Model Compression Techniques Comparison

Quantization

Reduce numerical precision

Size Reduction

4× (75% reduction)

Speed Improvement

2-4× faster

Accuracy Impact

<1% loss

Implementation

- Post-training: Easy
- QAT: Moderate
- Framework support: Good

Hardware Requirements

- CPU: Excellent
- GPU: Good (Tensor Cores)
- Mobile: Excellent

Best For

- CPU deployment
- Mobile/edge devices
- Memory-constrained environments

Knowledge Distillation

Train smaller student model

Size Reduction

2-3× (40-60%)

Speed Improvement

2-3× faster

Accuracy Impact

1-3% loss

Implementation

- Requires teacher model
- Training time: High
- Complexity: Moderate

Hardware Requirements

- CPU: Excellent
- GPU: Excellent
- Mobile: Good

Best For

- High-volume production
- Best accuracy-size trade-off
- Long-term deployment

Pruning

Remove unnecessary parameters

Size Reduction

1.5-2× (30-50%)

Speed Improvement

1.3-2× faster

Accuracy Impact

1-2% loss

Implementation

- Structured: Easier
- Unstructured: Complex
- Validation required

Hardware Requirements

- CPU: Good (structured)
- GPU: Good (structured)
- Sparse: Needs support

Best For

- Targeted optimization
- Removing redundant components
- Combine with others

Production deployments often combine multiple techniques:
Quantization + Distillation can achieve 6-8× cost reduction with <2% accuracy impact