Model Compression Techniques

Pruning

Remove unimportant weights/neurons

Structured or unstructured

Quantization

Reduce precision (FP32 → INT8)

~4× memory/speed improvement

Knowledge Distillation

Train small student to mimic large teacher

Transfer knowledge efficiently

Low-Rank Factorization

Decompose weight matrices into smaller components

Reduce parameter count

Weight Sharing

Group weights to reduce unique parameters

Efficient representation

Typical Compression Results

Compression Ratio	Accuracy Loss
5-10×	**<1%**

Mobile Deployment

Essential for phones, IoT devices

Tools

TensorFlow Lite · PyTorch Mobile · ONNX Runtime