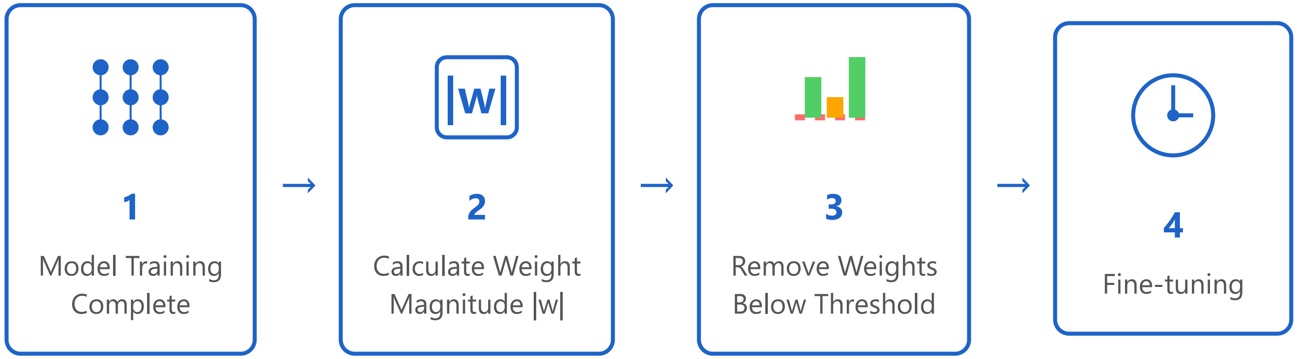
Magnitude Pruning

Magnitude-Based Pruning

Increase sparsity by setting small weight values to 0

Pruning Process

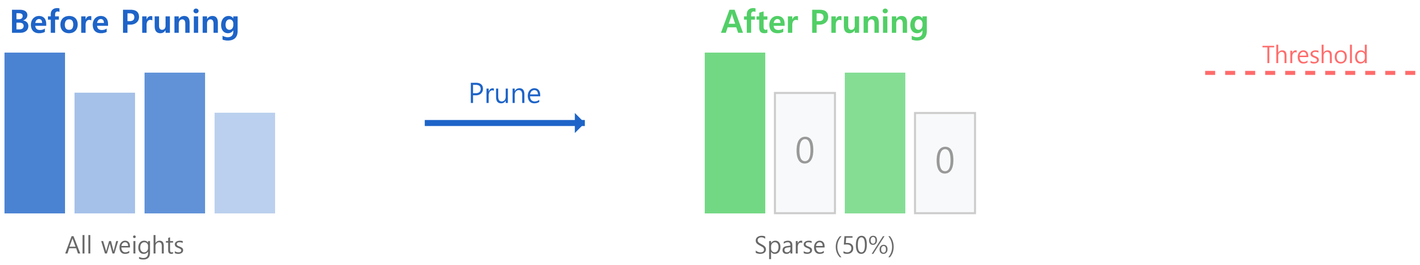


1 Model Training Complete →

2 Calculate Weight Magnitude |w| →

3 Remove Weights Below Threshold →

4 Fine-tuning



Before Pruning

After Pruning

Threshold Setting Strategies

θ

Global Threshold

Single threshold for entire network

θ_1 θ_2 θ_3

Layer-wise Threshold

Different threshold per layer

Top-k

Top-k Pruning

Keep only top k% weights

Sparsity Example

Dense: [0.8, 0.3, -0.5, 0.1, -0.9]

↓ Threshold = 0.4

Sparse: [0.8, 0, -0.5, 0, -0.9]

0.8 ✗ -0.5 ✗ -0.9

40% Sparsity Achieved