Progress Evaluation Report: EfficientNet-EdgeTPU with Extreme Quantization for Real-Time Inference

1. Introduction

1.1 Problem Statement

While EfficientNetV2 demonstrates significant improvements in training efficiency and model performance, its deployment in real-time inference scenarios remains challenging due to computational and memory constraints. Current quantization approaches primarily focus on standard INT8 precision, leaving substantial optimization potential unexplored in extreme quantization regimes (INT4, INT2, mixed-precision). The challenge lies in achieving aggressive model compression through extreme quantization while maintaining competitive accuracy for real-time inference applications.

1.2 Research Objectives

This project aims to:

- **Implement extreme quantization schemes** (INT8, INT4, INT2, mixed-precision) for EfficientNetV2 architectures
- **Achieve real-time inference** with latency on standard hardware accelerators
- **Maintain accuracy** above 70% ImageNet Top-1 while achieving >5x compression ratio
- **Compare performance** against EfficientNet-EdgeTPU

1.3 Motivation

Real-time AI inference is critical for applications requiring immediate response times, such as autonomous systems, industrial automation, and interactive AI applications. While EfficientNetV2 provides architectural improvements for training efficiency, its potential for extreme quantization and real-time deployment remains underexplored. This research addresses the gap between model efficiency improvements and practical deployment under severe computational constraints.

1.4 Research Gap

Current literature lacks comprehensive investigation of:

- **Extreme quantization** (sub-8-bit) specifically applied to EfficientNetV2 architectures
- **Layer-wise quantization sensitivity** analysis for EfficientNetV2's unique components (Fused-MBConv, progressive training benefits)
- **Hardware-agnostic quantization** schemes that optimize for inference speed across diverse accelerators

1.5 Challenges

Key technical challenges include:

- **Accuracy preservation** during aggressive quantization (INT4/INT2)
- **Quantization scheme selection** for different EfficientNetV2 components
- **Training stability** when applying extreme quantization with limited computational resources
- **Hardware optimization** ensuring quantization benefits translate to real speedups
- **Gradient flow preservation** during quantization-aware training with reduced precision

2. Literature Review

2.1 Baseline Model – EfficientNet-EdgeTPU

EfficientNet-EdgeTPU represents the current state-of-the-art for efficient inference, employing standard INT8 quantization and EdgeTPU-specific optimizations [1].
Key characteristics:

- **Quantization**: Uniform INT8 across all layers [2].
- **ImageNet accuracy**: 79.1% Top-1 with full precision, ~77.5% with INT8 [1].
- **Inference speed**: 3.0 ms on EdgeTPU, 45 ms on mobile CPU .
- **Compression ratio**: ~4× from FP32 baseline [3].
- **Architecture**: Standard EfficientNet-B0 with EdgeTPU optimizations.

The baseline demonstrates effective INT8 quantization but leaves room for more aggressive compression schemes and broader hardware compatibility [3].

2.2 EfficientNetV2 Improvements for Quantization

EfficientNetV2 introduces several architectural and training features that can improve quantization robustness [1]:

- **Fused-MBConv Blocks**: Reduce memory access patterns through operator fusion, leading to fewer intermediate activations that require high precision .
- **Progressive Training**: Adaptive image resolution training may enhance robustness under quantization by exposing models to variable computational loads.
- **Improved Training Efficiency**: Faster convergence enables more extensive quantization-aware training experiments within limited budgets .
- **Better Parameter Efficiency**: Reduced redundancy in parameters suggests improved tolerance to precision reduction in critical layers.

2.3 Extreme Quantization Opportunities

2.3.1 INT4 and Mixed-Precision Quantization

Recent advances in extreme quantization have explored sub-8-bit precision:

- **Uniform INT4**: All weights and activations quantized to 4-bit precision [4].
- **Mixed-precision schemes**: Critical layers (first/last) maintained at INT8, while intermediate layers use INT4 [5].
- **Asymmetric quantization**: Different precision for weights vs. activations [6].
- **Block-wise quantization**: Different precision for architectural blocks [5].

2.3.2 Quantization-Aware Training Methodologies

Advanced training techniques can further enhance extreme quantization performance:

- **Knowledge distillation**: A full-precision teacher guides a quantized student [6].
- **Progressive quantization**: Gradually reducing precision during training [3].
- **Learnable quantization**: Quantization parameters optimized as part of training [3].
- **Regularization techniques**: Preventing overfitting in the quantized parameter space [2].

2.3.3 Hardware-Aware Quantization

Different inference hardware architectures demand specific optimization strategies:

- **CPU optimization**: INT8/INT4 SIMD instruction utilization [5].
- **GPU acceleration**: Tensor Core mixed-precision support [7].
- **Custom accelerators**: EdgeTPU, NPUs optimized for low-precision inference.
- **FPGA deployment**: Flexible arithmetic units enabling variable precision [7].

2.4 Quantization Impact Analysis

Layer-wise sensitivity analysis highlights varying robustness levels across network components :

- **Depthwise convolutions**: Generally more tolerant to quantization [6].
- **Pointwise expansions**: Require higher precision to preserve information [4].
- **Squeeze-and-excitation modules**: Attention mechanisms are potentially sensitive to precision reduction [6].
- **Classification heads**: Final layers typically need higher precision for accuracy [3].

2.5 Summary

EfficientNetV2's architectural improvements provide a strong foundation for extreme quantization exploration. The combination of improved training efficiency and architectural optimizations creates opportunities for more aggressive compression than previously achievable with standard EfficientNet architectures.

3. Methodology Outline

3.1 Baseline Analysis and Quantization Profiling

The initial phase focuses on comprehensive analysis of EfficientNetV2 architectures and their quantization characteristics. Full-precision EfficientNetV2 models (S, M, L variants) will be trained and profiled to establish accuracy and computational baselines across different model scales.

Layer-wise quantization sensitivity analysis will be conducted by systematically applying different precision levels to individual components and measuring accuracy degradation. This analysis will identify quantization-tolerant layers (candidates for aggressive INT4/INT2 quantization) and precision-critical layers (requiring INT8 or higher precision).

Computational profiling will utilize hardware-specific tools to measure operator-level performance characteristics, identifying bottlenecks where quantization provides maximum speedup benefits. Memory access pattern analysis will guide quantization scheme selection to optimize both computation and memory bandwidth utilization.

3.2 Targeted Extreme Quantization Strategies

Following the sensitivity analysis, the project will implement progressive quantization schemes with increasing levels of compression, guided by EfficientNetV2's architectural characteristics and training improvements.

3.2.1 Architectural-Aware Quantization Schemes

Quantization strategies will be tailored to EfficientNetV2's specific architectural components. Fused-MBConv blocks will be analyzed for optimal precision allocation, leveraging their reduced memory access patterns to enable more aggressive quantization of intermediate

activations. The integrated nature of these blocks allows for fine-grained precision control at the sub-block level.

Mixed-precision quantization schemes will be developed based on architectural hierarchy: depthwise separable convolutions (candidates for INT4), pointwise expansions (INT8 for information preservation), and squeeze-and-excitation attention modules (INT8/FP16 for gradient stability). Progressive quantization will start with conservative mixed-precision allocation and systematically increase compression ratios.

Block-wise quantization strategies will exploit EfficientNetV2's compound scaling properties, applying different precision levels to early feature extraction layers (aggressive quantization acceptable) versus later semantic layers (precision preservation critical). Learnable quantization parameters will be integrated to automatically optimize bit allocation during training.

3.2.2 Quantization-Aware Training with EfficientNetV2 Improvements

Enhanced quantization-aware training will leverage EfficientNetV2's training efficiency improvements for extensive hyperparameter exploration. Progressive training strategies will be adapted for quantization: starting with higher precision and gradually reducing bit-widths while maintaining training stability through adaptive learning rate scheduling.

Knowledge distillation will employ full-precision EfficientNetV2 models as teachers for extremely quantized student networks. Multi-level distillation will match both intermediate feature representations and final predictions, with particular attention to preserving attention patterns in squeeze-and-excitation modules under quantization constraints.

Advanced regularization techniques will include quantization noise injection during training, simulating hardware quantization effects to improve robustness. Gradient clipping and normalization will be adapted for quantized parameter spaces, preventing gradient explosion in low-precision training regimes.

3.3 Incremental Quantization Experiments

Development will proceed through systematic quantization progression, with each compression level thoroughly validated before advancing to more aggressive schemes. Initial experiments will establish INT8 baselines across all EfficientNetV2 variants, replicating and improving upon EfficientNet-EdgeTPU quantization results.

Progressive compression experiments will systematically reduce precision: uniform INT8 → mixed INT8/INT4 → aggressive INT4 → experimental INT2/INT4 schemes. Each quantization level will be evaluated independently and in combination, enabling fine-grained attribution of accuracy-speed trade-offs to specific quantization decisions.

Hardware-specific optimization will target multiple inference platforms simultaneously, ensuring quantization benefits translate across CPU (Intel, ARM), GPU (NVIDIA, AMD), and specialized accelerators (EdgeTPU, Neural Processing Units). Platform-specific quantization schemes will be developed to maximize hardware utilization efficiency.

3.4 Comprehensive Real-Time Inference Evaluation

The final evaluation phase will demonstrate real-time inference capabilities through comprehensive benchmarking across multiple dimensions and hardware platforms. Performance evaluation will prioritize actual inference speedup over theoretical compression ratios.

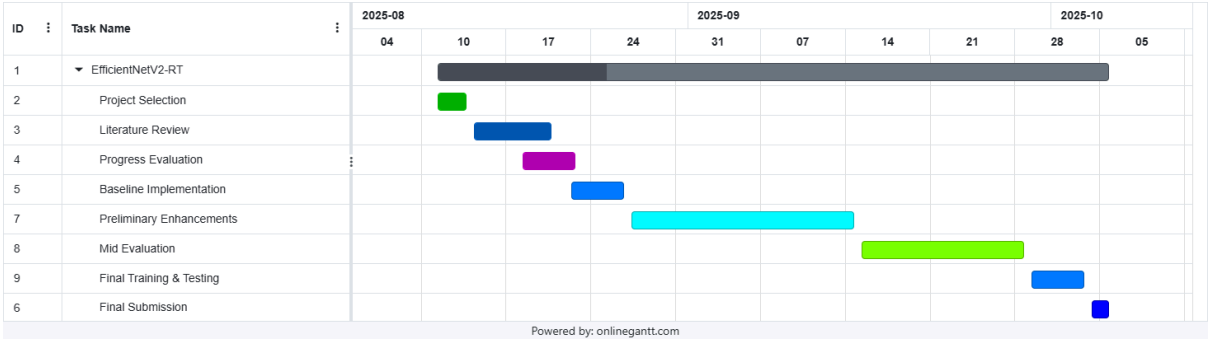
Inference Speed Metrics: Throughput optimization for batch processing scenarios and sustained inference workloads. Memory bandwidth utilization efficiency under different quantization schemes.

Accuracy Preservation Analysis: ImageNet classification accuracy across different quantization levels, with detailed analysis of accuracy-speed Pareto frontiers. Robustness evaluation on challenging samples and out-of-distribution data to verify quantization generalization.

Comparative Baseline Evaluation: Direct comparison with EfficientNet-EdgeTPU under equivalent quantization constraints, demonstrating advantages of EfficientNetV2 architectural improvements for extreme quantization scenarios.

All quantization schemes will be evaluated through comprehensive ablation studies, quantifying the individual and combined contributions of different compression techniques. Failed quantization approaches will be documented with analysis of failure modes, providing insights for future quantization research. The final optimized models will achieve the target real-time inference performance while maintaining competitive accuracy levels.

4. Project Timeline



5. Conclusion

This project addresses the critical challenge of achieving real-time inference through extreme quantization of EfficientNetV2 models. By leveraging the architectural improvements and training efficiencies introduced in the EfficientNetV2 paper, the project aims to push the boundaries of model compression while maintaining practical accuracy levels.

The systematic approach to extreme quantization, combined with hardware-aware optimization and comprehensive evaluation, will provide valuable insights for the broader

edge AI community. Success will demonstrate that aggressive quantization can achieve real-time inference requirements while preserving sufficient accuracy for practical deployment scenarios.

The research contributes to the fundamental understanding of quantization limits in modern efficient architectures and provides practical tools for deploying highly compressed models in resource-constrained environments.

6. References

[1] M. Tan and Q. V. Le, "EfficientNetV2: Smaller Models and Faster Training," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 139, pp. 10096–10106, Jul. 2021. https://arxiv.org/abs/2104.00298

[2] B. Jacob *et al.*, "Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference," presented at *CVPR 2018*, 2018. [1712.05877] Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference

[3] M. Nagel, M. Fournarakis, R. A. Amjad, Y. Bondarenko, M. van Baalen, and T. Blankevoort, "A White Paper on Neural Network Quantization," *arXiv*, preprint arXiv:2106.08295, Jun. 2021.(PDF) A White Paper on Neural Network Quantization

[4] R. Banner *et al.*, "Post-Training 4-bit Quantization of CNNs via a Frequency-Domain Approach," presented at *ICLR 2019*, 2019.[1810.05723] Post-training 4-bit quantization of convolution networks for rapid-deployment

[5] K. Wang *et al.*, "HAQ: Hardware-Aware Automated Quantization," presented at *CVPR 2019*, 2019.[1811.08886] HAQ: Hardware-Aware Automated Quantization with Mixed Precision

[6] B. Zhuang *et al.*, "Towards Effective Low-bitwidth Convolutional Neural Networks," presented at *CVPR 2018*, 2018.[1711.00205] Towards Effective Low-bitwidth Convolutional Neural Networks

[7] Z. Liu *et al.*, "Post-Training Quantization for Vision Transformer," presented at *NeurIPS 2021*, 2021.[2106.14156] Post-Training Quantization for Vision Transformer

[8] TensorFlow Quantization Guide, TensorFlow Model Optimization Toolkit Documentation, 2023.Post-training quantization | TensorFlow Model Optimization