

Article

Adaptive A-Semilogarithmic Gradient Quantization for Efficient Deep Neural Network Training

Stefan Panić , Milan Dubljanin * , Milan Savić  and Marko Smilić 

Faculty of Sciences and Mathematics, University of Priština, Lole Ribara 29, 38220 Kosovska Mitrovica, Serbia; stefan.panic@pr.ac.rs (S.P.); milan.savic1@pr.ac.rs (M.S.); marko.smilic@pr.ac.rs (M.S.)

* Correspondence: milan.dubljanin@pr.ac.rs

Abstract

This paper introduces an adaptive A -semilogarithmic gradient quantization framework aimed at reducing memory overhead and computational complexity during the training of deep neural networks. The approach employs a semilogarithmic companding function parameterized by a dynamically adjusted scaling factor A , which evolves in response to the statistical properties of gradients throughout the training process. Two distinct quantization strategies are proposed and evaluated: The switching piecewise A -quantizer, which adaptively toggles between low-bit uniform and high-bit semilogarithmic quantization according to an exponentially weighted moving-average (EMA) estimate of gradient variance; and the hybrid A -quantizer, which statically partitions the gradient domain, applying uniform quantization in low-magnitude regions and semilogarithmic companding in high-magnitude regions. The proposed methods are empirically evaluated on both multilayer perceptron (MLP) and convolutional neural network (CNN) architectures using tabular and image-classification benchmarks, including DCCC, CIFAR-10, CIFAR-100, and ImageNet. Quantitative results demonstrate that both models achieve comparable classification accuracy to full-precision (FP32) baselines while significantly reducing gradient reconstruction error. Notably, the hybrid A -quantizer consistently yields better validation accuracy, reduced RMSE, and improved convergence behavior relative to its switching counterpart. These findings underscore the effectiveness of hybrid semilogarithmic quantization as a robust and efficient solution for training deep models in resource-constrained or bandwidth-limited environments, with strong potential for scalable deployment across diverse hardware platforms.

Keywords: gradient quantization; semilogarithmic companding; efficient deep learning; adaptive quantization



Academic Editors: Roberto Montemanni and Jia-Bao Liu

Received: 20 May 2026

Revised: 23 June 2026

Accepted: 26 June 2026

Published: 29 June 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Quantization plays a critical role in the efficient training and deployment of deep neural networks (DNNs), particularly under the growing demands of large-scale models and resource-constrained environments. While the majority of quantization research has traditionally focused on weights and activations for inference optimization [1], the quantization of gradients during training has emerged as an equally important challenge. Full-precision gradient exchange using the 32-bit IEEE 754 floating point (FP32) format [2] imposes significant communication and memory burdens, especially in distributed and federated learning settings [3]. Each training iteration involves synchronization of high-resolution gradient tensors across devices, often saturating bandwidth, increasing latency,

and elevating energy consumption [4,5]. These issues are particularly pronounced in edge learning scenarios, where devices such as IoT nodes, mobile platforms, and embedded processors operate under tight hardware constraints [6,7].

Reducing the precision of DNN components has become a central strategy for improving computational and communication efficiency during training. Several methods have been proposed in this context, falling into three main categories: quantization-aware training (QAT), post-training quantization (PTQ), and gradient-based post-training quantization (GPTQ). QAT integrates quantization operations directly into the optimization loop using straight-through estimators (STE) [8,9] or differentiable approximations [10], allowing models to adapt to quantization during training. Despite their effectiveness, such approaches often suffer from increased training cost. PTQ techniques apply quantization after training without further optimization, leveraging calibration or statistical heuristics, but typically exhibit reduced accuracy in aggressive quantization settings [1]. GPTQ methods, such as BRECQ [11] and QDrop [12], represent a middle ground, using gradient signals and small calibration sets to fine-tune quantization parameters, enabling efficient compression with minimal accuracy degradation. While most of these approaches primarily target inference-time efficiency by quantizing weights and activations, gradient quantization has become increasingly important for reducing training-time communication overhead, particularly in distributed, federated, or edge-learning systems. Gradient quantization offers a practical solution to these challenges by reducing the bit-width of gradient representations. Lower-bit formats, typically ranging from 4 to 8 bits, enable faster training, lower communication cost, and more scalable deployment without significant performance loss. However, the highly dynamic and heavy-tailed distribution of gradients during backpropagation poses challenges to simple uniform quantizers. Numerous gradient compression methods have been proposed to reduce communication overhead during distributed and federated training. Representative examples include QSGD [13], TernGrad [14], SignSGD [15], and error-feedback-based approaches such as EF-SignSGD [16]. These methods reduce communication costs through various forms of stochastic quantization, ternary encoding, sign-based compression, and error compensation mechanisms. It should be noted that gradient quantization during training and post-training model quantization address fundamentally different objectives. Gradient compression methods, such as QSGD, TernGrad, and EFSignSGD, aim to reduce communication and memory requirements associated with gradient exchange during optimization, particularly in distributed and federated learning environments. In contrast, post-training quantization techniques, including GPTQ and related approaches, focus on compressing already trained model parameters for efficient inference and deployment. Since these methods operate at different stages of the machine learning pipeline and target different optimization objectives, direct comparisons are generally not considered appropriate. Consequently, the present work focuses on comparisons with representative gradient quantization and compression methods. Nonuniform quantization techniques rooted in classical signal processing and data compression address this by allocating finer resolution near zero and coarser resolution in the tails, improving signal-to-quantization-noise ratio (SQNR) [17,18]. In this work, we propose semilogarithmic A -law quantization schemes that leverage a tunable companding parameter to match the statistical properties of gradient distributions more closely. These schemes dynamically adjust the quantization resolution to provide finer granularity where gradients are small and more aggressive compression where gradients are large. Unlike uniform or purely logarithmic quantizers, the adaptive A -law strategy provides a flexible, structure-aware mechanism for efficient gradient encoding, yielding favorable trade-offs between communication cost and training fidelity.

Among nonuniform methods, semilogarithmic A -law quantization has long been favored for its simplicity, robustness, and hardware compatibility. It employs a piecewise linear, logarithmic companding curve parameterized by a scaling factor A , offering smooth compression behavior near the origin and more controlled resolution across a wide dynamic range [18]. Compared to the more aggressive μ -law companding, which is highly nonlinear and sensitive to small gradients, A -law provides better numerical stability and is well suited to gradient distributions observed during training [18]. Unlike uniform quantization, which treats all magnitudes equally, and purely logarithmic schemes, which excessively compress small values, semilogarithmic quantization offers a balanced approach that preserves both low-amplitude and high-amplitude gradients, making it particularly effective for heavy-tailed gradient distributions commonly observed in deep networks during training. Moreover, its mathematical form aligns with standardized telecommunication codecs, making it easier to implement in low-power and embedded hardware pipelines [19].

The idea of integrating signal processing-inspired quantization into learning systems is not new. Seminal works in scalar quantization have explored hybrid [20], switched [21], and adaptive companding strategies for Laplacian and Gaussian sources [22,23]. These approaches have demonstrated the benefits of combining linear and logarithmic quantization in piecewise or segment-based architectures. Inspired by these foundations, this paper introduces two novel A -law-based gradient quantization schemes for training DNNs: (1) a switching piecewise model that toggles between uniform and logarithmic modes based on gradient variance, and (2) a hybrid segmented model that statically partitions the gradient space and applies A -law compression to large-magnitude regions. Both models incorporate adaptive A -scaling, enabling fine-grained control over the quantization curve in response to local gradient statistics.

Compared to fixed-point and FP8-based quantization methods that operate in a post-training regime [24,25], our approach works inline during training and dynamically adjusts precision based on gradient behavior. It offers an efficient and interpretable mechanism to reduce bit-width while preserving convergence and generalization. Experimental validation on diverse datasets and neural network architectures, including DCCC, CIFAR-10, CIFAR-100, and ImageNet, suggests that the proposed A -law quantizers provide a favorable trade-off between predictive performance, gradient reconstruction fidelity, and compression efficiency. These contributions highlight the practical value of classical quantization theory when re-engineered for modern AI workloads.

Recent research has emphasized the importance of adaptive and resource-efficient learning strategies in dynamic and constrained environments. For instance, Yan et al. [26] proposed an online semi-supervised ensemble framework capable of handling uncertainty and missing data in evolving data streams, highlighting the need for robust adaptive mechanisms in learning systems. Similarly, Ning [27] investigated neural network-based pattern recognition in edge computing scenarios, demonstrating the importance of low-latency and computationally efficient models. In addition, Colbu et al. [28] explored long-term predictive modeling using machine learning techniques, further emphasizing the role of adaptive model behavior in complex real-world systems. These studies collectively motivate the development of lightweight and adaptive training mechanisms, such as the proposed gradient quantization schemes, which aim to balance efficiency and performance in modern deep learning pipelines.

In our recent work [29], an adaptive μ -law gradient quantization framework was introduced for efficient training of deep neural networks, in which the companding parameter μ is dynamically adjusted according to local gradient statistics. Two complementary strategies were proposed: a switching piecewise μ -law model, which adaptively toggles between low-bit uniform quantization and high-bit logarithmic quantization based on gradient

variance, and a hybrid μ -law model, which statically partitions the gradient domain and applies nonlinear compression to high-magnitude regions. Experimental results obtained using MLP and CNN architectures demonstrated that the hybrid variant achieves superior stability and accuracy, whereas the switching scheme enables more aggressive compression, achieving up to 87.5% communication savings with minimal performance degradation.

Although the μ -law framework provides strong logarithmic compression and demonstrates excellent performance in aggressive gradient compaction scenarios, the A -law companding model offers a complementary perspective that is particularly relevant for signals with moderate dynamic range and piecewise-linear characteristics. A -law companding has been extensively used in practical communication systems due to its reduced distortion in low-amplitude regions and improved numerical stability compared with purely logarithmic mappings. Motivated by these properties, it is natural to investigate whether similar adaptive mechanisms can be successfully transferred to gradient quantization in deep learning. In particular, A -law-based schemes are expected to provide a more balanced trade-off between compression and reconstruction fidelity, especially in regimes where gradient distributions exhibit mixed-scale behavior. Building on the previously established μ -law framework, this work explores adaptive A -law gradient quantization as a potentially more stable and interpretable alternative, with the aim of achieving comparable accuracy while improving robustness and control over quantization dynamics during training.

This paper introduces two adaptive semilogarithmic quantization schemes: switching A -law and hybrid A -law, that dynamically compress gradients based on their statistical properties. Section 2 presents the mathematical formulation of each quantization method and their integration into gradient-based optimization. Section 3 outlines the experimental setup and evaluation metrics, followed by a comprehensive performance analysis, and concluding remarks in Section 4. These contributions demonstrate the practical viability of semilogarithmic quantization for efficient and accurate deep neural network training.

2. System Model

The proposed switching piecewise A -semilogarithmic quantization framework is grounded in principles derived from robust companding models originally devised for scalar quantization of Laplacian sources, particularly those used in speech coding. This method introduces a dual mode adaptive quantization scheme that alternates between a low-bit uniform quantizer and a high-bit semilogarithmic quantizer, with the switching logic governed by local gradient statistics, primarily variance and amplitude. Unlike classical fixed quantization strategies, the method dynamically modulates both the companding parameter A and the bit-width allocation based on real-time gradient behavior, drawing direct theoretical inspiration from switched Jayant quantizers and variance-adaptive scalar compression techniques. In implementation, the piecewise design partitions the gradient space into two segments: (i) gradients of small amplitude falling below a threshold ϵ , which are quantized using low-resolution uniform quantization to preserve stability, and (ii) gradients exceeding ϵ , which are quantized using a high-resolution logarithmic A -law function. This companding law is parameterized by an adaptively computed factor A , bounded within $[A_{\min}, A_{\max}]$ and updated per batch as $A = \beta \cdot \sigma_g$, where σ_g is the standard deviation of the current gradient tensor and β is a scaling coefficient. To enhance robustness, an explicit support region tracking mechanism is also employed: if the maximum absolute gradient magnitude surpasses a fixed threshold τ , the model favors aggressive quantization to mitigate overload effects. The core quantization pipeline involves nonlinear compression,

$$g_{\text{comp}} = \text{sign}(g) \cdot \frac{\log(1 + A|g|)}{\log(1 + A)}, \quad (1)$$

uniform quantization within the normalized domain, and decompression via:

$$\hat{g} = \text{sign}(g_{\text{comp}}) \cdot \frac{\exp(g_{\text{comp}} \cdot \log(1 + A)) - 1}{A}. \quad (2)$$

This formulation eliminates the need for pre-defined codebooks by leveraging the real-time statistical structure of the gradients, enabling a hardware-efficient and memory-light implementation. In the switching A -law scheme, the choice between low- and high-bit quantization is guided by the local gradient variance. Rather than relying solely on the instantaneous variance estimate, an exponentially weighted moving average (EMA) is employed to provide a statistically stable reference for the switching decision. For a given parameter tensor $\mathbf{g}_t$, the variance reference is updated according to:

$$\bar{V}_t = \alpha \bar{V}_{t-1} + (1 - \alpha) \text{Var}(\mathbf{g}_t) \quad (3)$$

where $\bar{V}_t$ denotes the EMA estimate of the gradient variance, $0 < \alpha < 1$ is the smoothing factor, and $\kappa > 0$ is a tunable sensitivity coefficient. Based on this reference, an adaptive switching threshold is defined as:

$$\tau_t = \kappa \bar{V}_t \quad (4)$$

The system applies a low-bit quantizer (e.g., 4 bits) when the current gradient variance exceeds the adaptive threshold and reverts to a higher precision quantizer (e.g., 8 bits) otherwise:

$$b_{\text{switch}} = \begin{cases} b_{\text{low}}, & \text{Var}(\mathbf{g}_t) > \kappa \bar{V}_t \\ b_{\text{high}}, & \text{otherwise} \end{cases} \quad (5)$$

This switching logic enables dynamic adjustment of precision based on relative changes in gradient activity rather than absolute variance values. Consequently, the quantizer becomes more robust to non-stationary training dynamics, improving stability while preserving bandwidth efficiency during periods of elevated gradient variability.

The algorithm's theoretical foundation is directly motivated by results in distortion-constrained quantizer design, where granular and overload distortions are minimized via regime-specific encoding. As shown in [21], switched quantizers can optimally balance compression and fidelity by dynamically responding to input variance profiles. When applied to gradient compression in neural networks, this adaptive mechanism ensures fine-grained updates for small gradients while robustly regularizing noisy updates with larger amplitude. The result is a training process that benefits from both precision and generalization, especially in communication-constrained learning environments. Experimental evaluations on multiple datasets and architectures demonstrates that the proposed semilogarithmic quantization schemes provide a favorable trade-off between predictive performance and gradient fidelity, achieving competitive validation accuracy while significantly reducing quantization-induced distortion. These results confirm the practical utility of the method, supporting its deployment in scalable, low-memory deep learning systems while preserving model convergence and performance.

The switching approach is particularly suited for stochastic training scenarios where gradient statistics evolve throughout the optimization process. By employing an EMA-based variance reference, the method adapts the quantization precision to changing training dynamics, improving robustness while maintaining a favorable trade-off between gradient fidelity, compression efficiency, and predictive performance. As shown in the experimental results, the proposed semilogarithmic quantization schemes provide a favorable trade-off between gradient fidelity and predictive performance, while substantially reducing the effective gradient precision. In our implementation, the A -law companding parameter is not computed globally from the full gradient distribution but is instead locally adapted based on the standard

deviation of gradients within the high-magnitude segment only. This per-segment adjustment of A enables finer compression control in regions where logarithmic quantization is applied, further reducing gradient distortion while preserving gradient sensitivity.

The complete implementation of the proposed EMA-based switching piecewise A -semilogarithmic quantization procedure is summarized in Algorithm 1.

Algorithm 1 Piecewise A -Semilogarithmic Quantization with EMA-Based Adaptive Bit Allocation

Require: Gradient tensor $\mathbf{g}_t$, low bit-width b_{low} , high bit-width b_{high} , segmentation threshold ε , compression factor scale β , A -law limits $[A_{\min}, A_{\max}]$, EMA smoothing factor α , switching sensitivity factor κ , previous EMA variance estimate $\bar{V}_{t-1}$

- 1: Compute gradient absolute values: $|\mathbf{g}_t|$
- 2: Define piecewise masks:

$$\mathcal{M}_{\text{uniform}} = \{i : |g_{t,i}| < \varepsilon\}, \quad \mathcal{M}_{A\text{-law}} = \{i : |g_{t,i}| \geq \varepsilon\}$$

- 3: Compute current gradient variance:

$$V_t = \text{Var}(\mathbf{g}_t)$$

- 4: Update EMA variance estimate:

$$\bar{V}_t = \alpha \bar{V}_{t-1} + (1 - \alpha) V_t$$

- 5: Determine adaptive switching bit-width:

6: **if** $V_t > \kappa \bar{V}_t$ **then**

7: $b_{\text{switch}} \leftarrow b_{\text{low}}$

8: **else**

9: $b_{\text{switch}} \leftarrow b_{\text{high}}$

10: **end if**

- 11: **Uniform segment quantization:**

$$\tilde{\mathbf{g}}_u = \text{clip} \left(\frac{\text{round}(\mathbf{g}_t \cdot (2^{b_{\text{switch}}} - 1))}{2^{b_{\text{switch}}} - 1}, -1, 1 \right)$$

- 12: **Compute adaptive companding factor for the A -law segment:**

$$\sigma_{\log} = \text{std}(\mathbf{g}_{t,A\text{-law}}), \quad A_{\log} = \min(\max(\beta \sigma_{\log}, A_{\min}), A_{\max})$$

- 13: **A -law segment quantization:**

$$\mathbf{g}_c = \text{sign}(\mathbf{g}_t) \cdot \frac{\log(1 + A_{\log} |\mathbf{g}_t|)}{\log(1 + A_{\log})}$$

$$\tilde{\mathbf{g}}_c = \text{clip} \left(\frac{\text{round}(\mathbf{g}_c \cdot (2^{b_{\text{switch}}} - 1))}{2^{b_{\text{switch}}} - 1}, -1, 1 \right)$$

$$\tilde{\mathbf{g}}_a = \text{sign}(\tilde{\mathbf{g}}_c) \cdot \frac{\exp(|\tilde{\mathbf{g}}_c| \log(1 + A_{\log})) - 1}{A_{\log}}$$

- 14: Construct final quantized gradient:

$$\tilde{g}_{t,i} = \begin{cases} \tilde{g}_{u,i}, & i \in \mathcal{M}_{\text{uniform}} \\ \tilde{g}_{a,i}, & i \in \mathcal{M}_{A\text{-law}} \end{cases}$$

- 15: **return** Quantized gradient $\tilde{\mathbf{g}}_t$ and updated EMA estimate $\bar{V}_t$
-

The hybrid A -semilogarithmic quantization approach [20] generalizes the traditional A -law companding by combining a uniform quantization segment with a logarithmic segment in a single, continuous function. This model preserves high resolution near the origin, where gradient magnitudes are small and most informative, while applying logarithmic compression for larger values, thereby reducing bit-width requirements and improving quantization robustness. Unlike the switching model, which toggles quantization modes dynamically based on statistical properties of each gradient batch, the hybrid model statically partitions the input dynamic range into two distinct segments: a uniform region and a logarithmic region. The partition point, often referred to as $x_{\min}$, is derived from a fixed scaling relationship $x_{\min} = x_{\max} / A$, where A denotes the semilogarithmic parameter controlling the non-linearity. This structure does not require per-batch decisions, enabling efficient and deterministic implementation. The number of quantization levels in the uniform and logarithmic regions, denoted N_1 and N_2 , respectively, are adaptable and can be selected based on signal statistics or fixed through forward adaptation. This flexibility enables the hybrid quantizer to be fine-tuned for optimal signal-to-quantization noise ratio (SQNR) performance across a broad range of input variances. The hybrid model has demonstrated high performance when applied to Laplacian-like distributed gradients, commonly observed in training neural networks. It provides a favorable trade-off between complexity and accuracy, and although less adaptive than switching piecewise schemes, its structured design facilitates hardware implementation and reduces computational overhead. The complete implementation of the proposed adaptive hybrid A -semilogarithmic quantization procedure is presented in Algorithm 2.

Algorithm 2 Adaptive Hybrid Piecewise A -Semilogarithmic Quantization

Require: Gradient tensor g , threshold T , adaptive A , bit-widths b_u, b_c

```

1: Compute standard deviation  $\sigma_g = \text{std}(g)$ 
2: Update  $A = \min(\max(\beta \cdot \sigma_g, A_{\min}), A_{\max})$ 
3: for all  $g_i \in g$  do
4:   if  $|g_i| < T$  then
5:     Quantize using uniform scalar quantization:

$$q_i = \text{round}\left(\frac{g_i + T}{2T} \cdot (2^{b_u} - 1)\right)$$

6:     Dequantize:

$$\tilde{g}_i = \frac{q_i}{2^{b_u} - 1} \cdot 2T - T$$

7:   else
8:     Compress:

$$g_c = \text{sign}(g_i) \left(1 + \frac{\log(|g_i|/T)}{\log(A)}\right)$$

9:     Quantize:

$$q_c = \frac{\text{round}(g_c(2^{b_c} - 1))}{2^{b_c} - 1}$$

10:    Decompress:

$$\tilde{g}_i = \text{sign}(q_c) \cdot T \cdot A^{|q_c|-1}$$

11:   end if
12: end for
13: return Quantized gradient vector  $\tilde{g}$ 

```

The compression function $C(x)$ for the hybrid quantizer is defined piecewise as:

$$C(x) = \begin{cases} \frac{x}{x_{\min}}, & \text{for } |x| \leq x_{\min} \\ \text{sign}(x) \cdot \left(1 + \frac{\log\left(\frac{|x|}{x_{\min}}\right)}{\log A}\right), & \text{for } x_{\min} < |x| < x_{\max} \end{cases} \quad (6)$$

This function applies linear quantization within the uniform region $[-x_{\min}, x_{\min}]$ and logarithmic companding outside that region up to the maximal quantization bound $x_{\max}$. The transition is seamless and parameterized by $A = x_{\max}/x_{\min}$, allowing straightforward control over the compression aggressiveness. The use of this dual-mode quantizer results in reduced quantization error in low-gradient regions and more aggressive compression in the tails, thus maintaining fidelity where it is most critical while enabling substantial memory savings.

Computational Complexity Considerations

The proposed adaptive quantization schemes require the computation of gradient statistics, primarily the variance and standard deviation of the gradient tensor, in order to determine the companding parameter and, in the switching case, the active quantization mode. For a gradient tensor containing N elements, both operations exhibit linear computational complexity, $\mathcal{O}(N)$, since they involve a single pass through the tensor elements. The overall complexity of the proposed quantizers can therefore be expressed as:

$$\mathcal{C}_{\text{quant}} = \mathcal{O}(N), \quad (7)$$

which is negligible compared with the computational cost of the forward and backward propagation steps in modern deep neural networks. For example, convolutional and transformer-based architectures typically require millions or billions of floating-point operations per iteration, whereas the proposed statistical calculations involve only simple reduction operations over the gradient tensor.

In return for this modest computational overhead, the proposed methods substantially reduce the number of bits required to represent gradient updates. Consequently, the additional processing cost is often outweighed by the reduction in communication volume, particularly in communication-constrained environments where gradient transmission dominates the overall training cost. A more detailed runtime analysis, including distributed and federated learning scenarios, remains an important direction for future work.

3. Results and Discussion

To address the concerns raised during the review process, the experimental evaluation was substantially extended beyond the original manuscript. In addition to the DCCC tabular classification benchmark, three widely used image-classification datasets were considered: CIFAR-10, CIFAR-100, and ImageNet. Furthermore, the proposed Switching A-law and Hybrid A-law quantizers were compared against three established gradient compression baselines, namely QSGD, TernGrad, and EFSignSGD. All experiments were conducted using identical training protocols for the FP32 reference model and all compressed variants. Unlike the original formulation based on a fixed variance criterion, the revised switching mechanism employs an EMA-based variance reference, resulting in improved stability and more consistent bit-width adaptation throughout training.

The evaluation focuses on three complementary performance indicators: predictive accuracy, gradient reconstruction error (RMSE), and average effective gradient precision. While classification accuracy reflects the ultimate learning performance, the RMSE metric

quantifies the distortion introduced by gradient compression. Together, these metrics provide a comprehensive view of the accuracy–compression trade-off.

3.1. Experimental Setup

All experiments were implemented in Python (version 3.13.11) using the PyTorch deep learning framework (version 2.9.1+cu130). The proposed quantization schemes were evaluated on both tabular and image-classification tasks using four benchmark datasets: DCCC, CIFAR-10, CIFAR-100, and ImageNet. These datasets provide increasing levels of complexity, enabling assessment of the proposed methods across diverse learning scenarios.

The DCCC dataset was used to evaluate the proposed methods on a binary credit-risk classification task. A multilayer perceptron (MLP) consisting of four hidden layers with 50 neurons per layer and ReLU activation functions was employed. For CIFAR-10 and CIFAR-100, a convolutional neural network (CNN) architecture comprising three convolutional blocks followed by two fully connected layers was adopted. To further investigate scalability and performance on large-scale visual recognition tasks, additional experiments were conducted on the ImageNet benchmark using a ResNet-50 architecture. Training was performed using the Adam optimizer with an initial learning rate of 10^{-3} . Unless otherwise stated, mini-batches of 128 samples were used. The proposed Switching A-law and Hybrid A-law quantizers were compared against conventional full-precision (FP32) training and three widely used gradient compression methods: QSGD, TernGrad, and EFSignSGD. All methods were evaluated under identical training conditions to ensure a fair comparison. For the Switching A-law quantizer, the bit-width adaptation mechanism was driven by an exponentially weighted moving average (EMA) estimate of gradient variance, allowing the quantization precision to adapt to evolving optimization dynamics. The Hybrid A-law quantizer employed magnitude-based gradient partitioning, combining uniform quantization for low-amplitude gradients with semilogarithmic companding for high-amplitude gradients. All simulations were executed on a workstation equipped with NVIDIA GPU acceleration and CUDA support. For adaptive quantization schemes, the average number of bits used during training is computed as:

$$B_{\text{avg}} = \frac{1}{T} \sum_{t=1}^T b_t, \quad (8)$$

where b_t denotes the bit-width selected at iteration t and T is the total number of training iterations.

3.2. DCCC Credit Default Classification

The first experiment considers the DCCC credit default dataset using a multilayer perceptron architecture. The corresponding validation accuracy and validation loss curves are shown in Figures 1 and 2, respectively.

As shown in Figure 1, all investigated methods converge successfully and achieve similar predictive performance. The proposed Hybrid A-law quantizer attains the highest final test accuracy among the proposed compression approaches, while maintaining an extremely small gradient reconstruction error. In contrast, TernGrad and EFSignSGD exhibit noticeably larger quantization errors, despite using significantly fewer bits.

The numerical results are summarized in Table 1. The Hybrid A-law approach achieves a final test accuracy of 76.28%, which is comparable to both FP32 training and QSGD, while reducing the gradient representation from 32 bits to only 8 bits. Moreover, its RMSE is approximately two orders of magnitude smaller than that of QSGD and more than two hundred times smaller than that of EFSignSGD.

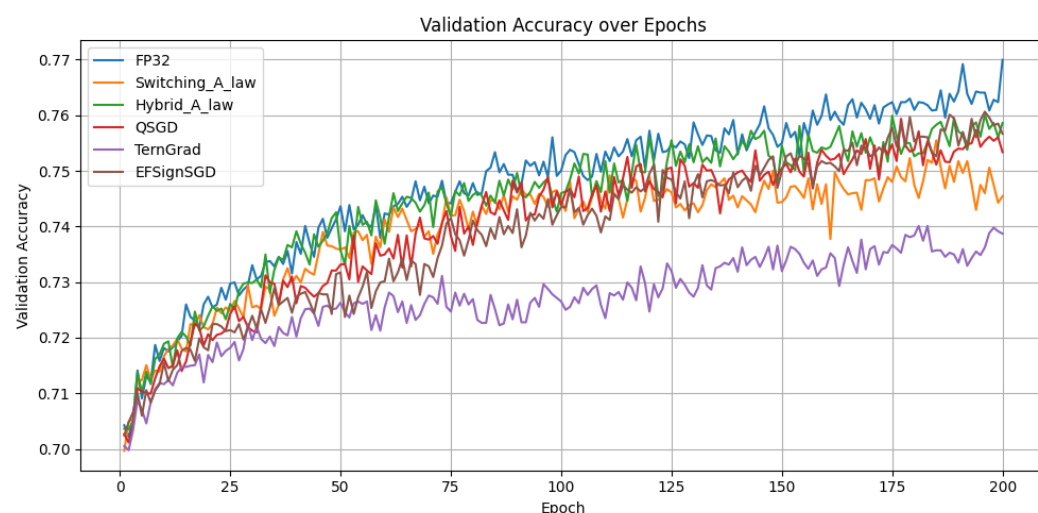


Figure 1. Validation accuracy evolution on the DCCC dataset.

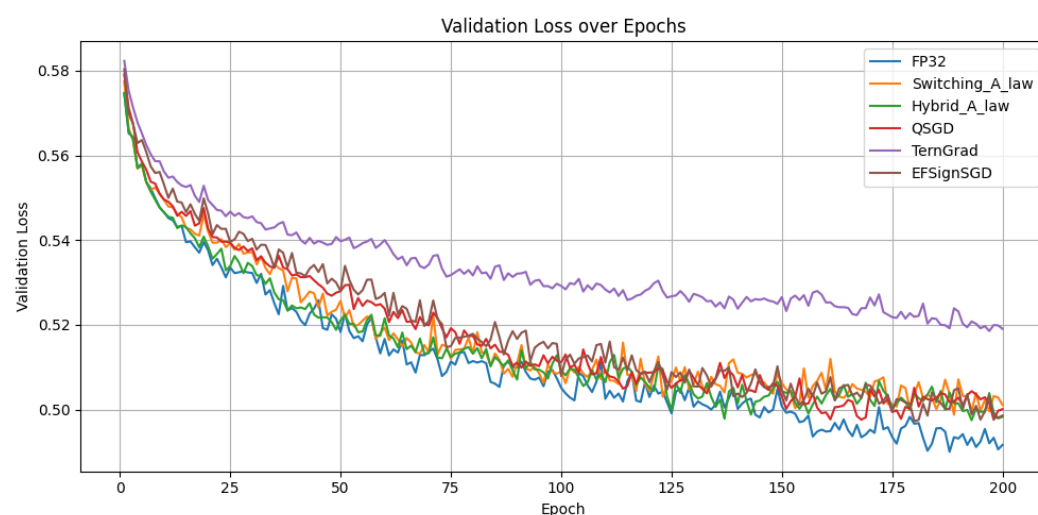


Figure 2. Validation loss evolution on the DCCC dataset.

Table 1. DCCC benchmark results.

Method	Test Acc.	RMSE	Avg. Bits
FP32	0.76097	0.000000	32.00
Switching A-law	0.75198	0.001548	7.15
Hybrid A-law	0.76279	0.000064	8.00
QSGD	0.76150	0.005305	4.00
TernGrad	0.74107	0.013718	2.00
EFSignSGD	0.75166	0.014349	1.00

These results demonstrate that semilogarithmic gradient compression is particularly effective for structured tabular learning tasks, preserving predictive performance while introducing only negligible gradient distortion.

3.3. CIFAR-10 Classification

The second experiment evaluates the proposed methods on the CIFAR-10 image classification benchmark. Figures 3 and 4 present the validation accuracy and validation loss trajectories, respectively.

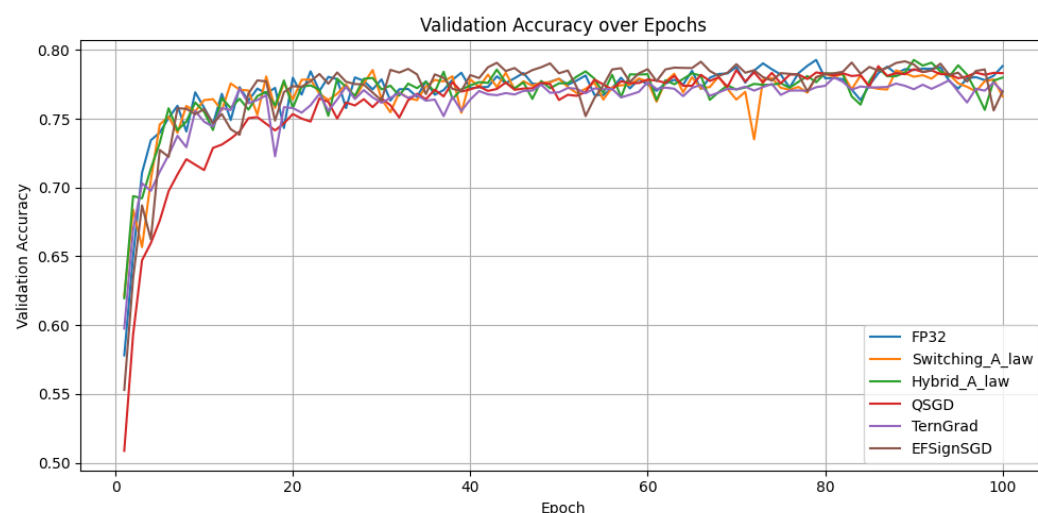


Figure 3. Validation accuracy evolution on CIFAR-10.

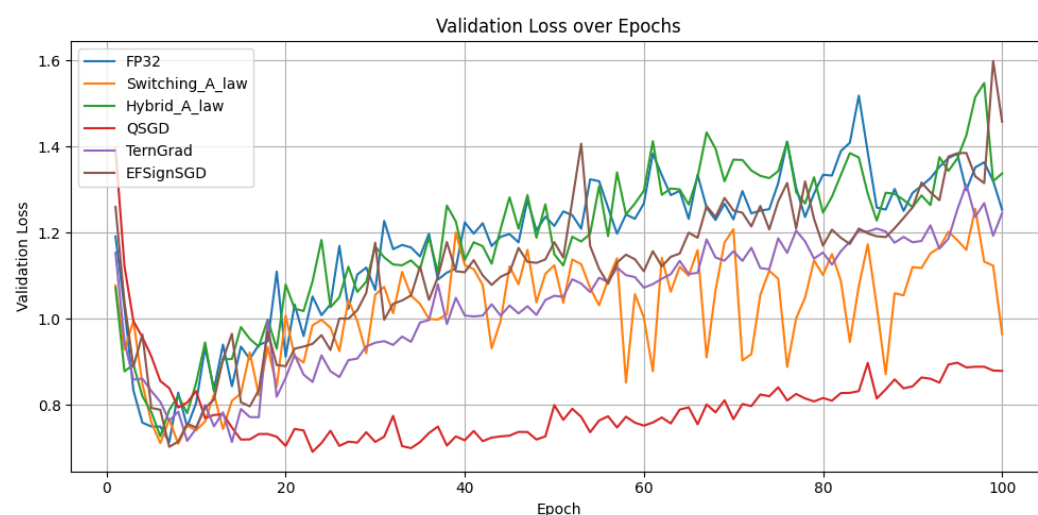


Figure 4. Validation loss evolution on CIFAR-10.

The validation accuracy curves indicate that the proposed Hybrid A-law quantizer follows the FP32 training trajectory almost exactly throughout the optimization process. The corresponding loss curves reveal similarly stable convergence behaviour.

The quantitative comparison is reported in Table 2. Notably, the Hybrid A-law quantizer achieves the same best validation accuracy as FP32 training (79.27%) while requiring only 8 effective gradient bits. Furthermore, the resulting gradient reconstruction error is only 7×10^{-6} , which is substantially lower than those obtained by QSGD, TernGrad, and EFSignSGD.

Table 2. CIFAR-10 benchmark results.

Method	Test Acc.	RMSE	Avg. Bits
FP32	0.79030	0.000000	32.00
Switching A-law	0.77220	0.000839	7.03
Hybrid A-law	0.79080	0.000007	8.00
QSGD	0.78510	0.004051	4.00
TernGrad	0.77490	0.005044	2.00
EFSignSGD	0.76900	0.001904	1.00

Among all compressed methods, the Hybrid *A-law* scheme provides the closest approximation to FP32 training, simultaneously achieving the highest classification accuracy and the lowest gradient distortion.

3.4. CIFAR-100 Classification

To evaluate robustness on a more challenging classification task, experiments were performed on CIFAR-100. The validation accuracy evolution is shown in Figure 5.

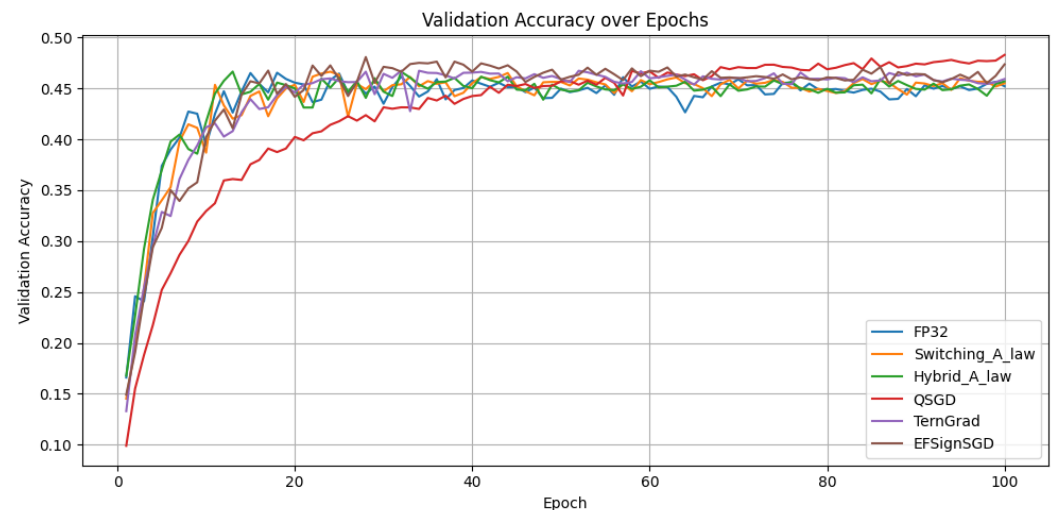


Figure 5. Validation accuracy evolution on CIFAR-100.

Compared with CIFAR-10, all methods operate in a considerably narrower accuracy range, reflecting the increased complexity of the dataset. Interestingly, QSGD and EF-SignSGD achieve slightly higher peak accuracies. Nevertheless, the proposed Hybrid *A-law* quantizer remains highly competitive while introducing dramatically lower gradient distortion.

The corresponding numerical results are summarized in Table 3.

Table 3. CIFAR-100 benchmark results.

Method	Test Acc.	RMSE	Avg. Bits
FP32	0.44480	0.000000	32.00
Switching <i>A-law</i>	0.45110	0.000514	7.78
Hybrid <i>A-law</i>	0.45400	0.000015	8.00
QSGD	0.47060	0.007242	4.00
TernGrad	0.45500	0.010322	2.00
EFSignSGD	0.46240	0.002897	1.00

An important observation emerges from this experiment. Although QSGD and EF-SignSGD achieve slightly higher classification accuracies, the Hybrid *A-law* quantizer produces gradient reconstruction errors that are two to three orders of magnitude lower. This finding suggests that semilogarithmic companding preserves gradient information substantially more faithfully than stochastic or sign-based compression approaches while maintaining highly competitive predictive performance.

3.5. ImageNet Validation

To address the reviewer's concern regarding large-scale validation, additional experiments were conducted on the ImageNet dataset using a ResNet-50 architecture. The resulting validation accuracy curves are shown in Figure 6. EFSignSGD failed to con-

verge under the considered ImageNet training configuration, resulting in near-random classification performance.

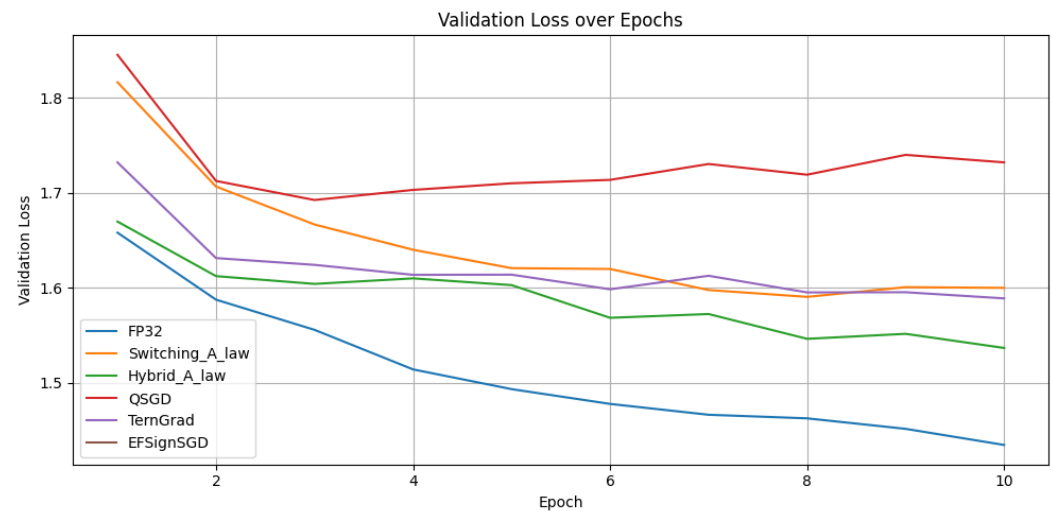


Figure 6. Validation loss evolution on ImageNet.

The quantitative results are presented in Table 4. The proposed Hybrid A-law quantizer achieves a Top-1 accuracy of 63.91% and a Top-5 accuracy of 85.77%, compared with 65.49% and 86.80% obtained using FP32 training. This performance is achieved while reducing the average gradient representation from 32 bits to approximately 8 bits.

Table 4. ImageNet benchmark results.

Method	Top-1	Top-5	RMSE	Bits
FP32	0.6549	0.8680	0.000000	32.0
Switching A-law	0.6211	0.8438	0.001386	7.39
Hybrid A-law	0.6391	0.8577	0.000092	8.00
QSGD	0.5901	0.8245	0.014288	4.00
TernGrad	0.6206	0.8447	0.031971	2.00
EFSignSGD	0.0010	0.0050	—	1.00

The ImageNet results are particularly significant because they demonstrate that the proposed semilogarithmic quantization framework scales beyond small benchmark datasets and remains effective for large-scale visual recognition tasks. Among all compressed methods, Hybrid A-law achieves the most favorable balance between predictive performance and gradient fidelity.

3.6. Overall Discussion

Across all evaluated datasets, three consistent observations can be made. First, both proposed semilogarithmic quantizers substantially reduce gradient precision while preserving predictive performance. Depending on the benchmark, the effective gradient representation is reduced from 32 bits to approximately 7–8 bits, corresponding to compression ratios of roughly $4\times$. In practical distributed training scenarios, such a reduction directly translates into lower communication overhead and reduced memory traffic during gradient synchronization. Second, the Hybrid A-law quantizer consistently achieves the lowest gradient reconstruction error among all compressed methods. On CIFAR-10 and CIFAR-100, the RMSE reduction relative to QSGD exceeds two orders of magnitude, indicating significantly improved preservation of gradient information. Third, the proposed methods remain competitive across fundamentally different learning scenarios,

including tabular classification (DCCC), medium-scale visual recognition (CIFAR-10 and CIFAR-100), and large-scale image classification (ImageNet). These findings support the use of semilogarithmic gradient quantization as a practical and communication-efficient compression mechanism for resource-constrained deep learning systems. A particularly important observation is the significantly lower gradient reconstruction error achieved by the proposed Hybrid A -law quantizer. Across all investigated datasets, the resulting RMSE was between one and three orders of magnitude lower than those produced by QSGD, TernGrad, and EFSignSGD. This indicates that semilogarithmic companding preserves the structure of the original gradients considerably more faithfully than stochastic or sign-based quantization schemes.

Unlike several competing methods whose performance strongly depends on the dataset and network architecture, the proposed Hybrid A -law quantizer exhibited remarkably consistent behavior across all considered learning scenarios. Competitive accuracy was maintained on tabular data, medium-scale image classification tasks, and large-scale ImageNet training without requiring dataset-specific tuning. An additional practical advantage of the proposed approach is that gradient reconstruction remains deterministic. Unlike stochastic quantization methods such as QSGD and TernGrad, the proposed semilogarithmic companding framework introduces no stochastic reconstruction noise, which contributes to the observed reduction in gradient distortion. Although certain baselines occasionally achieve slightly higher peak accuracies on individual datasets, the proposed Hybrid A -law quantizer consistently delivers the most favorable compromise between predictive performance, gradient fidelity, and communication efficiency.

4. Conclusions

This paper presents two adaptive semilogarithmic A -law gradient quantization methods: switching and hybrid piecewise schemes, designed to enable efficient and scalable training of deep neural networks. Both approaches dynamically adjust the companding parameter A based on gradient statistics, facilitating effective precision modulation and nonuniform compression during backpropagation. To address the limitations of a purely instantaneous variance-based switching strategy, the switching quantizer was enhanced through an exponentially weighted moving average (EMA) variance reference, enabling more stable and statistically meaningful adaptation of quantization precision throughout training. Experimental evaluations were conducted on a diverse set of learning tasks, including a tabular credit-risk prediction problem MLP-DCCC, CIFAR-10, CIFAR-100, and large-scale ImageNet image classification benchmarks. In addition to the conventional FP32 baseline, the proposed methods were compared against established gradient compression approaches, including QSGD, TernGrad, and EFSignSGD. The obtained results demonstrate that the proposed semilogarithmic quantization framework achieves competitive predictive performance while substantially reducing gradient reconstruction error. In particular, the Hybrid A -law quantizer consistently produced the lowest gradient distortion among the investigated compressed training schemes, while preserving validation and test accuracies close to full-precision training across multiple datasets and model architectures. Although certain competing methods occasionally achieved slightly higher peak accuracies on individual benchmarks, the proposed Hybrid A -law approach provided the most favorable compromise between predictive performance, gradient fidelity, and communication efficiency. Furthermore, unlike highly compressed sign-based approaches, the proposed framework remained stable across all considered experimental scenarios, including large-scale ImageNet training. The results confirm that structured semilogarithmic quantization offers a viable alternative to full-precision training, particularly in resource-constrained environments and communication-limited training scenarios. Although federated and

distributed learning represent natural application domains for gradient compression, such settings were not explicitly evaluated in the present study and remain part of future work. Although the experimental results demonstrate stable and competitive performance across multiple datasets and neural network architectures, the present study focuses primarily on the design and empirical evaluation of the proposed quantization schemes. A rigorous theoretical convergence analysis remains outside the scope of the current work and will be considered in future research. Future work will focus on establishing theoretical convergence guarantees, extending the framework to transformer-based architectures and large language models, investigating adaptive layer-wise quantization strategies, and developing hardware-aware implementations for next-generation distributed AI systems.

Author Contributions: Conceptualization, S.P., M.D. and M.S. (Milan Savić); methodology, S.P. and M.D.; software, M.D. and M.S. (Marko Smilić); validation, S.P., M.D. and M.S. (Milan Savić); formal analysis, S.P. and M.D.; investigation, S.P., M.D. and M.S. (Milan Savić); resources, S.P., M.D. and M.S. (Marko Smilić); data curation, M.D. and M.S. (Marko Smilić); writing—original draft preparation, S.P. and M.D.; writing—review and editing, S.P., M.D., M.S. (Milan Savić) and M.S. (Marko Smilić); visualization, M.D. and M.S. (Marko Smilić); supervision, S.P. and M.S. (Milan Savić); project administration, S.P. and M.D. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia (Grant numbers: 451-03-34/2026-03).

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DNN	Deep Neural Network
MLP	Multilayer Perceptron
CNN	Convolutional Neural Network
FP32	32-bit Floating Point
FP8	8-bit Floating Point
QAT	Quantization-Aware Training
PTQ	Post-Training Quantization
GPTQ	Gradient-Based Post-Training Quantization
STE	Straight-Through Estimator
SQNR	Signal-to-Quantization-Noise Ratio
RMSE	Root Mean Square Error
ReLU	Rectified Linear Unit
IoT	Internet of Things
AI	Artificial Intelligence

References

1. Gholami, A.; Kim, S.; Dong, Z.; Yao, Z.; Mahoney, M.W.; Keutzer, K. A Survey of Quantization Methods for Efficient Neural Network Inference. In *Low-Power Computer Vision*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2022; pp. 291–326. [CrossRef]
2. *IEEE Std 754-2019*; IEEE Standard for Floating-Point Arithmetic. IEEE: New York, NY, USA, 2019.
3. Zhang, A.; Lipton, Z.C.; Li, M.; Smola, A.J. *Dive into Deep Learning*; Amazon Science: Bellevue, WA, USA, 2020; pp. 5–6. [CrossRef]
4. Tagliavini, G.; Mach, S.; Rossi, D.; Marongiu, A.; Benini, L. A Transprecision Floating-Point Platform for Ultra-Low Power Computing. In Proceedings of the Design, Automation and Test in Europe Conference and Exhibition (DATE), Dresden, Germany, 19–23 March 2018; pp. 1051–1056. [CrossRef]

5. Cattaneo, D.; Di Bello, A.; Cherubin, S.; Terraneo, F.; Agosta, G. Embedded Operating System Optimization through Floating to Fixed Point Compiler Transformation. In Proceedings of the Euromicro Conference on Digital System Design (DSD), Prague, Czech Republic, 15–19 April 2018; pp. 172–176. [\[CrossRef\]](#)
6. Verucchi, M.; Brilli, G.; Sapienza, D.; Verasani, M.; Arena, M.; Gatti, F.; Capotondi, A.; Cavicchioli, R.; Bertogna, M.; Solieri, M. A Systematic Assessment of Embedded Neural Networks for Object Detection. In Proceedings of the IEEE International Conference on Emerging Technologies and Factory Automation (ETFA), Vienna, Austria, 8–11 September 2020; pp. 937–944. [\[CrossRef\]](#)
7. Chen, L.; Lin, S.; Lu, X.; Cao, D.; Wu, H.; Guo, C.; Liu, C.; Wang, F.-Y. Deep Neural Network Based Vehicle and Pedestrian Detection for Autonomous Driving: A Survey. *IEEE Trans. Intell. Transp. Syst.* **2021**, *22*, 3234–3246. [\[CrossRef\]](#)
8. Esser, S.K.; McKinstry, J.L.; Bablani, D.; Appuswamy, R.; Modha, D.S. Learned Step Size Quantization. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 26–30 April 2020. [\[CrossRef\]](#)
9. Bhalgat, Y.; Lee, J.; Nagel, M.; Blankevoort, T.; Kwak, N. LSQ+: Improving Low-Bit Quantization through Learnable Offsets and Better Initialization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 14–19 June 2020; pp. 2978–2985. [\[CrossRef\]](#)
10. Gong, R.; Liu, X.; Jiang, S.; Li, T.; Hu, P.; Lin, J.; Yu, F.; Yan, J. Differentiable Soft Quantization: Bridging Full-Precision and Low-Bit Neural Networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 October 2019; pp. 4851–4860. [\[CrossRef\]](#)
11. Li, Y.; Gong, R.; Tan, X.; Yang, Y.; Hu, P.; Zhang, Q.; Yu, F.; Wang, W.; Gu, S. BRECQ: Pushing the Limit of Post-Training Quantization by Block Reconstruction. *arXiv* **2021**, arXiv:2102.05426. [\[CrossRef\]](#)
12. Wei, X.; Gong, R.; Li, Y.; Liu, X.; Yu, F. QDrop: Randomly Dropping Quantization for Extremely Low-Bit Post-Training Quantization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 4224–4233. [\[CrossRef\]](#)
13. Alistarh, D.; Grubic, D.; Li, J.; Tomioka, R.; Vojnovic, M. QSGD: Communication-Efficient SGD via Gradient Quantization and Encoding. *arXiv* **2016**, arXiv:1610.02132. [\[CrossRef\]](#)
14. Wen, W.; Xu, C.; Yan, F.; Wu, C.; Wang, Y.; Chen, Y.; Li, H. TernGrad: Ternary Gradients to Reduce Communication in Distributed Deep Learning. *arXiv* **2017**, arXiv:1705.07878. [\[CrossRef\]](#)
15. Bernstein, J.; Zhao, J.; Azzadenesheli, K.; Anandkumar, A. signSGD with Majority Vote is Communication Efficient And Fault Tolerant. *arXiv* **2018**, arXiv:1810.05291. [\[CrossRef\]](#)
16. Kairouz, P.; McMahan, H.B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A.N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; et al. Advances and Open Problems in Federated Learning. *arXiv* **2019**, arXiv:1912.04977. [\[CrossRef\]](#)
17. Jayant, N.S.; Noll, P. *Digital Coding of Waveforms: Principles and Applications to Speech and Video*; Prentice-Hall: Englewood Cliffs, NJ, USA, 1984; pp. 26–84.
18. Sayood, K. *Introduction to Data Compression*, 5th ed.; Morgan Kaufmann: Cambridge, MA, USA, 2017; pp. 253–257.
19. Perić, Z.H.; Mosić, A.V.; Panić, S.R. Coding Algorithm Based on Loss Compression Using Scalar Quantization Switching Technique. *J. Inf. Sci. Eng.* **2010**, *26*, 967–976.
20. Savić, M.S.; Perić, Z.H.; Panić, S.R.; Mosić, A.V. Semi-Logarithmic and Hybrid Quantization of Laplacian Source in Wide Range of Variances. *J. Electr. Eng.* **2012**, *63*, 386–391. [\[CrossRef\]](#)
21. Mosić, A.V.; Perić, Z.H.; Panić, S.R. Switched Nonuniform and Piecewise Uniform Scalar Quantization of Laplacian Source. *Int. J. Comput. Commun. Control* **2012**, *7*, 115–122. [\[CrossRef\]](#)
22. Perić, Z.H.; Mosić, A.V.; Panić, S.R. Robust and Switched Nonuniform Scalar Quantization of Gaussian Source. *Autom. Remote Control* **2008**, *69*, 334–341. [\[CrossRef\]](#)
23. Perić, Z.H.; Dinčić, M.; Denić, D.; Jocić, A. Forward Adaptive Logarithmic Quantizer with New Lossless Coding Method. *Wirel. Pers. Commun.* **2011**, *59*, 625–641. [\[CrossRef\]](#)
24. Banner, R.; Nahshan, Y.; Hoffer, E.; Soudry, D. Post-Training 4-Bit Quantization of Convolution Networks for Rapid-Deployment. *arXiv* **2018**, arXiv:1810.05723. [\[CrossRef\]](#)
25. Micikevicius, P.; Narang, S.; Alben, J.; Diamos, G.; Elsen, E.; Garcia, D.; Ginsburg, B.; Houston, M.; Kuchaiev, O.; Venkatesh, G.; et al. Mixed Precision Training. In Proceedings of the International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018. [\[CrossRef\]](#)
26. Yan, L.; Zhao, T.; Xie, X.; Precup, R.E. OSSEFS: An Online Semi-Supervised Ensemble Fuzzy System for Data Streams Learning with Missing Values. *Expert Syst. Appl.* **2024**, *255*, 124695. [\[CrossRef\]](#)
27. Ning, J. Neural Network-Based Pattern Recognition in the Framework of Edge Computing. *Rom. J. Inf. Sci. Technol.* **2024**, *27*, 106–119. [\[CrossRef\]](#)

28. Colbu, S.; Bancila, D.M.; Popescu, D. Long-Term Power Generation Prediction in Photovoltaics Using Machine Learning-Based Models. *Rom. J. Inf. Sci. Technol.* **2025**, *28*, 39–50. [[CrossRef](#)]
29. Dubljanin, M.; Panić, S.; Savić, M.; Milosavljević, S. Adaptive μ -Law Gradient Quantization for Training MLPs and CNNs. *Adv. Electr. Comput. Eng.* **2026**, *26*, 55–64. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.