LAB 5: STUDENT WORKSHEET

# Arithmetic and Number Representations

**Name:**  **Student ID:**  **Date:**

# PART 1: Number Calculation

Q1: How to present 0.15625 in 32-bit IEEE 754 floating-point format, and how do you calculate?

Q2: What is the following IEEE half precision (IEEE FP16) number in decimal, and how do you calculate?

FP16: 1 10111 1010000000

Decimal Answer:

Q3: What is the decimal 5 in Brain Float (BF16), and how do you calculate?

# PART 2: Quantization

Complete the table below with the numerical values from your experiments:

1. **Quantization to Int**

|  |  |  |
| --- | --- | --- |
| Quantization method | MAE | MSE |
| Int8 |  |  |
| Int6 |  |  |
| Int4 |  |  |

* + 1. Which quantization method has the minimize MSE?

**Quantization to Float point**

|  |  |
| --- | --- |
| Quantization method | Average Absolute Errors |
| FP16 |  |
| BF16 |  |
| FP8\_E4M3 |  |
| FP8\_E5M2 |  |

* + 1. Can you explain the construction of FP8\_E4M3?
    2. Why are there reconstruction error when converting reduced-precision values back to FP32?
    3. Which quantization method has lower average absolute error, and why? Can you explain why FP8\_E4M3 and FP8\_E5M2 have different average absolute error, and what may cause that?

1. **Exploring Microscaling (MX) data formats for DNN**

**2.1 Preparation**

1. What is the size of the weight\_matirx and bias\_vector?

**2.2 Quantize and evaluate the output matrix with different MX data formats.**

|  |  |  |
| --- | --- | --- |
| Quantization method | MAE | MSE |
| mxint8 |  |  |
| mxint4 |  |  |
| mxint2 |  |  |
| mxfp8\_e5m2 |  |  |
| mxfp8\_e4m3 |  |  |
| mxfp6\_e3m2 |  |  |
| mxfp6\_e2m3 |  |  |
| mxfp4\_e2m1 |  |  |
| mxfp16 |  |  |
| mxbf16 |  |  |

1. What quantize method has the lowest and highest MAE and MSE when fixing the block size, and what might cause that?
   1. **Compare MX\_int8 with int8 matrix.**

|  |  |  |
| --- | --- | --- |
| Quantization method | MAE | MSE |
| mxint8 |  |  |
| int8 |  |  |

1. Comparing mxint8 and int8, which has the lower MAE and MSE, and what might cause that?
   1. **Test MX\_int8 for various block sizes**

|  |  |  |
| --- | --- | --- |
| Block size | MAE | MSE |
| 1 |  |  |
| 2 |  |  |
| 3 |  |  |
| 4 |  |  |
| 5 |  |  |
| 6 |  |  |
| 7 |  |  |
| 8 |  |  |
| 9 |  |  |
| 10 |  |  |

1. When we increase the block size, how do MAE and MSE of mx\_int8 change, what might cause that?
   1. **Test mx\_int quantization (mx\_int8/mx\_int4/mx\_int2) for various block sizes**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Block size | MAE | | | MSE | | |
| Mxint8 | Mxin4 | Mxint2 | Mxint8 | Mxin4 | Mxint2 |
| 2 |  |  |  |  |  |  |
| 4 |  |  |  |  |  |  |
| 8 |  |  |  |  |  |  |
| 12 |  |  |  |  |  |  |
| 16 |  |  |  |  |  |  |
| 24 |  |  |  |  |  |  |
| 32 |  |  |  |  |  |  |
| 40 |  |  |  |  |  |  |
| 48 |  |  |  |  |  |  |
| 56 |  |  |  |  |  |  |
| 64 |  |  |  |  |  |  |

1. Which quantization method has the lowest MAE and MSE? When we increase the block size, how do MAE and MSE of mx\_int change, what might cause that? Why does MAE of mxint become stable when increasing the blocksize?
   1. **Test mx\_fp quantization mx\_(fp8\_e5m2, fp8\_e4m3, fp6\_e3m2, fp6\_e2m3, fp4\_e2m1, fp16, bf16) for various block sizes**
2. Can you compare the MAE and MSE of different quantization method, such as fp8\_e5m2, fp8\_e4m3, fp6\_e3m2, fp6\_e2m3, fp4\_e2m1, fp16, bf16?
   1. **Evaluate quantized DNN with different MX data formats.**

|  |  |  |
| --- | --- | --- |
| Quantization method | Accuracy (blocksize=32) | Accuracy (custom) |
| Without Quant |  |  |
| mxint8 |  |  |
| mxint4 |  |  |
| mxint2 |  |  |
| mxfp8\_e5m2 |  |  |
| mxfp8\_e4m3 |  |  |
| mxfp6\_e3m2 |  |  |
| mxfp6\_e2m3 |  |  |
| mxfp4\_e2m1 |  |  |
| mxfp16 |  |  |
| mxbf16 |  |  |

1. How does different quantization method influence the accuracy? Can you change the block size to another number and record the accuracy in above table?

# PART 4: REFLECTION (100 words)

Write a short reflection on model selection criteria for different applications based on your experiment results.

# INSTRUCTOR COMMENTS

Grade: /