Efficient Computer Vision: Pushing the Frontier of Edge Computing

Michael Krah, Zach Gentile, and Alex Lavaee

March 2, 2025

Abstract

The Low Power Computer Vision Challenge (LPCVC) advances three critical directions in computer vision by focusing on model optimization for edge devices. Leveraging the Qualcomm AI-Hub ecosystem, this competition enables developers to deploy efficient vision models on mobile phones and AI PCs. Participants can submit models in various formats (PyTorch, TensorFlow, TFLite, ONNX), making the challenge globally accessible. Unlike traditional competitions that rely on cloud computing, LPCVC emphasizes practical applications that run directly on edge devices with minimal hardware requirements. This approach democratizes participation for developers, researchers, and students worldwide. The challenge provides open-source sample solutions as qualification benchmarks, with winning solutions and test datasets being publicly released to foster continued innovation in efficient computer vision. In this paper, we propose a solution to the LPCVC challenge that uses a combination of model architecture improvements such as quantization, pruning, token-efficient methods. In addition, we use enhancements to the training pipeline including 8-bit General Matrix Multiplication (GEMM) and Multi-head Latent Attention (MLA). Finally, we propose techniques for edge device optimization. Our solution is designed to be lightweight, efficient, and run directly on edge devices with minimal hardware requirements.

Introduction

Computer vision at the edge has emerged as a critical frontier in artificial intelligence, enabling real-time image analysis on resource-constrained devices without dependence on cloud connectivity. The 2025 IEEE Low-Power Computer Vision Challenge (LPCVC) - Track 1 addresses a fundamental challenge in this domain: developing accurate image classification models that perform well across varying lighting conditions and styles while maintaining high efficiency on edge devices.

This challenge is particularly significant as lighting variations remain one of the most common obstacles to robust computer vision deployment in real-world scenarios. Similarly, the ability to recognize objects across different artistic styles enables broader applications in mixed reality, content moderation, and accessibility tools. The competition leverages Qualcomm's AI-Hub ecosystem to deploy and evaluate models on Snapdragon 8 Elite hardware, providing a standardized benchmark for edge AI performance.

Our proposed approach focuses on optimizing neural network architectures specifically for edge deployment through a combination of model compression techniques, quantization-aware training, and hardware-specific acceleration. We aim to balance the trade-off between classification accuracy and inference speed, with a particular emphasis on maintaining robustness across lighting variations and style transfers. The core innovation of our work lies in developing specialized optimization techniques that preserve discriminative features critical for identifying the 64 target classes under challenging visual conditions.

By participating in this challenge, we seek to advance the state-of-the-art in efficient computer vision and contribute methodologies that can benefit researchers and developers working with limited computational resources across various domains including augmented reality, autonomous systems, and smart devices.

Related Work

Efficient computer vision for edge devices has seen significant advancements across several key areas. We organize our literature review around three main themes particularly relevant to the LPCVC challenge:

Model Compression and Quantization

Neural network compression has become crucial for deploying deep learning models on resource-constrained devices. Work by Han et al. [**han2016deep**] introduced deep compression, combining pruning, quantization, and Huffman coding to significantly reduce model size while maintaining accuracy. Building on this foundation, Jacob et al. [**jacob2018quantization**] developed quantization-aware training techniques that allow for 8-bit integer operations without significant accuracy degradation. More recently, Nagel et al. [**nagel2021white**] demonstrated data-free quantization methods that enable post-training conversion to lower precision formats, making deployment more accessible.

Architectures for Edge Deployment

Specialized neural network architectures designed explicitly for edge devices have shown remarkable efficiency. MobileNetV2 [**sandler2018mobilenetv2**] and MobileNetV3 [**howard2019searching**] introduced inverted residuals and linear bottlenecks that dramatically reduced computational costs. EfficientNet [**tan2019efficientnet**] proposed a principled scaling method that optimizes depth, width, and resolution dimensions simultaneously. Most relevant to our work, Lin et al. [**lin2021mcunet**] developed MCUNet for ultra-small microcontrollers, demonstrating techniques to adapt neural networks to extremely limited hardware.

Robustness to Lighting and Style Variations

Addressing variations in lighting conditions has been a longstanding challenge in computer vision. Histogram equalization and other traditional techniques have been supplemented by learning-based approaches as described by Lore et al. [**lore2017llnet**] who developed LLNet specifically for low-light image enhancement. For style robustness, Geirhos et al. [**geirhos2019imagenet**] demonstrated that CNNs are biased toward texture rather than shape and proposed methods to improve generalization across styles. Most recently, Li et al. [**li2022domain**] introduced adaptive normalization strategies that significantly improve performance across domain shifts including lighting and style variations.

Proposed Work

Our solution to the 2025 LPCVC Track 1 challenge builds on state-of-the-art techniques in efficient deep learning and leverages insights from recent low-power computer vision studies [2]. Our approach combines algorithmic innovations with hardware-specific optimizations to achieve superior accuracy and speed on the Snapdragon 8 Elite.

Model Architecture and Optimization

We begin with a lightweight backbone based on EfficientNet-B0, enhanced with attention mechanisms inspired by transformer architectures [4, 9] and efficient mobile designs [8]. Our approach includes:

- **Mixed-Precision Training:** We implement dynamic quantization that maintains full precision for critical parameters while quantizing less sensitive components. This approach is guided by prior work on efficient integer-arithmetic-only inference [6].

- **8-bit Matrix Multiplication:** Our model utilizes 8-bit integer matrix multiplication (Int8 MatMul) with calibrated scaling factors to minimize quantization error, reducing memory and computational complexity by approximately 75% compared to FP32 operations [6].

- **Structured Pruning:** To ensure hardware-friendly operations, we apply structured pruning that removes entire channels or filters. Our strategy is informed by recent frameworks for automatic DNN compression [7] and discrimination-aware channel pruning techniques [11].

- **Knowledge Distillation:** We train a large teacher model on cloud resources and distill its knowledge into a compact student model. This technique builds on seminal work in knowledge distillation [5] and recent advances in data-efficient transformer training [9].

- **Token-Efficient Vision Transformers:** We incorporate token-efficient strategies to further reduce computational overhead. Inspired by [10], our design represents images using only 32 tokens, drastically reducing the input dimensionality. In addition, we integrate token merging techniques as detailed in [1], which dynamically merge similar tokens during processing, maintaining performance while improving efficiency.

Hardware-Specific Optimizations

To fully leverage the Snapdragon 8 Elite's capabilities, we propose several hardware-specific enhancements:

- **Custom Kernels for the Qualcomm AI Engine:** We will develop custom convolutional kernels optimized for the Hexagon DSP and Adreno GPU. This work aligns with recent technical advancements in integrated hardware-software optimization [3].
- **GEMM and Multi-Head Latent Attention (MLA):** In addition to custom kernels, we integrate high-performance General Matrix Multiplication (GEMM) routines alongside novel Multi-head Latent Attention (MLA) mechanisms, as introduced in [3]. GEMM optimizations streamline batched computations, while MLA enhances the attention mechanism by leveraging latent representations, thus reducing computational overhead and improving feature extraction.
- **Adaptive Computation:** Early-exit mechanisms will be implemented to dynamically adjust computational resources based on input complexity, ensuring efficient processing for ”easy” samples.
- **Memory-Efficient Feature Extraction:** We optimize convolutional layers with in-place operations and specialized memory management techniques to minimize the activation footprint.
- **Sparse Computation:** Our design incorporates activation sparsity-aware computations that skip multiplications with zero or near-zero values, further reducing computational load.

Robustness to Lighting and Style Variations

Given the challenge's focus on varying lighting conditions and artistic styles, our approach emphasizes robust feature learning:

- **Illumination-Invariant Feature Learning:** We integrate a specialized preprocessing module that normalizes lighting conditions while preserving key discriminative features. This module leverages both token-efficient representations and robust feature merging techniques.

- **Data Augmentation Strategy:** A comprehensive augmentation pipeline will simulate diverse lighting conditions and artistic styles during training, enhancing the model's generalization.
- **Multi-Task Learning:** By jointly learning to classify objects and estimate lighting conditions, our network develops more robust internal representations.
- **Domain Adaptation Techniques:** Adversarial domain adaptation components are incorporated to minimize feature distribution shifts between different lighting and style domains.

Our initial experiments suggest that the combination of these approaches can achieve competitive accuracy while maintaining inference times well below the 10ms threshold specified by the challenge [2]. The final model is expected to demonstrate a favorable accuracy-to-efficiency ratio as measured by the competition metric:

$$\text{Score} = \frac{\text{Accuracy}}{\max(\text{ExecutionTime}/2, 1\text{ms})}. \tag{1}$$

This design not only meets the stringent requirements of the LPCVC challenge but also aligns with recent advancements in efficient deep learning and hardware-aware model design [3].

Datasets

Our project leverages the dataset provided by the 2025 IEEE Low-Power Computer Vision Challenge (LPCVC) Track 1, which consists of a carefully curated subset of images from the COCO (Common Objects in Context) dataset. The competition dataset has the following characteristics:

Dataset Composition

The dataset focuses on 64 object classes from the original 80 COCO detection classes. These classes span a diverse range of objects including vehicles (bicycles, cars, buses), animals (birds, cats, dogs), household items (bottle, wine glass, cup), and everyday objects (clock, laptop, cell phone). A comprehensive list of all 64 classes is provided in the challenge documentation.

Lighting and Style Variations

A key characteristic of this dataset is the deliberate inclusion of lighting variations for each class. The images encompass:

- **Natural Lighting Conditions:** Daylight, twilight, nighttime, indoor lighting
- **Challenging Lighting Scenarios:** Backlighting, low-light, high-contrast, and oversaturated conditions
- **Style Variations:** Some images have been generated using Stable Diffusion to create artistic style variations of the same objects

This diverse collection of lighting conditions and styles makes the dataset particularly challenging, as models must learn lighting-invariant and style-invariant features to achieve high classification accuracy.

Dataset Format and Preprocessing

The images in the dataset are standardized to RGB format with a resolution of 224×224 pixels. Input tensors are expected to have a shape of (batch, 3, 224, 224) with pixel values normalized to the range [0, 1] in float format. Our preprocessing pipeline will include:

- Data normalization using ImageNet mean and standard deviation values
- Random cropping, horizontal flipping, and color jittering for data augmentation
- Targeted augmentation techniques to simulate additional lighting conditions

Dataset Split and Evaluation

While the competition provides a sample dataset for development, the final evaluation will be conducted on a held-out test set. For our development process, we will:

- Use the official sample dataset as our primary training set
- Create a stratified validation split to ensure representative class distribution
- Augment the training data with additional COCO images and style-transferred variants where appropriate
- Evaluate on the competition's hidden test set for final submission

The diversity of lighting conditions and styles in this dataset aligns perfectly with our proposed methods for illumination-invariant feature learning and domain adaptation, allowing us to thoroughly evaluate and optimize our approach for robust performance across visual variations.

Evaluation

Our evaluation methodology follows the official metrics for the 2025 LPCVC Track 1 challenge while incorporating additional analyses to guide our development process:

Primary Metrics

In accordance with the challenge requirements, our primary evaluation metrics are:

- **Execution Time:** Our solution must execute within 10ms per image on the Snapdragon 8 Elite QRD platform to be considered valid. We will continuously monitor inference latency during development.
- **Classification Accuracy:** The percentage of correctly classified images across all 64 classes in the test dataset, with special attention to performance across different lighting conditions and styles.
- **Final Score:** The competition's official scoring formula of accuracy, as described in equation 1, which balances accuracy against computational efficiency.

Development Evaluation Strategy

During the development process, we will implement a more comprehensive evaluation approach:

- **Validation Set Evaluation:** We will measure performance on the validation set representing different lighting conditions and styles to identify potential weaknesses.
- **Confusion Matrix Analysis:** We will analyze the confusion matrix to identify frequently misclassified class pairs and target these specifically in our model refinement.
- **Hardware Performance Profiling:** Using Qualcomm AI Hub tools, we will profile memory usage, CPU/GPU/DSP utilization, and power consumption to guide our optimization efforts.

Baseline Comparisons

We will establish the following baselines for comparison:

- **Competition Sample Solution:** The performance of the provided sample solution will serve as our minimum benchmark.
- **Standard MobileNetV2:** We will evaluate a standard MobileNetV2 model to understand the improvements offered by our specialized approach.

- **EfficientNet-B0:** As we build upon this architecture, we will track improvements over the original implementation.

Timeline

We will structure our project with the following timeline:

Phase 1: Research and Exploration (Week 1, March 2-9)

- Literature review of efficient vision models and techniques for lighting/style invariance
- Setup development environment with Qualcomm AI Hub and test sample solution

Phase 2: Development and Optimization (Weeks 2-5, March 10-31)

- Implementation of core model architecture (identify base model to distill, e.g., DINOv2, I-JEPA)
- Exploring model hyperparameters, lower-token models, distillation techniques, quantization, etc.
- Initial hardware-specific optimizations and performance profiling
- Comprehensive evaluation and performance tuning
- Submission to IEEE Lower-Power Computer Vision Challenge

Phase 3: Refinement and Submission (Week 6-10, April 1-30)

- Continue refining algorithm, model, and hyperparameters
- Final model benchmarking and validation
- Documentation preparation and code cleanup
- Preparation of final project paper

Key Milestones

- March 10: Finish literature review and setup development environment
- March 17: Implement baseline, core model architecture, and develop initial codebase
- March 31: Submit working version of model to IEEE Lower-Power Computer Vision Challenge
- April 15: Finalize models and prepare paper and presentation

- May 1: Submit final code, paper, and presentation

Conclusion

This proposal outlines our approach to the 2025 IEEE Low-Power Computer Vision Challenge - Track 1, focusing on efficient image classification across diverse lighting conditions and styles. We have proposed a comprehensive solution that combines state-of-the-art model compression techniques, hardware-specific optimizations for the Snapdragon 8 Elite platform, and specialized methods for enhancing robustness to lighting and style variations.

Our strategy leverages 8-bit matrix multiplication, structured pruning, knowledge distillation, and custom optimized kernels to achieve both high accuracy and low inference latency for improving the accessibility of Deep Learning models.

References

[1] Daniel Bolya et al. *Token Merging: Your ViT But Faster*. 2023. arXiv: `2210.09461 [cs.CV]`. URL: `https://arxiv.org/abs/2210.09461`.

[2] Leo Chen et al. *2023 Low-Power Computer Vision Challenge (LPCVC) Summary*. 2024. arXiv: `2403.07153 [cs.CV]`. URL: `https://arxiv.org/abs/2403.07153`.

[3] DeepSeek-AI et al. *DeepSeek-V3 Technical Report*. 2025. arXiv: `2412.19437 [cs.CL]`. URL: `https://arxiv.org/abs/2412.19437`.

[4] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. arXiv: `2010.11929 [cs.CV]`. URL: `https://arxiv.org/abs/2010.11929`.

[5] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. *Distilling the Knowledge in a Neural Network*. 2015. arXiv: `1503.02531 [stat.ML]`. URL: `https://arxiv.org/abs/1503.02531`.

[6] Benoit Jacob et al. *Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference*. 2017. arXiv: `1712.05877 [cs.LG]`. URL: `https://arxiv.org/abs/1712.05877`.

[7] Ning Liu et al. *AutoCompress: An Automatic DNN Structured Pruning Framework for Ultra-High Compression Rates*. 2019. arXiv: `1907.03141 [cs.LG]`. URL: `https://arxiv.org/abs/1907.03141`.

[8] Junting Pan et al. *EdgeViTs: Competing Light-weight CNNs on Mobile Devices with Vision Transformers*. 2022. arXiv: `2205.03436 [cs.CV]`. URL: `https://arxiv.org/abs/2205.03436`.

[9] Hugo Touvron et al. *Training data-efficient image transformers and distillation through attention*. 2021. arXiv: `2012.12877 [cs.CV]`. URL: `https://arxiv.org/abs/2012.12877`.

[10] Qihang Yu et al. *An Image is Worth 32 Tokens for Reconstruction and Generation*. 2024. arXiv: 2406.07550 [cs.CV]. URL: https://arxiv.org/abs/2406.07550.

[11] Zhuangwei Zhuang et al. *Discrimination-aware Channel Pruning for Deep Neural Networks*. 2019. arXiv: 1810.11809 [cs.CV]. URL: https://arxiv.org/abs/1810.11809.