Structured Pruning

Structured Pruning

Structural compression by removing entire channels, filters, or layers

Types of Structured Pruning

Channel Pruning

[■][■][□][■]
Channel-level removal

Filter Pruning

Filter 1: ■■■■
Filter 2: XXXX
Filter 3: ■■■■

Layer Pruning

Layer 1 → ■
Layer 2 → X
Layer 3 → ■

Advantages of Structured Pruning

✓ Hardware-friendly: No special hardware required

✓ Real speed improvement: Maintains dense matrix operations

✓ Memory reduction: Actual reduction in parameter count

Medical Application: X-ray classification model with 40% channel removal → 2x faster inference, accuracy 98.5% → 97.8% (less than 1% decrease)