Tensor Operations and Hardware Optimization

Tensor: Multi-dimensional array (Foundation of deep learning)

Common Tensor Operations

matmul
Matrix multiplication

Parallelizable

conv2d
2D convolution

Parallelizable

attention
Self-attention

Parallelizable

Coalesced Access

Sequential memory reads

⚡ Fast & efficient

Uncoalesced Access

Random memory reads

⚠ Slow & inefficient

Optimization Strategies

- ✓ Batch operations together
- ✓ Maintain contiguous memory

Tensor Cores

Specialized hardware for mixed-precision matmul

✓ Minimize CPU-GPU transfers