Model Optimization Comparison Results

Cherry Mathew Roy - Lab 2

Model Architecture Summary

- **Input size**: 224 × 224 × 3 (RGB image)
- **Key layers**: GlobalMaxPooling2D, Dense, BatchNormalization, Dropout
- **Base model**: EfficientNet with ImageNet pre-trained weights

Training Results

Epochs	Test Accuracy	Comments
3	≈ 91%	Fast but not fully converged
15	≈ 96%	Accuracy plateaus after ~12 epochs; no overfitting with early-stopping

Optimization Results

Model Type	Test Accuracy (%)	File Size (MB)	Size Reduction	Notes
Baseline FP32	≈ 96%	≈ 29	1× (reference)	Full precision
Float-16	≈ 95.8%	≈ 14-15	≈ 2× smaller	Only weight precision changes; activations use float math
Dynamic Range Quantization	≈ 95.4%	≈ 8	≈ 3.5× smaller	Weights become int8, activations stay float
Integer Quantized (Int8)	≈ 94%	≈ 7	≈ 4× smaller	Both weights and activations quantized

Pruned (50% sparsity)	≈ 96%	-	-	Minimal accuracy impact
Pruned (70% sparsity)	≈ 95.5%	≈ 20	≈ 1.5× smaller	Good accuracy, modest size reduction
Pruned (80% sparsity)	≈ 95%	-	-	Higher sparsity increases compression but costs accuracy
Pruned (50%) + Int8	-	-	≈ 6-8× smaller	Best size-to-accuracy ratio

Model Size & Accuracy Comparison

Key Findings

1. **Quantization Effects**:
    - Quantization applies uniform precision changes to all parameters
    - Accuracy impact is typically small and predictable
    - Provides computational speedups as int8/FP16 operations are faster than FP32
2. **Pruning Effects**:
    - Creates sparse networks by removing connections
    - Minimal accuracy impact up to ~70% sparsity
    - Sharp accuracy decline beyond 70% sparsity as network capacity diminishes
    - Requires specialized sparse kernel libraries for runtime speedups
3. **Optimal Solution for Edge Deployment**:
    - Hybrid approach: moderate pruning (≈50%) + INT8 quantization
    - Achieves 6-8× size reduction with <2pp accuracy loss
    - Maintains >94% accuracy for binary classification tasks
    - Best choice for constrained devices like ESP32-S3 microcontrollers
    - Alternative: standard INT8 quantization if sparse kernels aren't supported
    - Float-16 is ideal for GPUs/NPUs that support half-precision arithmetic when accuracy is critical