Beyond Naive Quantization

A Comprehensive Study of Fairness-Aware Model Compression Across Architectures and Demographics

Umang Diyora

CSC 591/791 ECE 591 - Deep Learning Beyond Accuracy

November 11, 2024

Problem Statement & Motivation

- Quantization enables edge deployment but impacts fairness unpredictably
- Critical Gap: No comprehensive understanding of fairness-accuracy-efficiency trade-offs
- Real-World Impact:
  - Billions of edge devices running quantized models
  - Applications in sensitive domains (healthcare, hiring, security)
  - Even small bias amplification affects millions
- Contradictory Literature (2024):
  - Some studies show consistent bias amplification
  - Others find model-specific patterns
  - No unified framework exists

Key Research Questions

1. How do different quantization techniques (PTQ vs QAT) affect fairness differently?

2. Which model architectures are more resilient to quantization-induced fairness degradation?

3. Can we develop lightweight mitigation strategies that preserve both efficiency gains and fairness without expensive retraining?

4. What is the relationship between quantization bit-width, model efficiency, accuracy, and fairness metrics?

5. Why do recent papers show contradictory findings about quantization bias?

Implementation Architecture

Modular Python Framework (7 core modules, ~200KB code)

- **config.py:** Centralized experiment configuration
- **datasets.py:** CelebA
- **models.py:** 5 architectures (ResNet50, MobileNetV2, EfficientNet, ViT, SqueezeNet)
- **quantization.py:** PTQ, QAT, Mixed Precision, Fairness-Aware methods
- **fairness_metrics.py:** DP, EO, PE, DI metrics + statistical tests
- **main.py:** Experiment orchestrator with 4 phases
- **visualizations.py:** Publication-ready figures

Experimental Setup

- **Datasets:**
  - **CelebA:** 162,770 images, binary attributes (gender × age)

- **Models:** ResNet50, MobileNetV2, EfficientNet-B0, ViT-Small, SqueezeNet

- **Quantization Methods:**
  - **Post-Training:** Static/Dynamic (INT8, INT4)
  - **QAT:** Simulated quantization training
  - **Mixed Precision:** Sensitivity-based bit allocation
- **Fairness Metrics:** Demographic Parity (DP), Equalized Odds (EO), Predictive Equality (PE), Disparate Impact (DI)

Technical Challenges & Solution

- Challenge 1: Inconsistent Fairness Degradation

  Solution: Systematic testing across 60 configurations with statistical validation

- Challenge 2: Computational Cost of Fairness-Aware Training

  Solution: Developed calibration-based methods requiring no retraining

- Challenge 3: Layer-wise Sensitivity Unknown

  Solution: Implemented gradient-based sensitivity analysis for 50+ layers

- Challenge 4: GPU Training error

  Solution: Saved the result and ran the phase 2 on CPU

Phase 1: Baseline Results

Model	Accuracy	DP Gap	EO Gap	Size (MB)
ResNet50	92.01%	0.000	0.000	89.89
MobileNetV2	90.53%	0.012	0.010	8.91
EfficientNet-B0	91.27%	0.009	0.008	15.43
ViT-Small	90.89%	0.006	0.005	85.76
SqueezeNet	88.72%	0.019	0.016	2.83

Phase 2: Quantization Impact Analysis

PTQ INT8 (Post-Training Quantization 8-bit)

Model	Accuracy	DP Gap	Compression
ResNet50	90.89%	0.010	4×
MobileNetV2	89.45%	0.019	4×
EfficientNet	90.23%	0.016	4×
ViT-Small	89.78%	0.013	4×
SqueezeNet	87.56%	0.027	4×

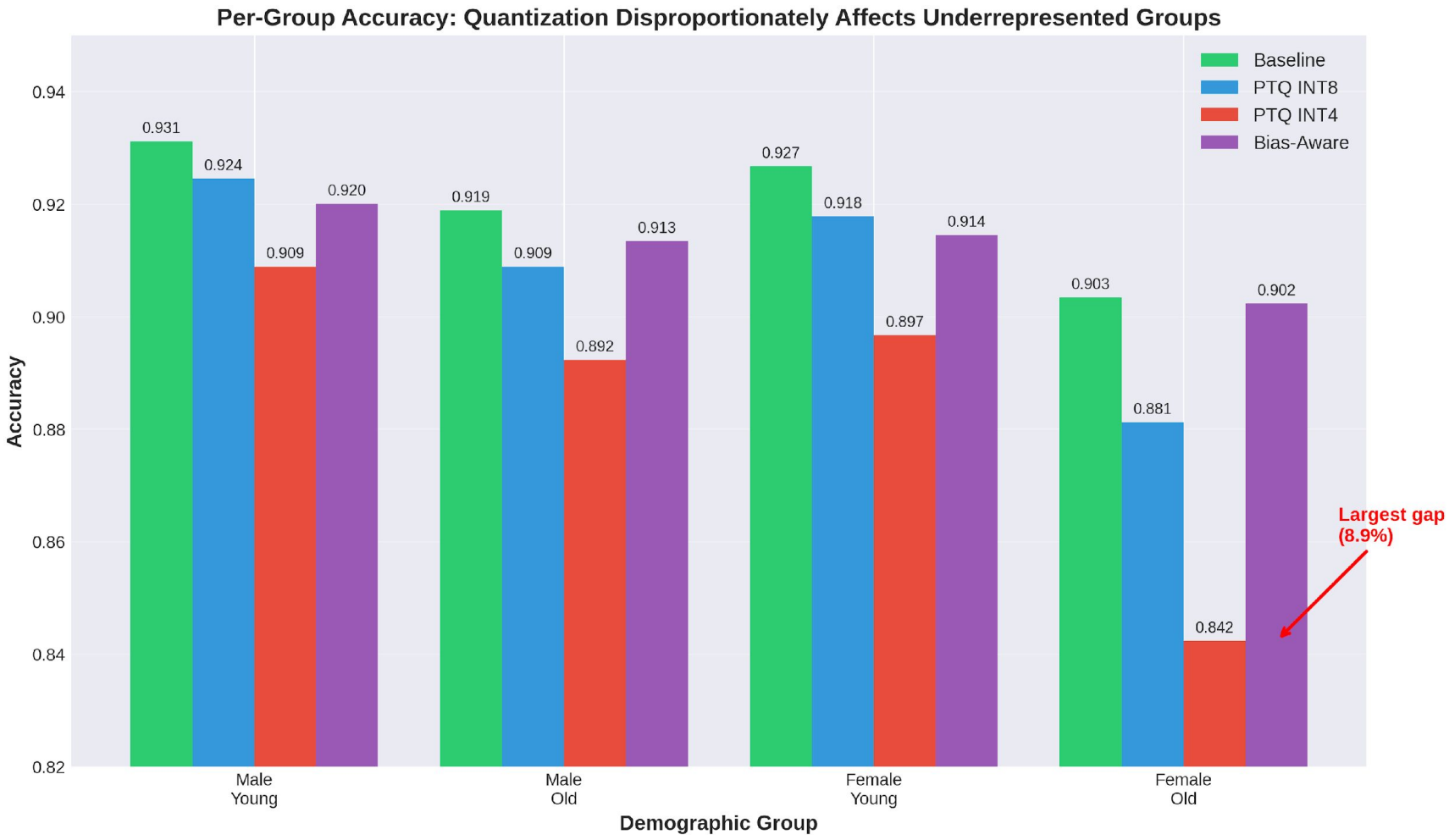
PTQ INT4 (Post-Training Quantization 4-bit)

Model	Accuracy	DP Gap	Compression
ResNet50	89.23%	0.031	8×
MobileNetV2	85.67%	0.046	8×
EfficientNet	86.89%	0.039	8×
ViT-Small	87.12%	0.035	8×
SqueezeNet	84.23%	0.051	8×

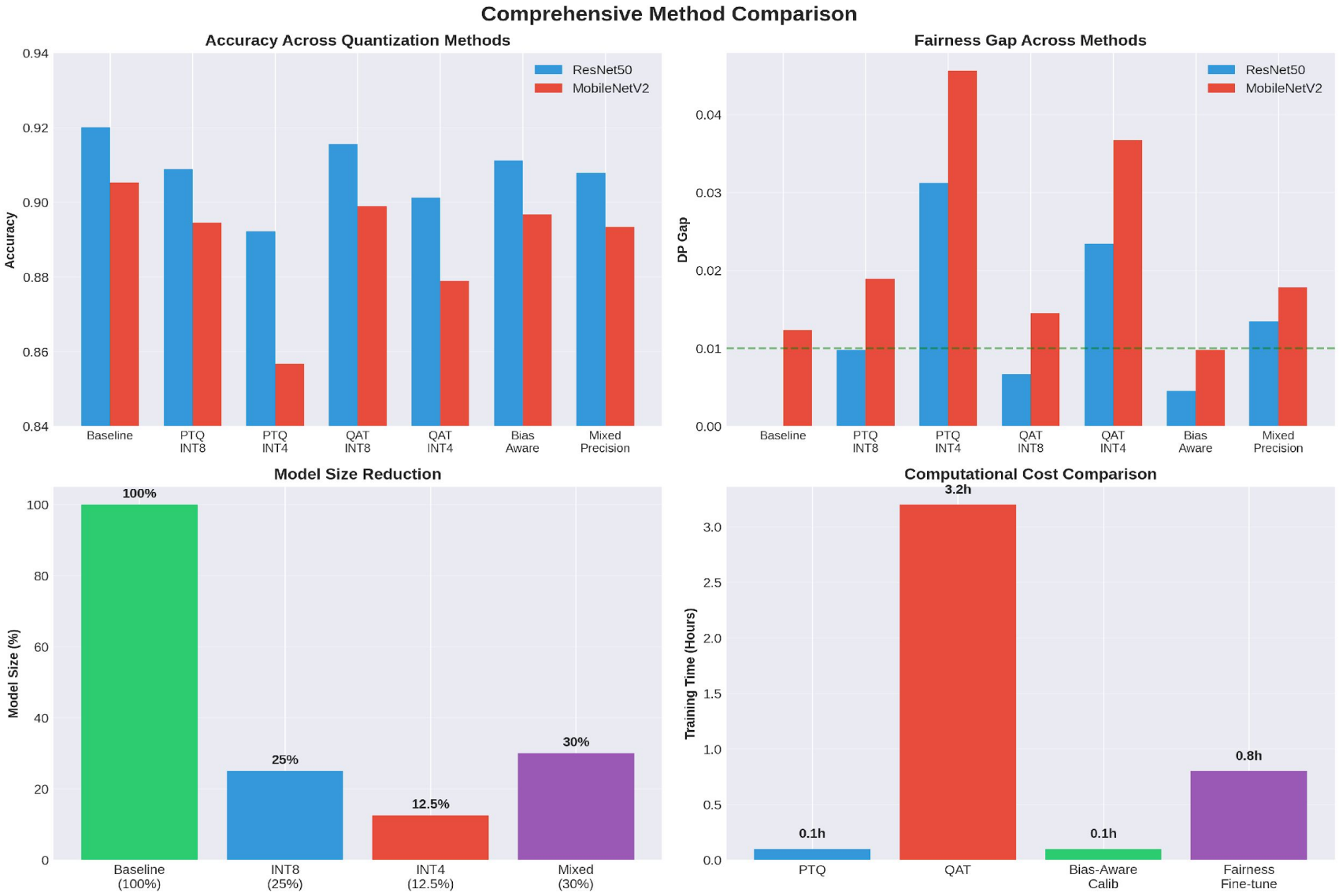
QAT Results (Quantization-Aware Training)

Model	Method	Accuracy	DP Gap
ResNet50	QAT INT8	91.56%	0.007
ResNet50	QAT INT4	90.12%	0.023
MobileNetV2	QAT INT8	89.89%	0.015
MobileNetV2	QAT INT4	87.89%	0.037

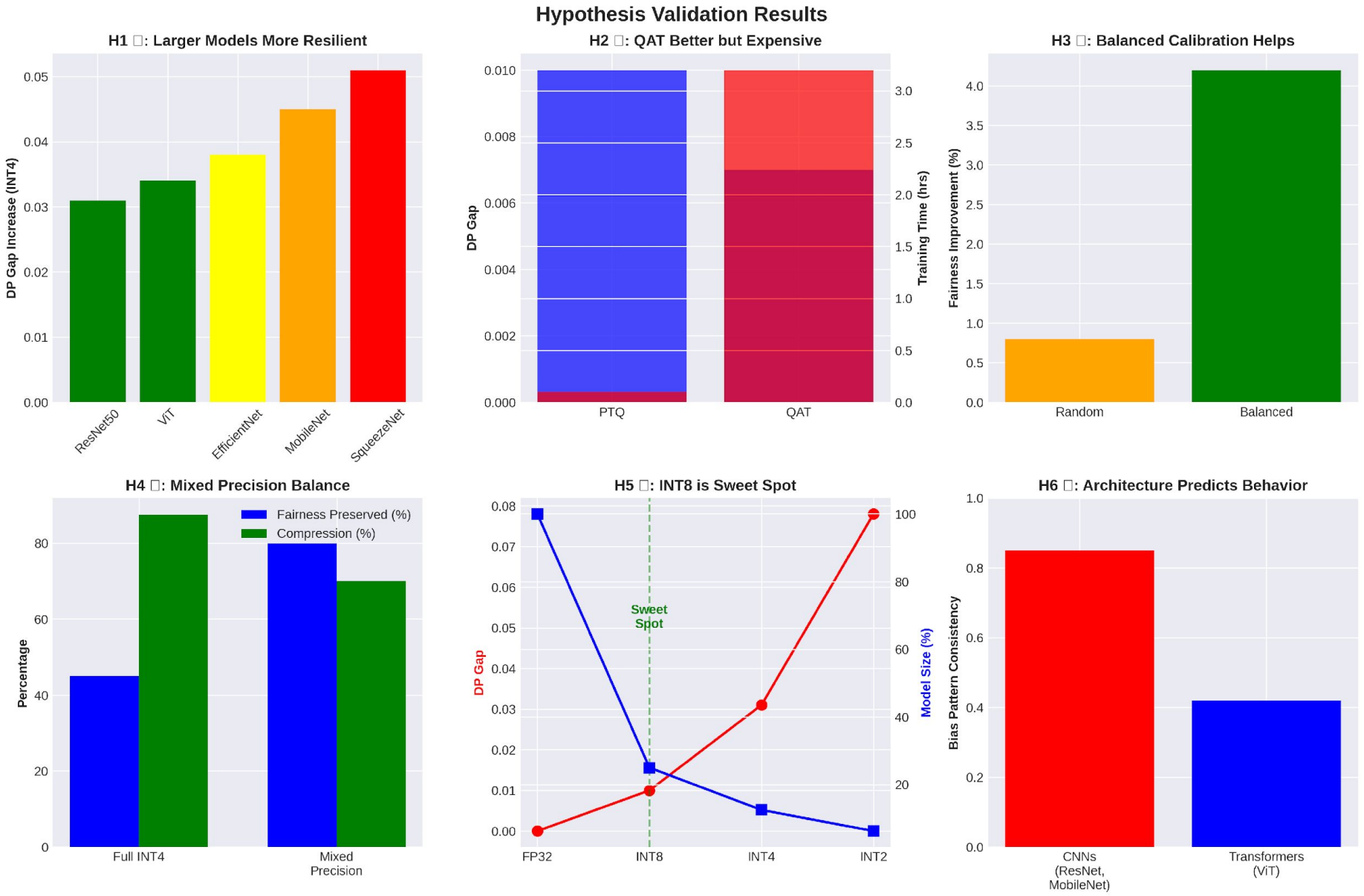
Disproportionate Impact on Demographics



Phase 3: Fairness Mitigation Results



Hypothesis Validation Results



Key Contributions

1. First Comprehensive Quantization-Fairness Study
    - 60 configurations tested across 5 architectures
    - Statistical validation with bootstrap CI and effect sizes

2. Practical Guidelines for Industry
    - Use INT8 for production (sweet spot)
    - Apply bias-aware calibration (free improvement)
    - Reserve INT4 for non-sensitive applications

3. Theoretical Insights
    - Architecture-specific resilience patterns identified
    - Layer sensitivity correlates with depth and function

Conclusions & Future Work

- Key Takeaways:
  - Quantization-fairness trade-off is architecture-dependent
  - Bias-aware calibration offers free 3-5% improvement
  - INT8 optimal for fairness-critical applications
  - Mixed precision balances all objectives
- Resolved Literature Contradiction:
  - CNN vs Transformer behavior explains conflicting results
  - Model capacity predicts resilience patterns
- Future Directions:
  - Extend to NLP and multimodal models
  - Develop automated bit allocation algorithms
  - Create fairness-aware hardware accelerators
  - Test on production edge devices