# Hardware Architecture for Deep Learning - CS6490. Spring 2024-25.

## Dept. of CSE, IIT Hyderabad

### Assignment-1: Number representation and Quantization

--------------------------------------------------------------------------------------------------------------------

We discussed various common representations for numbers. Let us identify examples and trade-offs in using these representations. For the given pair of types, answer the questions given later:

Pair of datatypes:

| Sl. No. | TYPE-1 | TYPE-2 |
|---|---|---|
| 1 | INT32 | standard FP32 (floating point 32-bit) |
| 2 | INT32 (fixed-point with 18-bit for integer and 12-bit for fractions) | standard FP32 |
| 3 | INT32 (fixed-point with 14-bit for integer and 18-bit for fractions) | standard FP16 |
| 4 | INT32 (fixed-point with 14-bit for integer and 18-bit for fractions) | bfloat16 |
| 5 | standard FP16 | bfloat16 |
| 6 | INT8 | standard FP16 |

**Questions to answer for each pair:**

a. How many unique numbers can each of them represent
b. Compare the two types in terms of supported range and precision
c. Show and explain an example number which can be represented in TYPE-1 but not in TYPE-2
d. Show and explain an example number which can be represented in TYPE-2 but not in TYPE-1

**General guidelines:**

1. This assignment is to be done individually by all students crediting the course
2. Audit students are not required to do this assignment
3. Submit a pdf file containing your answers. Write necessary details only
4. This should be your own work and not copied from any other source
5. Sufficient time is given, so NO late submission is expected. If submitting late, a 10% penalty for a delay of every 24 hours or part of it.