Insights in
Computer Science

https://acadpub.com/ics

Impact of Fixed-Point Weight Quantization Bit-Width on MNIST Classification Accuracy

Ali Siddique

Department of Computer Science, University of Manchester, Manchester- M13 9PL, United Kingdom

About the Article

Open Access

Systematic Review

How to Cite: A. Siddique, "Impact of fixed-point weight quantization bit-width on MNIST classification accuracy," *Insights Comput. Sci.*, Vol. 1, pp. 10-14. 2025.

Keywords:
ASIC, Edge AI, fixed-point, FPGA, MNIST, quantization, word length

Corresponding author:
Ali Siddique,
Department of Computer Science, University of Manchester, Manchester- M13 9PL, United Kingdom

ABSTRACT

Background and Objective: This systematic review examines how fixed-point weight bit-width influences the classification accuracy of convolutional neural networks deployed in edge and embedded systems.

Materials and Methods: Studies evaluating uniform min–max quantization of weights across 1-32 bits were reviewed, focusing on work that isolates weight precision while keeping activations in float32 and maintaining consistent network architecture, training and evaluation procedures. Research relevant to Field Programmable Gate Array (FPGA) and Application-Specific Integrated Circuit (ASIC) implementations was prioritised.

Results: Across the literature, weight precisions of 5-6 bits consistently provide a strong balance between accuracy and hardware efficiency for MNIST-level tasks. Accuracy deteriorates below 5 bits, while higher precisions offer negligible gains relative to increased resource use.

Conclusion: Fixed-point weight bit-width is a key parameter for efficient CNN deployment in constrained hardware environments. Simple word-length sweeps offer practical guidance for selecting precision and complement existing work on hardware-centric neural network design.

INTRODUCTION

Modern deep neural networks deliver high accuracy across a wide range of applications, including image classification, disease diagnosis, neuromorphic sensing and smart agriculture[1,2]. However, this performance often requires substantial computational and memory resources, which are difficult to accommodate in edge and embedded platforms where area, latency and power budgets are highly constrained. To address these limitations, an expanding body of research explores dedicated hardware solutions[3-5].

Siddique *et al.*[6] have developed several FPGA-based neuromorphic and deep-learning architectures targeting spiking neural networks and low-cost accelerators. These include a Tempotron-based neuromorphic computer enabling high-throughput online SNN learning with low synaptic overhead, a supervised SNN engine based on the HaSiST scheme and specialised systems such as SpikoPoniC and N-AquaRAM for real-time aquaponics monitoring[6-9]. These platforms estimate fish length and weight or monitor aquaponic conditions while consuming only modest FPGA resources. Additionally, a hardware-based deep-learning system for disease diagnosis demonstrates that carefully chosen activation functions and low-cost hardware structures can achieve high accuracy on Virtex-6 devices[10].

Another line of work by Siddique *et al.*[11] presents a 218-GOPS accelerator employing a cost-efficient surrogate-gradient method to address the dying-ReLU problem. In digital agriculture, user-centred systems such as AgFAB illustrate how efficient edge inference can integrate seamlessly with farmer-facing mobile interfaces to support practical field deployment[12].

Despite their diverse applications, these works share several common features: they optimise neuron models, learning rules and dataflow for hardware efficiency and they frequently rely on reduced-precision arithmetic.

ACADEMIA Publications

Nevertheless, most studies examine only a limited set of bit-width choices for weights and activations, fixing these values without systematically evaluating their impact on performance. Comprehensive analyses of how classification accuracy varies with weight word length-particularly on standard benchmarks-remain comparatively rare.

This study addresses that gap by isolating a single parameter: The weight word length in a fixed-point representation. We quantify its direct effect on classification accuracy in a standard MNIST convolutional neural network, with an emphasis on clarity, reproducibility and practical relevance rather than maximising accuracy or hardware throughput. The findings provide a straightforward reference for selecting weight precision in more complex neuromorphic and accelerator architectures.

The main contributions of this work are as follows:

- A controlled post-training quantization study in which only the weight bit-width is varied from 1 to 32 bits, while all training and evaluation conditions remain fixed.
- An empirical demonstration that accuracy saturates above 8 bits, whereas 3-4 bits offer a practical balance between accuracy and storage efficiency for MNIST-level tasks
- A concise discussion relating these findings to hardware-oriented design decisions in low-cost neuromorphic and deep-learning accelerators, where bit-width directly influences memory usage, bandwidth requirements and computational cost

Hardware oriented neural and neuromorphic systems: Several studies conducted by Siddique *et al.*[6-8] focus on the development of low-cost, high-throughput neuromorphic and deep-learning systems that are directly relevant to fixed-point design. A neuromorphic computer based on the Tempotron learning rule and population coding achieves an approximate 15× speedup on a general-purpose device and processes millions of samples per second on a Virtex-6 FPGA while maintaining low synaptic hardware cost[6]. The NME-HaSiST system introduces a supervised SNN backpropagation scheme that avoids computationally expensive operations such as error normalization and weight-threshold balancing, achieving ~97.5% accuracy on MNIST using 158,800 synapses[7]. Its corresponding inference engine employs a hard-sigmoid-based training scheme and attains giga-synaptic-operations-per-second performance with very low slice-register and LUT utilization per synapse.

SpikoPoniC and N-AquaRAM extend these principles to smart agriculture applications. SpikoPoniC uses SNNs for fish length and weight estimation in aquaponic systems, achieving over 84 million classifications per second with fewer than 1.1k slice registers[8]. N-AquaRAM implements a hardware-efficient smooth activation function in a neuromorphic accelerator for aquaponic monitoring, reaching approximately 40 million classifications per second with only a few thousand slice registers[9]. A related deep-learning hardware design for disease diagnosis employs cost-efficient activation functions to achieve ~98.23% accuracy on medical datasets while remaining significantly more affordable than many specialized hardware platforms[10].

The 218-GOPS neural accelerator introduces a surrogate-gradient mechanism to mitigate gradient vanishing and the dying-ReLU problem, enabling the use of ReLU across all network layers[11]. This accelerator reaches ~98.39% accuracy on MNIST with fewer than 159k synapses and delivers approximately 218 GOPS on a low-end Virtex-6 device with low per-synapse hardware cost.

Collectively, these systems demonstrate that activation-function design, learning rules and dataflow must be tailored to the underlying hardware platform. They also highlight bit-width as a critical design parameter; however, none of these studies systematically isolates and analyses the accuracy-bit-width trade-off. The present work addresses this targeted but practically important gap by evaluating the impact of fixed-point weight precision on the accuracy of a simple CNN.

Human centred digital agriculture and edge intelligence: The AgFAB study focuses on a farmer-centred digital agriculture platform designed to support smallholder farmers in developing countries[12]. Using human-centred design principles, the authors identify key user requirements for mobile and computing applications and evaluate prototype interfaces using the System Usability Scale (SUS). The AgFAB prototype achieved an average SUS score of 72.37, indicating an acceptable and user-friendly design and a paired t-test further suggested strong potential for real-world adoption.

Although, AgFAB does not directly address neural network hardware, it provides important context for environments in which lightweight fixed-point models may eventually be deployed. Modern digital agriculture systems increasingly integrate user-facing tools with on-device sensing, classification and predictive analytics. In such scenarios, fixed-point neural accelerators-such as those used in SpikoPoniC, N-AquaRAM, or the disease-diagnosis engine-could process sensor data efficiently, while users interact with the system through accessible interfaces like AgFAB.

Viewed in this broader context, fixed-point word-length analyses such as the present study contribute a foundational component to a larger design space that spans hardware efficiency, computational constraints and user-experience considerations.

MATERIALS AND METHODS

Network and dataset: The classification model employed in this study is a compact convolutional neural network composed of two convolutional layers followed by two fully connected layers. The first convolutional layer processes the single-channel MNIST input and generates 32 feature maps using 3×3 kernels. The second convolutional layer increases the representation to 64 feature maps, also with 3×3 kernels. A 2×2 max-pooling operation is applied to reduce the spatial resolution. The resulting feature maps are flattened into a 9,216-dimensional vector and passed to a fully connected layer comprising 128 ReLU-activated units. A final fully connected layer produces 10 output logits corresponding to the MNIST digit classes.

The dataset used is the standard MNIST handwritten digit benchmark, consisting of 60,000 training images and 10,000 test images. All images are 28×28-pixel greyscale samples and input normalization is performed by scaling pixel intensities to the range [0, 1] via division by 255.

Training protocol: The baseline model is trained in float32 with exactly the same hyper-parameters as the reference script:

- **Optimizer:** Adam (weight decay = 0)
- **Learning rate:** 1×10^{-3}
- **Batch size:** 64
- **Epochs:** 3
- **Normalization:** Mean = 0.1307, std = 0.3081 (standard MNIST)

Quantization setup: Only the network weights are quantized after training; all activations and intermediate computations remain in full float32 precision. For a weight tensor x and a target bit-width b (where b<32), uniform min–max quantization is applied as follows:

$$q_{levels} = 2^b$$

$$\text{Scale} = \frac{x_{max} - x_{min}}{q_{levels} - 1 + 10^{-8}}$$

$$x_q = \frac{\text{Round}\left(x - x_{min}\right)}{\text{Scale}}$$

$$x = x_q \times \text{scale} + x_{min}$$

The quantized weights are written back into the original float32 model, which is then evaluated directly. In this dequantized configuration, inference incurs no quantization error on the weights. Although no hardware prototype is implemented in this study, we define two hardware-relevant metrics to contextualize the results:

Model size (weights only): The memory required to store all network weights at a given bit-width, excluding minor overheads such as scale parameters.

Relative bandwidth proxy: A quantity proportional to model size, under the assumption that weight transfers dominate traffic during model loading or update.

Both metrics scale linearly with the number of bits per weight, providing a straightforward estimate of storage and bandwidth savings: Reducing bit-width directly reduces memory footprint and data-movement cost.

RESULTS AND DISCUSSION

Figure 1 illustrates the effect of weight quantization bit-width on classification accuracy. The results indicate that a precision of 3-4 bits provide an effective compromise between accuracy and storage efficiency, whereas precision above 8 bits yields no measurable improvement, as accuracy has already saturated.

Overall, the findings support the following guideline for MNIST-scale classification tasks:

- 8 bits: Negligible accuracy loss relative to float32
- 5-6 bits: Moderate degradation (approximately 1-5% points, depending on the run)
- ≤4 bits: Substantial accuracy deterioration

These observations align with prior research on hardware-efficient neural and neuromorphic systems by Siddique *et al.*[11]. The 218 GOPS accelerator demonstrates that the selection of activation functions and surrogate gradient strategies can substantially enhance hardware efficiency in ReLU-based networks without compromising accuracy[11]. Likewise, Tempotron-based neuromorphic computing, the HaSiST engine, SpikoPoniC, N-AquaRAM and FPGA-based diagnostic systems all rely on reduced-precision arithmetic to achieve high throughput and efficient resource utilization[6-10].

The present results can inform early-stage precision selection in neuromorphic accelerator design. For example, in systems with approximately 155 k synapses for MNIST, reducing weight precision from 8 bits to 6 bits can decrease synaptic storage by roughly one quarter and reduce memory bandwidth requirements. The corresponding reduction in accuracy is generally small and may be further mitigated through limited quantization-aware fine-tuning.

Neuromorphic platforms employing Tempotron learning, supervised SNN architectures such as NME-HaSiST and application-specific accelerators including SpikoPoniC and N-AquaRAM, as well as medical diagnostic accelerators, all depend on efficient arithmetic and carefully selected bit-widths[8-10]. The 218 GOPS accelerator further underscores the importance of optimized

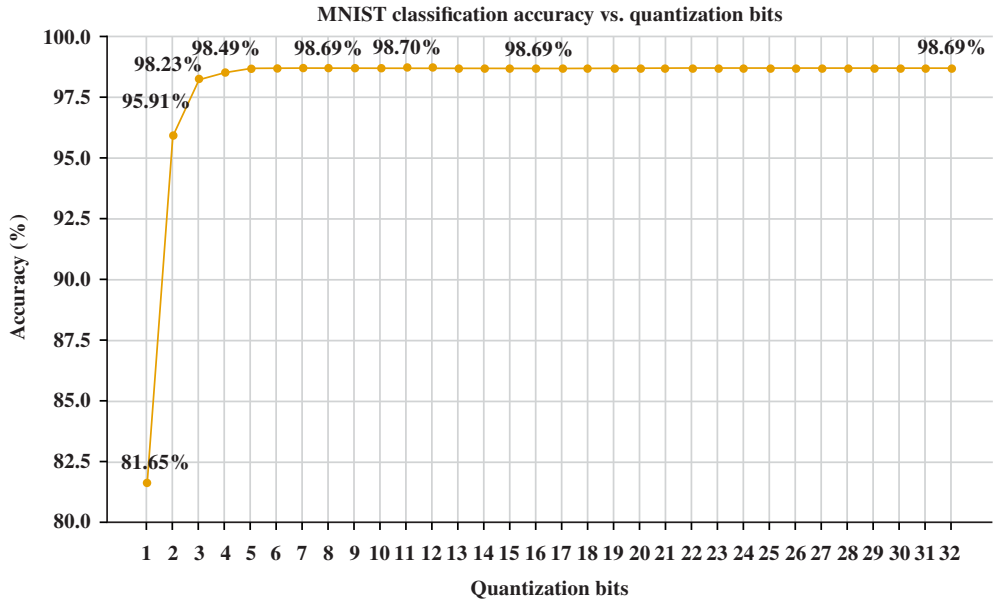

Fig. 1: Impact of weight size on MNIST classification accuracy

activation and gradient design[11]. In parallel with these developments, a concise and transparent word-length analysis such as the present study can serve as a practical tool for guiding precision choices prior to more complex hardware–algorithm co-design stages. These insights are also relevant to human-centered digital agriculture platforms such as AgFAB, which emphasize usability and cost-efficiency for smallholder farming applications[12].

CONCLUSION

This study systematically examined the influence of fixed-point weight quantization on the classification accuracy of a convolutional neural network applied to the MNIST dataset. By varying the weight bit-width from 1 to 32 bits while keeping all other architectural and training parameters constant, the results demonstrate a clear trade-off between precision and model performance. Accuracy remains effectively unchanged relative to float32 when using 8 bits or more and remains largely stable at 6 bits, with only moderate degradation observed at 5-6 bits. In contrast, accuracy declines markedly at 4 bits and collapses rapidly below this threshold. These findings indicate that 3-4 bits represent the lower limit at which a meaningful balance between storage efficiency and acceptable accuracy can still be achieved.

The results reinforce the importance of bit-width selection as a simple yet powerful design parameter for edge-oriented hardware accelerators. Reducing weight precision for example, from 8 bits to 6 bits can yield substantial savings in memory footprint and bandwidth (e.g., approximately one-quarter reduction for a 155k-synapse neuromorphic accelerator), with minimal accuracy loss, particularly when paired with quantization-aware fine-tuning. Although deliberately focused on a compact CNN, this empirical sweep aligns with broader research into hardware-efficient neural network design.

Future work should extend this analysis to larger convolutional networks, spiking neural networks and implementations more tightly coupled to FPGA or ASIC platforms. Overall, the study highlights that weight bit-width remains a critical, tunable parameter for balancing accuracy, computational cost and storage efficiency and merits explicit consideration alongside more prominent algorithmic innovations.

REFERENCES

[1] G. Rutishauser, J. Mihali, M. Scherer and L. Bonini, "xTern: Energy-efficient ternary neural network inference on RISC-V-based edge systems," in *2024 IEEE 35th International Conference on Application-specific Systems, Architectures and Processors (ASAP)*, Jul. 2024. Hong Kong: IEEE, 2024, pp. 206-213.

[2] I. Westby, X. Yang, T. Liu and H. Xu, "FPGA acceleration on a multi-layer perceptron neural network for digit recognition," *J. Supercomput.*, vol. 77, pp. 14356-14373. 2021.

[3] M. Molina, J. Mendez, D. P. Morales, E. Castillo, M. L. Vallejo and M. Pegalajar, "Power-efficient implementation of ternary neural networks in edge devices," *IEEE. Internet Things J.*, vol. 9, no. 20, pp. 20111-20121. 2022.

[4] N. Zheng and P. Mazumder, "A low-power hardware architecture for on-line supervised learning in multi-layer spiking neural networks," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, May. 2018. Florence, Italy: IEEE, 2018, pp. 1-5.

[5] C. Sun *et al*, "An energy efficient STDP-based SNN architecture with on-chip learning," *IEEE. Trans. Circuits Syst. I: Regul. Pap.*, vol. 69, no. 12, pp. 5147-5158. 2022.

[6] A. Siddique, M. I. Vai and S. H. Pun, "A low-cost, high-throughput neuromorphic computer for online SNN learning," *Cluster Comput.*, vol. 27, pp. 2447-2464. 2023.

[7] A. Siddique, M. I. Vai and S. H. Pun, "A low cost neuromorphic learning engine based on a high performance supervised SNN learning algorithm," *Sci. Rep.*, vol. 13, 2023. [Online]. Available: 10.1038/s41598-023-32120-7

[8] A. Siddique, J. Sun, K. J. Hou, M. I. Vai, S. H. Pun and M. A. Iqbal, "SpikoPoniC: A low-cost spiking neuromorphic computer for smart aquaponics," *Agric.*, vol. 13, no. 11, 2023. [Online]. Available: 10.3390/agriculture13112057

[9] A. Siddique, M. A. Iqbal, J. Sun, X. Zhang, M. I. Vai and S. Siddique, "N-AquaRAM: A cost-efficient deep learning accelerator for real-time aquaponic monitoring," *Agric. Res.*, vol. 14, pp. 591-604. 2024.

[10] A. Siddique, M. A. Iqbal, M. Aleem and J. C.-W. Lin, "A high-performance, hardware-based deep learning system for disease diagnosis," *PeerJ Comput. Sci.*, vol. 8, 2022. [Online]. Available: 10.7717/peerj-cs.1034

[11] A. Siddique, M. A. Iqbal, M. Aleem and M. A. Islam, "A 218 GOPS neural network accelerator based on a novel cost-efficient surrogate gradient scheme for pattern classification," *Microprocess. Microsyst.*, vol. 99, 2023. [Online]. Available: 10.1016/j.micpro.2023.104831

[12] M. A. Iqbal, B. B. Posadas, F. Qin, B. Liu and A. Siddique, "AgFAB - A farmer-centered agricultural bower," *EAI Endorsed Trans. Ind. Networks Intell. Syst.*, vol. 10, no. 1, 2023. [Online]. Available: 10.4108/eetinis.v10i1.2714