Part 2/3:

Quantization and Pruning

1. INT8/INT4 Quantization

2. Mixed Precision Strategies

3. Structured Pruning

4. Magnitude Pruning

5. Lottery Ticket Hypothesis

6. Dynamic Sparsity