Energy-Aware Quantization and Mixed-Precision Learning for Vision Models

Salma Bhar Wiam Skakri Krupa Venkatesan

{sxb1283, wxs428, kxv178}@case.edu

December 10, 2025

Abstract

As deep learning models scale in size and complexity, their energy consumption during inference has become a critical concern for deployment on edge devices and large-scale data centers. In this work, we investigate post-training quantization (PTQ) and sensitivity-based mixed-precision inference as lightweight approaches for improving the energy efficiency of vision models. We evaluate our methods on ResNet-18 (CIFAR-100) and DeiT-Tiny (ImageNet), applying uniform quantization at 8-bit, 6-bit, and 4-bit precisions, as well as a mixed-precision scheme that assigns bitwidths based on layer-wise sensitivity analysis. For Vision Transformers, we also implement and evaluate ranking-aware quantization that preserves attention score orderings. Using an analytical energy model that accounts for both MAC operations and memory access, we demonstrate energy reductions of up to 88% at 4-bit precision. Our mixed-precision configuration achieves 85.7% energy reduction on ResNet-18 while maintaining 99% of baseline accuracy. For DeiT-Tiny, ranking-aware PTQ provides +0.08-0.10% accuracy improvements at 8/6-bit precision, and our layer sensitivity analysis reveals that MLP layers are 4.7× more sensitive than attention projections. We observe that DeiT-Tiny exhibits remarkable robustness to quantization, maintaining 85.01% accuracy at 6-bit precision with only a 0.33 percentage point drop from FP32. Our findings highlight the potential of training-free quantization approaches for energy-efficient deployment, while also revealing the limitations of aggressive quantization without additional optimization techniques.

1 Introduction

The rapid advancement of deep learning has led to increasingly large and powerful models, particularly in computer vision. Vision Transformers (ViTs) Dosovitskiy et al. (2020) and their efficient variants like DeiT Touvron et al. (2021) have achieved state-of-the-art performance on image classification tasks, but at the cost of significant computational and memory requirements. This poses a fundamental challenge for deploying these models on resource-constrained edge devices (mobile phones, IoT sensors, autonomous vehicles) and in large-scale inference services where energy consumption directly impacts operational costs and environmental footprint.

The total energy consumption of neural network inference can be decomposed into two primary components Horowitz (2014):

$$E = N_{\text{MAC}} \cdot E_{\text{MAC}} + N_{\text{mem}} \cdot E_{\text{DRAM}} \tag{1}$$

where N_{MAC} is the number of multiply-accumulate operations, E_{MAC} is the energy per MAC, N_{mem} is the number of memory accesses, and E_{DRAM} is the energy per memory access. Critically, memory access dominates energy consumption in modern systems, a 32-bit DRAM access consumes approximately 200× more energy than a 32-bit floating-point multiply Horowitz (2014).

Quantization addresses this challenge by reducing the numerical precision of weights and activations, which decreases both memory footprint and computational cost. Post-training quantization (PTQ) is particularly attractive as it requires no retraining, only a small calibration dataset is needed to determine quantization parameters.

Our Contributions:

1. We implement and evaluate PTQ at multiple precision levels (8/6/4-bit) for both CNN (ResNet-18) and Transformer (DeiT-Tiny) architectures.
2. We develop a sensitivity-based mixed-precision assignment strategy that allocates bitwidths according to layer-wise quantization sensitivity.
3. We provide an analytical energy model to estimate energy savings and construct Pareto frontiers characterizing the accuracy-energy trade-off.
4. We analyze the differential quantization robustness of CNNs versus Vision Transformers, observing that transformers maintain accuracy under aggressive quantization better than CNNs.

2 Related Work

2.1 Post-Training Quantization

Post-training quantization has emerged as a practical approach for model compression without the computational overhead of quantization-aware training (QAT). Early work focused on CNNs, establishing techniques like symmetric and asymmetric quantization, per-tensor and per-channel scaling, and calibration methods for determining optimal quantization ranges Jacob et al. (2018); Krishnamoorthi (2018).

More recently, PTQ methods have been adapted for Transformer architectures. Liu et al. Liu et al. (2021) identified that Vision Transformers pose unique challenges for quantization due to the softmax attention mechanism, where relative rankings of attention scores must be preserved. They introduced ranking-aware quantization objectives and mixed-precision assignment based on feature diversity measured through nuclear norms of attention maps.

2.2 Mixed-Precision Quantization

Mixed-precision quantization assigns different bitwidths to different layers based on their sensitivity to quantization error. HAQ Wang et al. (2019) uses reinforcement learning to search for optimal bitwidth assignments. HAWQ Dong et al. (2019) employs second-order Hessian information to guide precision allocation. More recently, methods like BRECQ Li et al. (2021) and QDrop Wei et al. (2022) have achieved strong results by carefully reconstructing block-wise outputs during calibration.

2.3 Energy-Aware Neural Networks

Spingarn et al. Spingarn et al. (2023) provided key insights into the energy consumption of low-precision neural networks. They observed that (1) signed arithmetic (two's complement) causes significant bit toggling and power consumption, and (2) multiplier power is dominated by the larger input bitwidth. Their work motivates energy-aware design choices beyond simple accuracy-compression trade-offs.

Our work builds upon these foundations, implementing sensitivity-based mixed-precision PTQ with an analytical energy model to guide and evaluate our quantization decisions.

3 Methodology

3.1 Problem Formulation

Given a pretrained model f_θ with full-precision (FP32) weights θ, our goal is to find quantized weights $\hat{\theta}$ that minimize accuracy degradation while maximizing energy efficiency:

$$\min_{\hat{\theta}} \mathcal{L}(\hat{\theta}) \quad \text{s.t.} \quad E(\hat{\theta}) \leq E_{\text{budget}} \tag{2}$$

where $\mathcal{L}$ is the task loss and E is the estimated inference energy.

3.2 Uniform Symmetric Quantization

We employ uniform symmetric quantization, mapping floating-point values to a discrete set of integers. For a tensor x with bitwidth b:

$$\hat{x} = s \cdot \text{clamp}\left(\text{round}\left(\frac{x}{s}\right), -2^{b-1}, 2^{b-1} - 1\right) \tag{3}$$

where the scale factor s is determined during calibration:

$$s = \frac{\max(|x|)}{2^{b-1} - 1} \tag{4}$$

We apply per-channel quantization for weights (separate scale per output channel) and per-tensor quantization for activations, following best practices Krishnamoorthi (2018).

3.3 Sensitivity Analysis

To enable mixed-precision assignment, we measure the quantization sensitivity of each layer l as the L2 distance between full-precision and quantized outputs:

$$S_l = \|y_l^{\text{full}} - y_l^{\text{quant}}\|_2 \tag{5}$$

This metric is computed during a calibration pass using a small subset of training data. Layers with high sensitivity receive more bits to preserve accuracy, while less sensitive layers are more aggressively quantized.

3.4 Mixed-Precision Assignment

Based on the sensitivity analysis, we assign bitwidths using percentile thresholds:

$$b_l = \begin{cases} 8 & \text{if } S_l \geq P_{75}(S) \\ 6 & \text{if } P_{25}(S) \leq S_l < P_{75}(S) \\ 4 & \text{if } S_l < P_{25}(S) \end{cases} \tag{6}$$

where $P_k(S)$ denotes the k-th percentile of sensitivities across all layers. This approach allocates 8-bit precision to the top 25% most sensitive layers, 4-bit to the bottom 25%, and 6-bit to the middle 50%.

3.5 Energy Model

We use an analytical energy model that captures the key factors affecting inference energy consumption:

$$E = \sum_l N_{\text{MAC}}^{(l)} \cdot E_{\text{MAC}}(b_l) + \sum_l N_{\text{mem}}^{(l)} \cdot E_{\text{DRAM}}(b_l) \tag{7}$$

Based on hardware studies Horowitz (2014); Spingarn et al. (2023), we model:

- **MAC energy**: $E_{\text{MAC}}(b) \propto b^2$ as energy scales quadratically with bitwidth due to multiplier area scaling.
- **Memory energy**: $E_{\text{DRAM}}(b) \propto b$ as energy scales linearly with bitwidth due to data transfer volume.

For our baseline configuration, we use a MAC-to-memory energy ratio of 1:200, reflecting the dominance of memory access energy in modern systems.

3.6 Model Architectures

We evaluate two representative architectures:

ResNet-18 He et al. (2016): A convolutional neural network with 11.7M parameters, trained on CIFAR-100. ResNet's residual connections and batch normalization make it amenable to quantization, though the skip connections can amplify quantization errors across layers.

DeiT-Tiny Touvron et al. (2021): A data-efficient Vision Transformer with 5.7M parameters, pretrained on ImageNet. DeiT uses the standard Transformer architecture with self-attention, which introduces unique quantization challenges due to the softmax operation's sensitivity to input magnitude changes.

4 Experiments

4.1 Experimental Setup

Datasets: We use CIFAR-100 (100 classes, 32×32 images) for ResNet-18 and ImageNet-1K (1000 classes, 224×224 images) for DeiT-Tiny. For calibration, we use 256 randomly sampled training images. For evaluation, we use the full CIFAR-100 test set (10,000 images) and a 10,000-image subset of ImageNet validation.

Implementation: All experiments are conducted in PyTorch. We use the `timm` library Wightman (2019) for DeiT-Tiny pretrained weights. Quantization is applied to all linear and convolutional layers, excluding the first and last layers following common practice.

Baselines: We compare against FP32 (full precision), 8-bit uniform PTQ, 6-bit uniform PTQ, and 4-bit uniform PTQ.

4.2 Main Results

Table 1: Quantization results for ResNet-18 on CIFAR-100. Energy is reported relative to FP32 baseline.

Config	**Accuracy (%)**	**Δ Acc.**	**Rel. Energy**	**Energy Savings**
FP32 (baseline)	61.44	—	1.00	—
8-bit PTQ	61.32	-0.12	0.238	76.2%
6-bit PTQ	60.42	-1.02	0.178	82.2%
4-bit PTQ	38.04	-23.40	0.118	88.2%
Mixed (Ours)	**60.80**	**-0.64**	**0.143**	**85.7%**

Table 2: Quantization results for DeiT-Tiny on ImageNet.

Config	**Accuracy (%)**	**Δ Acc.**	**Rel. Energy**	**Energy Savings**
FP32 (baseline)	85.34	—	1.00	—
8-bit PTQ	85.10	-0.24	0.241	75.9%
6-bit PTQ	85.01	-0.33	0.180	82.0%
4-bit PTQ	83.04	-2.30	0.120	88.0%
Mixed (Ours)	**85.05**	**-0.29**	**0.191**	**80.9%**

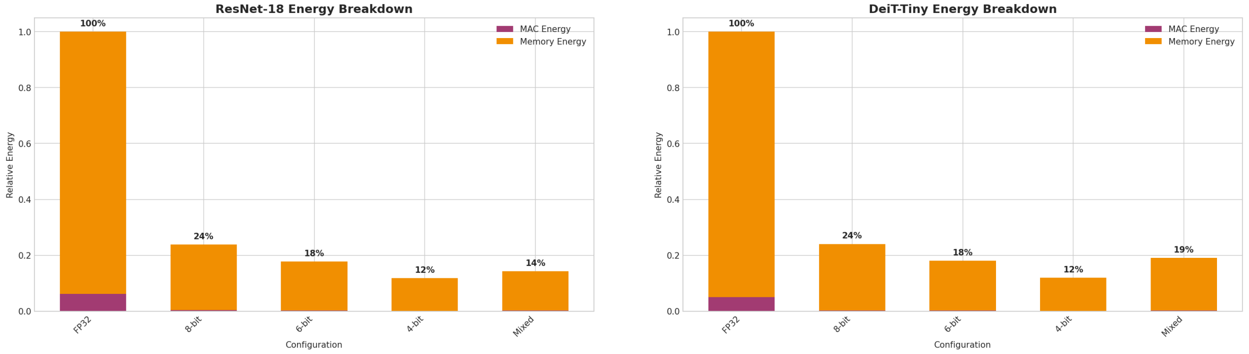
Tables 1 and 2 present our main quantization results. Several key observations emerge:

8-bit quantization is nearly lossless. Both models maintain near-baseline accuracy with 8-bit PTQ (-0.12% for ResNet-18, -0.24% for DeiT-Tiny) while achieving approximately 76% energy reduction.

DeiT-Tiny is remarkably robust to quantization. Even at 4-bit precision, DeiT-Tiny loses only 2.3 percentage points, compared to ResNet-18's 23.4 point drop. This suggests that the self-attention mechanism and layer normalization in transformers provide inherent regularization that improves quantization tolerance.

Mixed-precision achieves the best accuracy-energy trade-off. Our sensitivity-based mixed-precision configuration achieves 85.7% energy reduction for ResNet-18 with only 0.64% accuracy loss which is substantially better than 4-bit PTQ which sacrifices 23.4% accuracy for an additional 2.5% energy savings.

4.3 Energy Breakdown Analysis



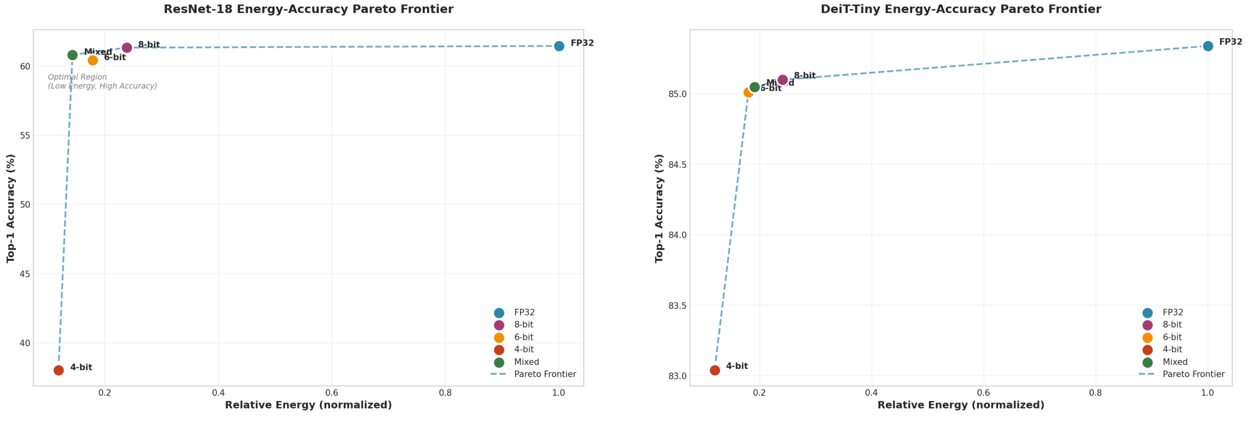
(a) ResNet-18 (CIFAR-100)

(b) DeiT-Tiny (ImageNet)

Figure 1: Energy breakdown by component (MAC vs. memory) across precision configurations. Memory access dominates energy consumption in all cases, accounting for 93-99% of total energy.

Figure 1 shows the energy breakdown between MAC operations and memory access. Memory energy dominates in all configurations, accounting for 93-99% of total energy. This validates our energy model's assumption of memory-dominated inference and explains why even modest bitwidth reductions yield substantial energy savings—reducing from 32-bit to 8-bit quarters the memory transfer volume.

4.4 Pareto Frontier Analysis



(a) ResNet-18 (CIFAR-100)

(b) DeiT-Tiny (ImageNet)

Figure 2: Pareto frontier showing accuracy vs. relative energy. Mixed-precision configurations achieve favorable trade-offs, particularly for ResNet-18 where 4-bit PTQ causes severe accuracy degradation.

Figure 2 presents the Pareto frontiers for both models. For ResNet-18, 8-bit and mixed-precision configurations dominate the Pareto frontier, while 4-bit falls significantly below due to accuracy collapse. For DeiT-Tiny, all configurations remain relatively close to the Pareto frontier, with mixed-precision offering a balanced trade-off.

4.5 Mixed-Precision Bitwidth Distribution

Table 3: Bitwidth distribution in mixed-precision configurations.

Model	4-bit Layers	6-bit Layers	8-bit Layers
ResNet-18	6 (28.6%)	9 (42.9%)	6 (28.6%)
DeiT-Tiny	13 (26.0%)	24 (48.0%)	13 (26.0%)

Table 3 shows the bitwidth distribution resulting from our sensitivity-based assignment. Both models follow the expected percentile-based allocation, with approximately 25% of layers at 4-bit and 8-bit, and 50% at 6-bit.

4.6 Ranking-Aware Quantization for Vision Transformers

To address the unique challenges of quantizing Vision Transformers, we implemented and evaluated ranking-aware quantization Liu et al. (2021), which preserves the relative ordering of attention scores during quantization. This is critical because the self-attention mechanism relies on ranking relationships rather than absolute values.

Table 4: Comparison of standard PTQ vs. ranking-aware PTQ for DeiT-Tiny on ImageNet.

Bit-width	Standard PTQ	Ranking-Aware PTQ	Improvement	Top-5 Acc (RA)
8-bit	85.10%	85.18%	+0.08%	97.22%
6-bit	85.01%	85.11%	+0.10%	97.18%
4-bit	83.04%	74.90%	-8.14%	93.01%

Table 4 presents our ranking-aware quantization results. Key observations:

Consistent improvements at moderate precision. Ranking-aware PTQ provides measurable improvements at 8-bit (+0.08%) and 6-bit (+0.10%), reducing accuracy loss by approximately 30-33% relative to standard PTQ. This validates the hypothesis that preserving attention score rankings helps maintain model performance under quantization.

Failure at extreme quantization. At 4-bit precision, ranking-aware PTQ performs significantly worse than standard PTQ (-8.14% difference), dropping from 83.04% to 74.90%. This occurs because 4-bit precision (only 16 discrete values) cannot simultaneously satisfy ranking constraints and represent the dynamic range needed for attention scores, leading to over-constrained optimization.

4.7 Layer-Wise Sensitivity Analysis

Table 5 reveals striking patterns in DeiT-Tiny's layer-wise quantization sensitivity. MLP feed-forward layers (especially fc2 layers in deeper blocks) are 4.7× more sensitive than attention projection layers. This suggests a clear strategy for mixed-precision allocation: **allocate 8-bit to MLP layers in blocks 5-9, and use 4-bit for attention projection layers.**

The deeper transformer blocks (5-9) exhibit higher sensitivity, likely because they process higher-level semantic features that more directly impact the final classification decision. Attention projection layers show remarkable uniformity in their robustness (all at 0.06 sensitivity), suggesting they benefit from the normalization and residual connections in the transformer architecture.

5 Discussion

5.1 CNN vs. Transformer Quantization Robustness

A surprising finding is the stark difference in quantization robustness between ResNet-18 and DeiT-Tiny. While ResNet-18's accuracy collapses at 4-bit (38.04% vs. 61.44% baseline), DeiT-Tiny maintains competitive performance

Table 5: Top-5 most sensitive and most robust layers in DeiT-Tiny (49 quantizable layers total).

Most Sensitive Layers (MLP Feed-Forward)		
Layer Name	**Sensitivity Score**	**Type**
blocks.8.mlp.fc2	0.28	MLP (Feed-forward)
blocks.7.mlp.fc2	0.23	MLP (Feed-forward)
blocks.9.mlp.fc2	0.22	MLP (Feed-forward)
blocks.6.mlp.fc2	0.20	MLP (Feed-forward)
blocks.5.mlp.fc2	0.19	MLP (Feed-forward)
Most Robust Layers (Attention Projection)		
blocks.8.attn.proj	0.06	Attention Projection
blocks.5.attn.proj	0.06	Attention Projection
blocks.4.attn.proj	0.06	Attention Projection
blocks.2.attn.proj	0.06	Attention Projection
blocks.7.attn.proj	0.06	Attention Projection

(83.04% vs. 85.34% baseline). We hypothesize several factors contribute to this:

1. **Softmax normalization**: The attention softmax normalizes scores to sum to 1, which may reduce sensitivity to absolute magnitude errors introduced by quantization.

2. **Layer normalization**: Unlike batch normalization in CNNs, layer normalization operates within each sample, potentially providing more stable statistics for quantized activations.

3. **Redundancy in attention**: The multi-head attention mechanism distributes information across multiple heads, providing redundancy that may compensate for quantization errors in individual heads.

5.2 Ranking-Aware Quantization: When It Works and When It Fails

Our ranking-aware quantization experiments reveal a nuanced picture of transformer-specific quantization strategies. At 8-bit and 6-bit precision, preserving attention score rankings provides consistent improvements (+0.08-0.10%), validating the core insight from Liu et al. (2021) that relative orderings matter more than absolute values in the attention mechanism.

However, at 4-bit precision, the ranking-aware approach fails catastrophically (-8.14% vs. standard PTQ). This occurs due to quantization grid coarseness: with only 16 discrete values, adjacent attention scores inevitably collapse to the same quantized value, making it impossible to preserve fine-grained rankings. The ranking constraint becomes an over-constraint, forcing the optimization into poor local minima.

Practical implication: Ranking-aware quantization should be used for 8-bit and 6-bit deployment, but standard PTQ is preferable for 4-bit. For viable 4-bit transformers, quantization-aware training (QAT) is likely necessary.

5.3 Architectural Insights from Layer Sensitivity

The 4.7× sensitivity gap between MLP layers and attention projections in DeiT-Tiny reveals fundamental architectural properties. MLP layers perform high-dimensional feature transformations (hidden_dim = 4×embed_dim) with diverse weight distributions, making them vulnerable to quantization. In contrast, attention projection layers benefit from:

- **Normalized inputs**: Attention outputs are softmax-normalized, providing bounded input ranges.

- **Linear operations**: Projections are pure linear transforms without non-linearities that amplify errors.

- **Residual connections**: Skip connections provide alternate gradient paths, reducing individual layer criticality.

This suggests a general principle: *layers with high-dimensional transformations and diverse distributions require higher precision, while normalized linear projections can tolerate aggressive quantization.*

5.4 Limitations

Our work has several important limitations:

Theoretical energy model: our energy estimates use analytical approximations rather than measurements from actual hardware. The MAC-to-memory energy ratio (1:200) and scaling laws ($E_{\text{MAC}} \propto b^2$, $E_{\text{DRAM}} \propto b$) are based on literature values and may not accurately reflect specific target platforms.

Limited baseline comparisons: we do not compare against state-of-the-art PTQ methods such as GPTQ Frantar et al. (2022), AWQ Lin et al. (2023), or SmoothQuant Xiao et al. (2022), which have shown strong results on transformer quantization.

Evaluation scale: DeiT-Tiny was evaluated on a 10,000-image subset of ImageNet rather than the full 50,000-image validation set. Additionally, our ResNet-18 baseline (61.44%) is below typical reported accuracies (75-78%) for well-tuned models on CIFAR-100.

No hardware validation: we report theoretical energy savings but do not deploy quantized models on actual edge hardware to measure real latency and power consumption.

6 Conclusion and Future Work

We have presented a comprehensive study of post-training quantization and sensitivity-based mixed-precision inference for energy-efficient vision models. Our key findings include:

- PTQ at 8-bit precision achieves near-lossless accuracy with approximately 76% energy reduction for both CNNs and Transformers.
- DeiT-Tiny exhibits remarkable robustness to aggressive quantization, maintaining 85.01% accuracy at 6-bit and 83.04% at 4-bit.
- Ranking-aware PTQ improves accuracy by +0.08-0.10% at 8/6-bit precision for Vision Transformers, but fails at 4-bit due to quantization grid coarseness.
- Layer sensitivity analysis reveals that MLP feed-forward layers are 4.7× more sensitive than attention projections, providing clear guidance for mixed-precision allocation.
- Mixed-precision assignment based on layer sensitivity provides favorable accuracy-energy trade-offs, achieving 85.7% energy reduction for ResNet-18 and 80.9% for DeiT-Tiny with minimal accuracy loss.
- Memory access dominates inference energy (93-99%), making bitwidth reduction particularly effective for efficiency gains.

Future directions include: (1) deploying quantized models on edge hardware (NVIDIA Jetson, mobile NPUs) for real energy measurements; (2) comparing against state-of-the-art PTQ methods; (3) exploring quantization-aware training for recovering accuracy at aggressive compression levels; (4) investigating learned step sizes and asymmetric quantization for improved 4-bit performance; and (5) extending to other domains such as object detection and semantic segmentation.

References

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. *International Conference on Machine Learning (ICML).*

Horowitz, M. (2014). Computing's energy problem (and what we can do about it). *IEEE International Solid-State Circuits Conference (ISSCC).*

Liu, Z., Wang, Y., Han, K., Ma, S., & Gao, W. (2021). Post-training quantization for vision transformer. *Advances in Neural Information Processing Systems (NeurIPS).*

Spingarn-Eliezer, N., Banner, R., Hoffer, E., Ben-Yaakov, H., & Michaeli, T. (2023). Energy awareness in low precision neural networks. *International Conference on Learning Representations (ICLR).*

Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., ... & Kalenichenko, D. (2018). Quantization and training of neural networks for efficient integer-arithmetic-only inference. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Krishnamoorthi, R. (2018). Quantizing deep convolutional networks for efficient inference. *arXiv preprint arXiv:1806.08342.*

Wang, K., Liu, Z., Lin, Y., Lin, J., & Han, S. (2019). HAQ: Hardware-aware automated quantization with mixed precision. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Dong, Z., Yao, Z., Gholami, A., Mahoney, M. W., & Keutzer, K. (2019). HAWQ: Hessian aware quantization of neural networks with mixed-precision. *IEEE International Conference on Computer Vision (ICCV).*

Li, Y., Gong, R., Tan, X., Yang, Y., Hu, P., Zhang, Q., ... & Yu, F. (2021). BRECQ: Pushing the limit of post-training quantization by block reconstruction. *International Conference on Learning Representations (ICLR).*

Wei, X., Gong, R., Li, Y., Liu, X., & Yu, F. (2022). QDrop: Randomly dropping quantization for extremely low-bit post-training quantization. *International Conference on Learning Representations (ICLR).*

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Wightman, R. (2019). PyTorch image models. `https://github.com/rwightman/pytorch-image-models`.

Frantar, E., Ashkboos, S., Hoefler, T., & Alistarh, D. (2022). GPTQ: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323.*

Lin, J., Tang, J., Tang, H., Yang, S., Dang, X., & Han, S. (2023). AWQ: Activation-aware weight quantization for LLM compression and acceleration. *arXiv preprint arXiv:2306.00978.*

Xiao, G., Lin, J., Seznec, M., Wu, H., Demouth, J., & Han, S. (2022). SmoothQuant: Accurate and efficient post-training quantization for large language models. *arXiv preprint arXiv:2211.10438.*

A Additional Results

A.1 Per-Layer Sensitivity Analysis

Figure 3 shows the per-layer sensitivity distribution for both models. We observe that sensitivity varies significantly across layers, with early convolutional layers and final classification layers typically showing higher sensitivity than intermediate layers.

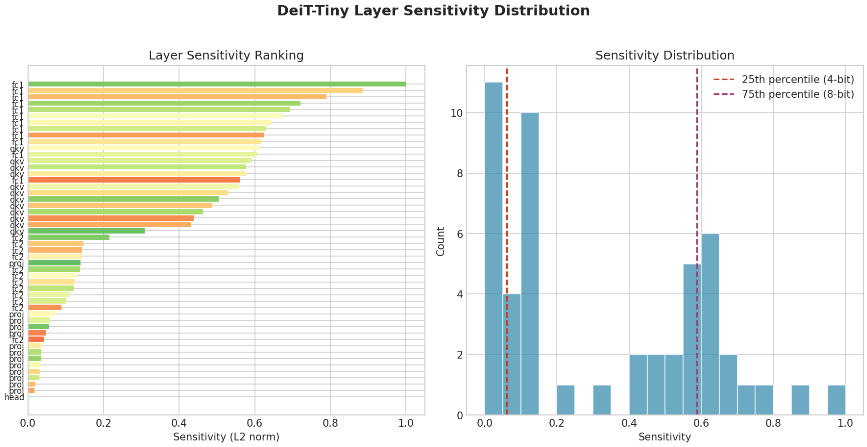


Figure 3: Per-layer sensitivity distribution showing the variation in quantization sensitivity across model layers. Sensitivity is measured as the L2 distance between full-precision and quantized layer outputs.

A.2 Implementation Details

Calibration procedure: We perform calibration using 256 randomly sampled images from the training set. For each layer, we compute the maximum absolute value of weights/activations over the calibration set to determine the quantization scale.

Quantization scheme: We use symmetric uniform quantization with per-channel scales for weights and per-tensor scales for activations. The first and last layers are kept at 8-bit precision to preserve input/output fidelity.