

Rescaling-Aware Training for Efficient Deployment of Deep Learning Models on Full-Integer Hardware

Lion Müller[✉], Alberto García-Ortiz[✉], Ardalan Najafi[✉], Adam Fuks and Lennart Bamberg[✉]

Abstract—Integer AI inference significantly reduces computational complexity in embedded systems. Quantization-aware training (QAT) helps mitigate accuracy degradation associated with post-training quantization but still overlooks the impact of integer rescaling during inference, which is a hardware costly operation in integer-only AI inference. This work shows that rescaling cost can be dramatically reduced post-training, by applying a stronger quantization to the rescale multiplicands at no model-quality loss. Furthermore, we introduce Rescale-Aware Training, a fine tuning method for ultra-low bit-width rescaling multiplicands. Experiments show that even with $8\times$ reduced rescaler widths, the full accuracy is preserved through minimal incremental retraining. This enables more energy-efficient and cost-efficient AI inference for resource-constrained embedded systems.

Index Terms—Quantization, Edge AI, Hardware Accelerator

I. INTRODUCTION

Hardware-efficient full-integer-quantized AI inference is the de-facto standard in today’s embedded systems. A crucial step of full-integer quantization is the rescaling of high bitwidth accumulators to low bitwidth outputs. This operation is implemented in integer-only hardware by multiplying the 32-bit accumulator with a wide unsigned integer, m , followed by a shift-right-and-round and casting to 8-bit [1]. Consequently, commercial embedded integer neural processing units (NPUs) such as [2] integrate multiple wide multipliers for data-parallel processing (see Fig. 1)—causing significant hardware cost.

One approach to reduce rescaling overhead is to restrict the rescaler to a power-of-two, allowing the operation to be implemented using simple bit-shifting [3]–[6]. However, this requires coarse discretization of the quantization scales, which can lead to noticeable degradation in model accuracy.

We aim to optimize the cost-quality trade-off by reducing the bit-width of m from 32 bits to a smaller value. This also decreases the width of the multiplication output, reducing the cost of the subsequent shift-and-cast operation. We show that one can minimize the rescaling width to 8 bit, from today’s 32 bit, without impairing model accuracy.

To enable even lower bitwidths for m , we propose a fine-tuning method called *Rescale-Aware Training*. This method emulates the quantization errors of the rescale hardware during the forward pass. Proper gradient propagation during the backward pass is ensured by Straight-Through Estimation. By leveraging the inherent noise resilience of deep neural networks, our approach enables an $8\times$ reduction in the bitwidth of the rescaler multiplicand while preserving model accuracy, all within a reasonable retraining effort.

Lion Müller, Ardalan Najafi, and, Adam Fuks Lennart Bamberg are with NXP Semiconductors. Alberto García-Ortiz is with University of Bremen.

This work is funded by the IPCEI ME/CT initiative by the European Union.

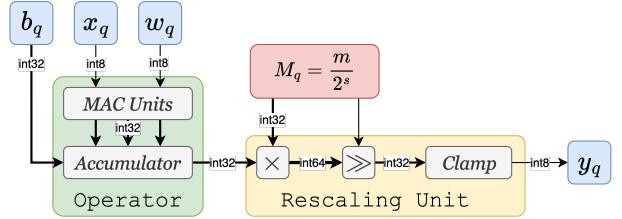


Fig. 1: Visualization of full-integer arithmetic proposed in [1].

A. Related Work: Integer-only arithmetic for efficient AI inference on resource-constrained devices was introduced by Jacob et al. [1], along with Quantization-Aware Training (QAT), enabling full-integer quantization with 8-bit weights and 32-bit rescaling while maintaining accuracy.

Subsequent works refined quantization techniques to improve accuracy and hardware efficiency. Nagel et al. [7] proposed weight equalization and bias correction to reduce quantization-induced bias. Li et al. [3] introduced Additive Powers-of-Two (APoT) weight quantization, enabling shift-and-add based, low-cost dot-product. Efficient rescaling has been addressed by training all quantization thresholds under Power-of-Two (PoT) constraints [8]. This was extended with hardware-compliant schemes with better resolution near zero to improve accuracy [5].

Existing power-of-two methods have advanced the field of integer-only inference, by a fixed trade-off of cost for model quality. A fine-grain cost-quality trade-off in rescaling remains not explored in sufficient detail. This work addresses that gap by proposing *Rescale-Aware Training*, a technique that adapts pre-trained models to low-bitwidth rescaling multiplicands without compromising accuracy.

II. METHODOLOGY

Our implementation is based on the LiteRT framework, which compiles models into a hardware-agnostic intermediate representation. We modify this framework to enable accurate inference modeling under specific rescaler configurations. The same approach is used to implement rescale-aware training.

A. Preliminaries: This section formalizes the quantized inference setting and introduces the relevant assumptions, quantization schemes, and computational models.

1. Quantization Scheme: We assume an integer-only inference pipeline with the following quantization configurations:

- **Weights:** Symmetric, per-channel, 8-bit, scale S_w
- **Activations:** Affine, 8-bit, scale S_x , zero-point Z_x
- **Bias:** Symmetric, per-channel, 32-bit, scale $S_x \cdot S_w$

This format aligns with all widely adopted deployment practices (e.g., LiteRT, TOSA, Executorch) and enables efficient integer-arithmetic hardware.

2. Integer Quantized Inference: Let \vec{x}_q , y_q , and \vec{w}_q denote the 8-bit quantized input/output activations and weights, with scales (S_x, S_y, S_w) and zero-points (Z_x, Z_y) . The core scaled-and-biased dot-product operation (used to realize matrix multiplications, convolutions, etc. in AI) is computed as:

$$y_q = \text{Sat}_{\text{int}8}(\lfloor M_q \sum_i (x_{q,i} \cdot w_{q,i}) + b_{\text{eff}} \rfloor) \quad (1)$$

Here, b_{eff} is the effective bias, incorporating both the neural bias and the impact of input/output zero-points as described in [1]. The 32-bit symmetrically quantized accumulator, a_q , holds the biased results and the dot-product at scale $S_x \cdot S_w$. To cast the accumulator output to the desired scale S_y , a rescaling factor $M = \frac{S_x \cdot S_w}{S_y}$ is applied. Since the bit-width is reduced from 32 to 8 bits, M falls within the range $(0, 1]$.

The final saturation to the int8 range not only casts the result to the target format but also implements activation functions that are a mere clamping (e.g., ReLU, ReLU6), commonly used in integer-only inference.

3. Quantization of the rescale factor: The dyadic quantization of the rescale factor into an integer multiplier and associated right-shift-and-round is derived from [1]:

$$M_q = m \cdot 2^{-s} \quad (2)$$

In [1], m is a 32-bit integer and s is chosen to have a leading 1-bit in m to maximize the effective precision gained. We propose keeping the multiplier width as $k < 32b$, investigating how small we can go without loss of accuracy.

This work proposes a computationally efficient way to obtain the k -bit value m and the shift amount s compared to [1], to speed up our proposed rescale-aware training. The real-valued IEEE 754 floating-point rescale value M is represented by a fractional part (mantissa) and exponent as: $r \simeq (1 + fr) \cdot 2^{exp}$. The multiplier m can be obtained without arithmetic operations by appending the k most significant bits of mantissa fr to a leading 1 (hidden bit). Similarly, s is obtained from exp by adding a simple integer constant.

Note that even for small k , s will be positive in all cases, as $m \geq 1$ and $M \leq 1$. Thus, a narrow *uint* is used for s . With low- k inference, s can use as few as $\log_2(32 + k - 8)$ bits. The multiplication results of the accumulator and m is of width $32 + k$, shifted right by at most $32 + k - 8$ to fully exploit the 8-bit output range. Consequently, reducing k not only lowers the complexity of the multiplication stage but also simplifies the subsequent shift and rounding operations.

4. Rescale Error Model: The quantization error of the quantized accumulator value before rescaling, a_q , relative to the true value a , is obtained by dequantization the value to \tilde{a} (i.e., multiplying with the quantization scale) as

$$\epsilon_a = a - \tilde{a} = a - S_x S_w a_q. \quad (3)$$

This error comes from the weight, input, and bias quantization. During rescaling, the quantized accumulation result is multiplied by the quantized rescaler M_q , defined in Eq. (2). After rescaling the quantized values use scale S_y . Dequantizing, to \tilde{y} allows obtaining the output error:

$$\epsilon_y = y - \tilde{y} = y - S_y \cdot \text{Sat}_{\text{int}8}(\lceil M_q a_q \rceil) \quad (4)$$

Algorithm 1 MultiplyByQuantizedMultiplier

Ensure: Gradient propagation of all rounding operations

```

1: function MULTIPLYBYQUANTIZEDMULTIPLIER( $x$ , multiplier, shift)
2:    $rnd \leftarrow 2^{(\text{shift}-1)}$                                 ▷ enable round-to-nearest
3:    $mul \leftarrow x \times \text{multiplier} + rnd$ 
4:    $res \leftarrow \text{Floor}(mul \times 2^{-\text{shift}})$ 
5:   return  $res$ 
6: end function
```

Comparing the dequantized outputs before and after rescaling, yields the error introduced by the rescaling operation:

$$\epsilon_r = \epsilon_a - \epsilon_y = \tilde{y} - \tilde{a} = S_y a_q (M_q - M) + S_y \delta_r \quad (5)$$

Active saturation under valid quantization ranges ties the error typically to 0 (ReLU activation as 0 is always quantized error-free). Thus, we assume no saturation (i.e., $a = y$), to consider the worst-case error. The derived formulation of the rescale error ϵ_r consists of two parts. One due to the discrepancy between the ideal rescaling factor M and its dyadic approximation M_q . After rescaling, this error is scaled by the output quantization scale S_y times the accumulator value a_q , and is bounded by $|M_q - M| \cdot \max(|\tilde{y}|)$. The second component originates from the rounding of the right-shift operation (*cf.* Algorithm 1). This rounding error is uniformly distributed in the range $(-\frac{S_y}{2}, \frac{S_y}{2})$.

For reduced rescale bit-widths, the component of the error associated with quantization of the rescale factor exhibits exponential growth, with its magnitude distributed proportionally to the accumulator. This occurs when $|M_q - M| \cdot \max|S_y a_q| > \frac{S_y}{2}$, caused by more severe quantization of the rescale factor. In contrast, when $|M_q - M| \rightarrow 0$, the total rescale error is instead dominated by the rounding/quantization error from going down to 8-bit.

B. Validation (Emulate Rescaling Step): To validate the proposed method, we guarantee output parity between the forward path during retraining and quantized inference. This is achieved by recompiling LiteRT with modified reference kernels. In detail, the global routine to quantize floating-point rescale factors is altered to reflect a specific rescaling bit-width. Similarly, the routine to perform the rescaling operation is adapted (*cf.* Algorithm 1).

C. Rescale-Aware Finetuning: To support effective training under quantization constraints, our method adopts a rescale-aware training approach that emulates integer inference using floating-point arithmetic. This ensures consistency with the behavior of LiteRT’s quantized deployment path while maintaining accurate differentiability. The key considerations and mechanisms of this training procedure are outlined below.

1. Hardware Modeling: Although LiteRT compiles models into a generic intermediate representation, it implicitly assumes specific hardware behaviors during quantized inference. To accurately replicate these behaviors in the training loop, we adopt the following assumptions:

- **Accumulator Precision:** 32-bit accumulators are used.
- **Rounding Strategy:** Rounding follows a hardware-efficient round-to-nearest (round-half-up) convention.
- **Output Saturation:** Rescaling is performed with int64 intermediate results to avoid overflow. The result is stored in int32 and then saturated to the output data type (e.g., int8).

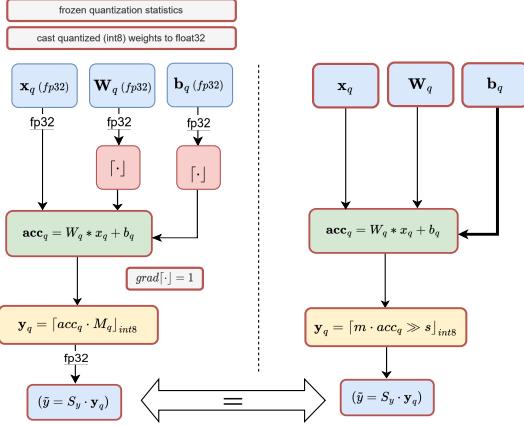


Fig. 2: Left: training graph with emulated integer-only arithmetic. Right: integer quantized inference graph.

2. Emulation of Integer Inference during Training: Training under these hardware assumptions requires floating-point emulation of integer arithmetic (shown in Figure 2). We derive a quantization-aware training loop that mirrors integer inference behavior while supporting gradient-based updates:

- **Weight Preparation:** Pre-trained LiteRT models are used for initialization. Integer weights are cast directly to floating-point format without undergoing dequantization. Given that quantization scales are assumed to be fixed, the corresponding rescale factors are likewise treated as constant.
- **Numerical Stability:** Accumulations and rescaling operations are performed in float64 to match the precision of int32. This prevents rounding instabilities that may arise due to the limited 24-bit mantissa of float32.
- **Fake Quantization:** All parameters are subjected to fake quantization during the forward pass. Thereby, the outputs produced during training are representative of those observed during inference.
- **Gradient Flow via STE:** Since rounding is a non-differentiable, discretized operation, we use the Straight-Through Estimator (STE) [9] during backpropagation.

D. Model Re-Deployment: After fine-tuning, the updated floating-point weights are re-quantized by rounding them to integers. These updated values are directly substituted into the LiteRT flatbuffer model, effectively replacing the original quantized weights with those optimized through training, while keeping the original quantization parameters.

III. EXPERIMENTAL RESULTS

A. Post-Training Rescaler Bit-Width Reduction: First, we quantify model robustness to post-training quantization of the rescale-multiplicand bit-widths. This point not only serves as a reference point for our fine-tuning method, but also defines for hardware architects the cost reduction opportunity over the state-of-the-art even without model fine-tuning.

We consider standard efficient-AI models trained on the ImageNet1000 classification task. We define the degradation point as the rescaling bit-width at which the classification accuracy declines by more than 0.5%. The results are presented in Table I.

TABLE I: Impact of post-training rescaler bit-width reduction on different LiteRT ImageNet classification models.

Model	Base Accuracy	Accuracy per rescaler bit-width				
		8	6	5	4	3
EfficientNet-Lite0	71.28%	70.93%	69.83%	69.20%	65.39%	21.72%
MobileNet-V1	70.51%	70.51%	70.41%	67.24%	63.67%	05.11%
MobileNet-V2	71.09%	71.10%	70.82%	69.28%	54.55%	23.26%

TABLE II: EfficientNet-Lite0 with a 4-bit rescaler.

Evaluation	Accuracy	Weight Change [%]	
		ratio	Abs. diff.
Original	71.28%		
Post-training	65.39%		
re-trained (2 epochs)	71.62%	0.66	0.71

Our findings indicate that reducing the rescaling width to 8 bits (representing a 4× reduction over the state-of-the-art) does not compromise model accuracy in all cases.

B. Rescale-Aware Fine Tuning: This subsection analyzes the effect of rescale-aware fine-tuning of model weights, using the *EfficientNet-Lite0* architecture as a case study. We investigate both, the proportion of weights that were modified during the process and the mean absolute deviation of the updated weight values relative to those in the original quantized model.

All experiments were conducted using Stochastic Gradient Descent (SGD) without any modification to the default learning rate provided by TensorFlow. Although the weights were directly taken from the quantized model—thus retaining large absolute values due to the absence of dequantization—the optimizer proved stable and effective.

The assumption that only minimal weight changes would be required, is confirmed by the empirical results. Weight changes following 2 epochs of retraining for a 4-bit rescaler are minor. Specifically, Table II indicates that merely 0.66% of the weights were updated, with the average absolute change remaining below 0.71%. Despite the limited extent of these changes, they were sufficient to yield a substantial improvement in classification accuracy—approximately 6% after two epochs, resulting in full accuracy recovery.

All weight updates observed after a single epoch of training were of magnitude ± 1 . Furthermore, only 50% (25/50) of the layers containing weights were affected—predominantly those in the later stages of the network. This observation aligns with prior findings in aggressive quantization approaches [10], which frequently exclude the first layer from quantization due to its heightened sensitivity.

A second epoch revealed additional adaptivity in the upper layers and introduced minor modifications in three more layers. In particular, the first depth-wise convolutional layer exhibited consistent weight changes of ± 2 across a substantial number of parameters. Overall, we notice for this layer a reduced number of zero-valued weights (sparsity) as well as a shift in mean value. This suggests a compensatory mechanism that addresses the mean activation shift introduced during rescaling. Such dynamics are consistent with the hypothesis of biased output distributions in small-kernel layers due to quantized weights, as discussed by Finkelstein et al. [11].

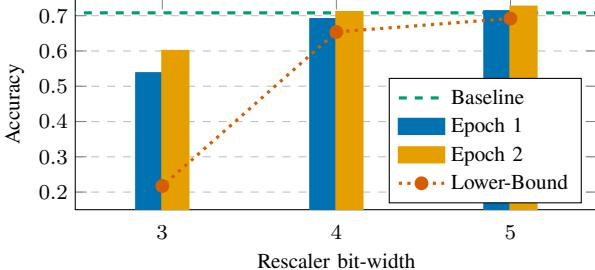


Fig. 3: EfficientNet-Lite0 accuracy recovery.

An alternative explanation is that the network compensates for systematic output bias through weight adjustments, similar in spirit to the Incremental Bias Correction method, which shows strong results with few calibration samples (typically 8–64). While bias correction explicitly adjusts layer biases, the observed behavior here suggests that incremental weight fine-tuning may achieve similar ends indirectly.

Further insight is obtained by analyzing the relative changes in weight values. After one epoch of training, the majority of weight updates exhibit a magnitude of ± 1 , suggesting that most updates originated from previously zero-valued (i.e., ineffective) weights.

C. Evaluation of Performance Recovery: Building on the near-output parity between the integer-only arithmetic emulated inference path and the customized LiteRT implementation, as well as the previously established lower bounds on rescaler bit-width, this section evaluates the effectiveness of retraining in restoring model performance.

The *EfficientNet-Lite0* model was successfully retrained using rescaler bit-widths of 4 and 5 bits. As shown in Figure 3, performance significantly improved even after a single epoch of rescale-aware training. In the 4-bit case, the accuracy nearly reached the original baseline after the first epoch. Following a second epoch, the accuracy not only closed the gap but slightly exceeded the original baseline accuracy of 71.28%.

This marginal improvement over the baseline may be attributed to further training or subtle differences in the training and validation datasets used in this study versus those used during the original model’s development. Therefore, any interpretation of out-performance should be made with caution. In the more aggressive 3-bit case—where performance degraded substantially after quantization (see Figure 3)—retraining improves accuracy but can no longer recover the original quality.

Note: After retraining, the model will only reach its maximum accuracy for the specific rescaler bit-width or hardware implementation employed during fine tuning. The assumptions made for these experiments were presented in 1.

D. Hardware efficiency: To assess the impact of reducing the rescaler multiplicand bit-width on hardware cost by means of a commercial embedded-NPU architecture, we synthesize the systolic dot-product-array including the rescaler from [2] in a commercial 16-nm technology. As shown in Figure 1, scaling down from 32 to 4 bits yields area reductions of 58.5%, 47.5%, and 34.8% for dot-product-engines deploying 4, 8, and 16 parallel MACs, respectively. These savings solely stem from reduced rescaling multiplicand width.

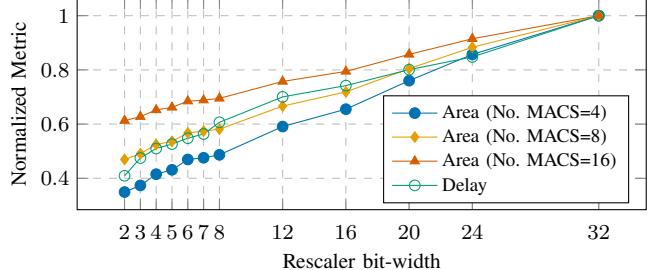


Fig. 4: Silicon area of the processing unit from Figure 1.

Relative power savings correlate with area reductions. Timing analysis shows nearly 50% improvement in the critical path across all MAC trees—comprising multiplication, adder tree, accumulation, and rescaling operations—since the rescaler multiplication consistently determines the critical path. Overall, the proposed approach reduces the area-delay product of integer-only AI accelerators by over 4 \times . For 8-bit rescale-multiplicands (to not require fine-tuning) instead of 4, all savings are approximately halved, implying still an area-delay product improvement by over 2 \times .

IV. CONCLUSION

This work shows that rescale-aware quantization and incremental training enable highly efficient embedded NPU architectures. Reducing the rescaler bit-width from 32 to 8 bits yields over 2 \times area-delay product improvement without accuracy loss over standard W8A8 model quantization. With lightweight fine-tuning, even aggressive 4-bit rescaler quantization achieves over 4 \times improvement while preserving model quality. Rescaling optimization thus offers a promising path to lower complexity and energy consumption in quantized inference. In future work, we will extend this approach to emerging Gen-AI and Agentic-AI applications at the edge.

REFERENCES

- [1] B. Jacob *et al.*, “Quantization and training of neural networks for efficient integer-arithmetic-only inference,” in *CVPR, Salt Lake City, UT, USA, June 18-22, 2018*. IEEE Computer Society, 2018, pp. 2704–2713.
- [2] L. Bamberg *et al.*, “eIQ Neutron: Redefining edge-AI inference with integrated npu and compiler innovations,” 2025. [Online]. Available: <https://arxiv.org/abs/2509.14388>
- [3] Y. Li, X. Dong, and W. Wang, “Additive powers-of-two quantization: An efficient non-uniform discretization for neural networks,” in *ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.
- [4] M. Elhoushi *et al.*, “Deepshift: Towards multiplication-less neural networks,” in *IEEE/CVF CVPR Workshops 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021, pp. 2359–2368.
- [5] D. Przewlocka-Rus *et al.*, “Power-of-two quantization for low bitwidth and hardware compliant neural networks,” 2022. [Online]. Available: <https://arxiv.org/abs/2203.05025>
- [6] X. Li *et al.*, “Denseshift : Towards accurate and efficient low-bit power-of-two quantization,” in *IEEE/CVF ICCV 2023, Paris, France, October 1-6, 2023*. IEEE, 2023, pp. 16964–16974.
- [7] M. Nagel *et al.*, “Data-free quantization through weight equalization and bias correction,” in *IEEE/CVF ICCV, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019, pp. 1325–1334.
- [8] S. Jain *et al.*, “Trained quantization thresholds for accurate and efficient fixed-point inference of deep neural networks,” in *Proc. of Machine Learning and Systems*, vol. 2, 2020, pp. 112–128.
- [9] Y. Bengio, N. Léonard, and A. Courville, “Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation,” 2013. [Online]. Available: <http://arxiv.org/abs/1308.3432>
- [10] I. Hubara *et al.*, “Quantized neural networks: Training neural networks with low precision weights and activations,” 2016. [Online]. Available: <https://arxiv.org/abs/1609.07061>
- [11] A. Finkelstein, U. Almog, and M. Grobman, “Fighting quantization bias with bias,” 2019. [Online]. Available: <https://arxiv.org/abs/1906.03193>