INT8/INT4 Quantization

Integer Quantization

Converting floating-point (FP32/FP16) to integers (INT8/INT4) to reduce model size

Bit Count Comparison

FP32

32 bits
4 bytes
Default training precision

INT8

8 bits
1 byte
75% memory reduction

INT4

4 bits
0.5 bytes
87.5% memory reduction

Benefits of Quantization

- Reduced memory usage
- Improved inference speed
- Lower power consumption

Accuracy Trade-off: INT8 typically shows less than 1% accuracy degradation
INT4 requires Quantization-Aware Training