point MAC per cycle, and that the memory bandwidth is 400 MB/s. *Assume ideal data reuse, i.e., each activation, weight, and output is read/written only once.* [6 pts]

(1) Compute the **peak compute performance** (in MAC/s) of the ESP32.
(2) For the CNN layer in TinyCNN, determine whether execution is **compute-bound** or **memory-bound**, using the roofline model.
(3) Repeat the same analysis for the Fully Connected (FC) layer.

(d) If either the CNN or the FC layer is identified as **memory-bound**, what **software technique** could be applied to increase the utilization of the available peak compute power? What challenges might you foresee in implementing such optimizations on the ESP32 microcontroller? [3 pts]

(e) Suppose the ESP32 has **no cache memory** but **50 floating-point registers**. Re-evaluate whether the **CNN layer** would be compute-bound or memory-bound under this assumption. **Specify the computational technique** you would employ that leads to your conclusion. [3 pts]

2. **Roofline analyses of CPU, GPU and TPU**

Consider the roofline plots in Figure 1 for (a) Google TPU die, (b) Intel Haswell CPU die, and (c) Nvidia K80 GPU die:

Answer the following:

(a) Identify the **memory bound** and **compute bound** regions for TPU, CPU and GPU given their ridge points as 1350 MAC Ops/Byte, 13 MAC Ops/Byte and 9 MAC Ops/Byte respectively. [3 pts]

(b) Which of the models (LSTM0, LSTM1, MLP0, MLP1, CNN0, CNN1) are **bottlenecked by memory bandwidth** in: [3 pts]

(1) TPU.
(2) CPU.
(3) GPU.

(c) Which of the models (LSTM0, LSTM1, MLP0, MLP1, CNN0, CNN1) are **bottlenecked by peak computation rate** in: [3 pts]

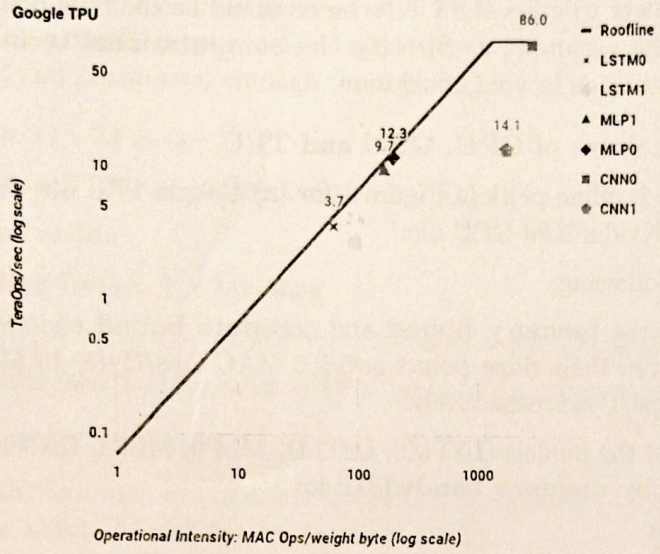
(1) TPU.
(2) CPU.
(3) GPU.

(d) What does the long slanted part of roofline model for TPU (Figure 1(a)) indicate about its performance? [2 pts]

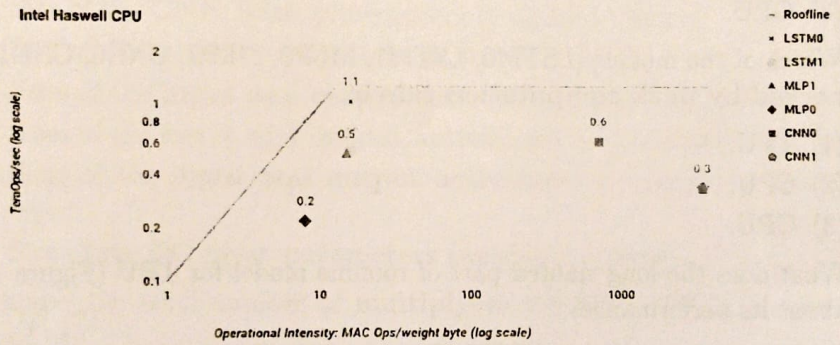
(e) In the roofline model for TPU (Figure 1(a)), CNN1, despite having high operational intensity, is running at only 14.1 TOPS/s compared to CNN0 running at 86 TOPS/s. What could be the reason for the low TPU performance on CNN1? [3 pts]

(f) What does the gap between actual TOPS/s and the ceiling (roofline) indicate as we see in the figures for some of the models? [3 pts]

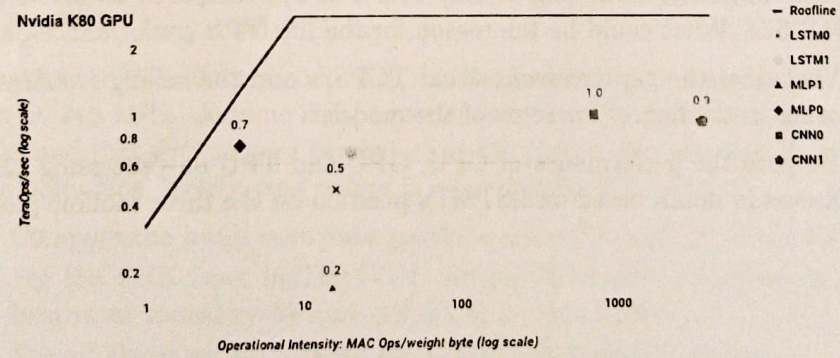
(g) Compare the performance of CPU, GPU and TPU on processing LSTM1 model. Discuss in detail based on LSTM1's position on the three roofline plots. [3 pts]



(a)



(b)



(c)

Figure 1: Roofline plots of (a) Google TPU. (b) Intel Haswell CPU. (c) Nvidia K80 GPU.