Grenze International Journal of Engineering and Technology, June Issue

Q-Net Compressor: Deep Learning Model Compression with Quantization

Yash Bansal[1], Sejal Garg[2], Kartik Saroha[3] and Manu Singh[4]
[1-4]Department of Computer Science, ABES Engineering College, Ghaziabad, UP, India
Email: yashbansal81003@gmail.com, {sejalgarg1306, kartiksaroha1112, drmanuajaykumar}@gmail.com

***Abstract*— The rapid advancement of deep learning has brought about remarkable achievements in various fields, particularly by utilizing Deep Neural Networks.**
It offers a high rate of accuracy, extensive parameters, and intensive computational requirements that often characterize these networks. Due to high memory consumption and energy usage, this results in significant challenges when deploying DNNs on hardware-constrained devices. Researchers have extensively explored various compression techniques (Al-Qurabat 2022) to address these challenges, quantifying the merging as a promising solution. However, the extensive use of quantization in previous works necessitates a comprehensive survey that provides a complete study, analysis, and comparison of various quantization approaches.
This paper introduces the "Q-Net Compressor," a novel solution for installing substantial neural (Tao, Hou, et al. 2022) network training on constrained resources edge devices. Utilizing adaptive precision levels to maintain model functionality while reducing memory and computation demands. The Q-Net Compressor, implemented with TensorFlow and PyTorch, offers flexibility across diverse deep-learning architectures, striking a balance between significant model size reduction and the preservation of predictive power. This innovative approach contributes to the evolving deep learning model compression field with quantization, bridging the gap between computational demands and resource constraints.
We hope to address this gap in this survey study by providing an integrative report that investigates the ideas and methods of quantization, (Hu and Wen 2021) with a particular emphasis on its use in image classification problems.

***Index Terms*— Deep Learning, Quantization Compression, Adaptive Quantization, Uniform Quantization, etc.**

I. INTRODUCTION

In recent years, the escalating complexity and size of neural networks (NNs) have pushed deep learning into unprecedented realms of accuracy and performance (Mary Shanthi Rani, Chitra et al. 2022). The deployment of these over-parameterized models in resource-constrained environments poses a significant challenge. Realizing the transformative potential of deep learning in applications (Mishra and Gupta 2023) such as intelligent healthcare monitoring, autonomous driving, and speech recognition necessitates a shift towards efficient, real-time inference that balances low energy consumption with high accuracy.
This paper addresses a pivotal aspect of mitigating the challenges posed by large NNs—quantization. As NNs have evolved to possess tens to hundreds of millions of parameters, the precision with which these parameters are represented has become a critical consideration. Quantization, defined as the reduction of precision in representing

Grenze ID: 01.GIJET.10.2.57

neural network parameters, emerges as a compelling solution. Despite its promise, quantization introduces a critical trade-off between precision and accuracy.

The significance of NN quantization is underscored by its historical roots in the 1990s, driven by the need to expedite training times on available machines. In the current landscape, the resurgence of interest in NN quantization arises from the imperative to deploy increasingly complex NNs in real-world hardware, such as edge devices. This paper embarks on a survey of the diverse techniques (Pei, Yao et al. 2023) that have emerged in the last decade to address the challenges associated with NN quantization.

This innovative approach addresses the intricacies of training DNN models with quantization (Oh, Sim et al. 2022) by incorporating advanced optimization algorithms, including stochastic gradient descent. Through iterative precision optimization guided by the loss function gradient, the Q-Net Compressor strikes an optimal balance between significant model size reduction and the preservation of predictive power.

A. TYPES OF QUANTIZATION

Quantization comes in various forms, each serving different purposes in different fields. Here are some common types of quantization and summaries of each:

Scalar Quantization: Scalar quantization (Bradley, Brislawn, and Hopper 1993) is a fundamental process in signal processing, particularly in the analog-to-digital conversion and data compression domains.

Vector Quantization: Vector quantization (VQ) is a signal processing technique that extends the principles (Rahebi 2022) of scalar quantization to groups of samples or vectors.

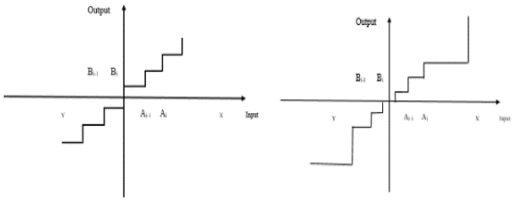

Figure 1. Comparison between Scalar and Vector Quantization

Uniform Quantization: Uniform quantization is a straightforward method where the range of possible input values is divided into a fixed number of equally spaced intervals, and each interval is represented by a unique quantization level. This technique is commonly applied in analog-to-digital conversion (ADC) and digital-to-analog conversion (DAC) processes.

Non-Uniform Quantization: On the contrary, non-uniform quantization allows for different step sizes between quantized levels. This means that the intervals between adjacent quantization levels can be adjusted based on the characteristics of the incoming signal (Liu, Cheng et al. 2022). Non-uniform quantization is especially useful when the input signal has non-uniform features, it comes at the cost of introducing quantization errors, especially when dealing with low bit-depth quantization.

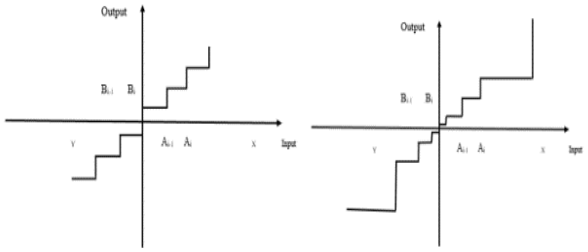

Figure 2. Comparison between Uniform and Non-Uniform Quantization.

B. INTRODUCTION OF QUANTIZATION USING FINE-TUNING

Post-training Quantization (PTQ) and Quantization Aware Training (QAT): Quantization fine-tuning is a critical step in the deployment of neural networks that have undergone the quantization process to reduce precision for efficient inference on resource-constrained devices. Two prominent approaches in this fine-tuning process are Post Training Quantization (PTQ) and Quantization Aware Training (QAT) (Finkelstein, Fuchs et al. 2022).

Post-training Quantization (PTQ): It is a technique used to adapt pre-trained neural network models to lower precision after the initial training phase (Oh, Sim et al. 2022). During the initial training, models are often trained with high precision, such as 32-bit floating-point numbers, to achieve maximum accuracy.
Quantization Aware Training (QAT): It is an approach that integrates quantization considerations directly into the model training process (Kirtas, Oikonomou et al. 2022). Unlike PTQ, which applies quantization after training, QAT simulates quantization effects during the training phase itself. This allows the model to learn representations that are more robust to lower precision.
In summary, both PTQ and QAT are valuable techniques for quantization fine-tuning, and addressing. In the deployment of artificial neural networks on resource devices, there is a compromise between model accuracy and computing efficiency.

C. Q-NET COMPRESSOR CASE STUDY

With the evolution of deep learning model compression techniques, contemporary approaches have emerged to address the challenges (Nori, Ge, and Kim 2023) posed by deploying models on resource-constrained edge devices. One such innovative method is the Q-Net Compressor, (Xiao, Lin, et al. 2023) a quantization approach designed to establish a compromise between model size reduction and accuracy preservation.

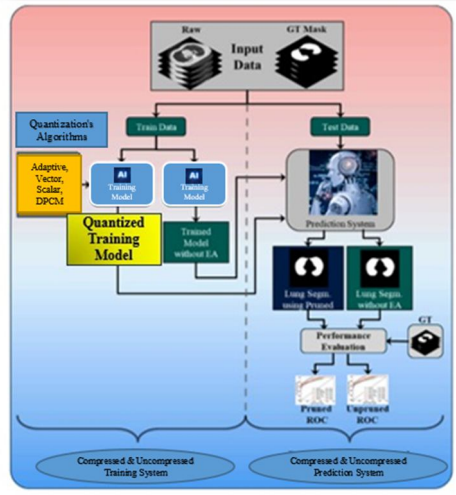

Figure 3. Diagrammatic representation of Quantization Model set

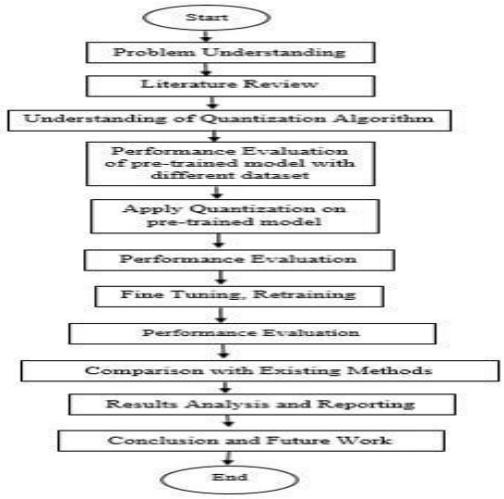

Figure 4. Flowchart representation of Quantization Model set.

The goal of the project is to create the Q-Net Compressor, a deep-learning model compression tool designed for edge devices with limited analysis of the literature on the optimization quantization, and compression issues related to installing DNNs on edge devices for deep-learning models (Inés, Díaz-Pinto et al. 2023). Pre-processing is done on a variety of datasets (Tian, Luan, and Zheng 2023), such as MNIST, CIFAR, and SVHN, to make sure they function with the frameworks TensorFlow and PyTorch. Adaptive quantization and gradient-guided accuracy optimization are the main goals of the algorithm design, which places a strong emphasis on modular architecture for flexibility.
Based on the algorithm design, the Q-Net Compressor is constructed and has an easy-to-use interface for configuration and integration into pre-existing deep learning pipelines. The system is rigorously trained and evaluated on many DNN models, with its complexity gradually rising. The next steps involve optimization and fine-tuning, utilizing evaluation results to achieve the best possible balance between accuracy preservation and model size reduction.
Comprehensive documentation and a user guide are developed, detailing methodology, algorithms, and implementation. The Q-Net Compressor is tested on edge devices, such as mobile phones and embedded systems, and real-world performance is analyzed. Results are thoroughly examined, and a technical report is prepared, emphasizing the Q-Net Compressor's advantages in addressing edge device deployment challenges.

II. LITERATURE REVIEW

To cultivate a comprehensive understanding of quantization techniques and their impact on deep neural networks, this survey paper meticulously reviews and synthesizes insights from a diverse array of research works. Noteworthy studies, including “Adaptive Quantization of Neural Networks” (Khoram and Li 2018), “Model

Compression for Deep Neural Networks: A Survey" (Li, Li and Meng 2023), and "Deep Network Quantization via Error Compensation," contribute to the overarching discourse.

"Adaptive Quantization of Neural Networks" by (Khoram and Li 2018), presented at ICLR, employs a trust region algorithm to compress neural networks by assigning unique precision to each parameter based on importance.

"Model Compression for Deep Neural Networks: A Survey" by (Li, Li, and Meng 2023) delves into various compression techniques, including model pruning, parameter quantization, low-rank decomposition, knowledge distillation, and lightweight model design. Despite discussing the pros and cons of these techniques, the study identifies gaps in the absence of a unified framework and the quantification of trade-offs.

In the realm of hardware implementation, the "FPGA-Based Vehicle Detection and Tracking Accelerator" (Zhai, Li et al. 2023) implements a low-power FPGA-based vehicle detector and tracker, achieving higher accuracy and throughput. However, gaps exist in the robustness of the proposed solution to environmental conditions.

Another study, "Comprehensive SNN Compression Using ADMM Optimization and Activity Regularization" by (Deng, Wu et al. 2021), addresses challenges in Spiking Neural Network (SNN) compression with ADMM optimization and activity regularization. Notably, gaps include considerations for real-world deployment, comparisons with existing methods, and generalizability.

These studies, alongside others surveyed, contribute diverse perspectives to the understanding of model compression techniques, spanning areas such as hardware acceleration, pruning methods (De Leon and Atienza 2022), and innovative approaches like spiking neural networks (Deng, Wu et al. 2021) and error compensation. This literature review sets the stage for a comprehensive exploration (Deng, Wu et al. 2021) of the state of the art in deep learning model compression.

TABLE I. SUMMARISATION

S.NO	Title	Author	Year	Journal/ Conference	Technology/ Algorithm	Features	Gap
01.	Adaptive Quantization of Neural Network	(Khoram and Li 2018)	2018	ICLR	Trust region	It compresses neural netwrks by assigning unique precision.	Didn't compare methods to other Quantization technique.
02.	Deep Learning Model Compression Technique	(Msuya and Maisen 2023)	2023	Tanzania Journal of Engineering and Technology	Various Compression Technique	Model compression enhances efficiency with accuracy loss.	Did not discuss the challengesof compression in DL Models.
03.	Model Compression and Accerlation in DNN	(Cheng, Wang et al. 2018)	2018	IEEE Signal Processing Magazine	Pruning Parameters, Knowledge Distillation.	Review with DNN size and speed optimization.	It doesn't have trade-offs and challenges in real world.
04.	Model Compression for DNN	(Li, and Meng 2023)	2023	Computers	Lightweight model design, Low rank factorization, Knowledge Distillation.	Discusses model compression techniques with pros and cons.	Gaps include the lack of a unified framework and quantization of trade-offs.
05.	FPGA-Based vehicles Detection and Tracking	(Zhai, Li et al. 2023)	2023	Sensors	Neural Network, Hardware accerlator, Vehicle Detection.	It achieves high accuracy and throughput.	It is not a robust to the efficient environmental conditions.

III. Q-NET COMPRESSOR COMPARISON STUDY ON DIFFERENT ALGORITHMS:

In a Q-NET compressor comparison study, various quantization algorithms are evaluated to assess their effectiveness in compressing neural network models. This study aims to analyze and compare the performance of different quantization methods in terms of model size reduction, inference speed, and accuracy preservation. By examining the impact of each algorithm, researchers can identify the most suitable compression technique based on specific requirements and trade-offs in neural network applications.

TABLE II. STUDY ON DIFFERENT ALGORITHMS

S.NO	Method / Algorithm	Q-Net Compressor	Comparison Criteria
01.	Adaptive Quantization of Neural Network	Implements Adaptive Precision Optimization.	It dynamically optimizes the precision ratio of each parameter based on its importance.
02.	Model Compression for Deep Neural Networks	It utilizes a flexible architecture and adaptive quantization.	It stands out in the landscape of model compression techniques through its utilization of a flexible architecture.
03.	SNN Compression Using ADMM Optimization	Addresses Challanges in Sniking Neural Network (SNN) compression with ADMM optimization.	It serves as a beacon in shedding light on practical application, providing valuable insight.
04.	Quantization via Distillation and Constrastive Learning	Q-Net as a pratical demonstration.	It takes centre stage offering tangible insight into real world applications.

TABLE III. COMPARISON STUDY

Criteria	Q-Net Compressor	Classic Quantization	Uniform Quantization	Vector Quantization
Quantization Efficiency	✓	x	x	x
Algorithmic Simplicity	x	✓	✓	x
Memory Utilisation	✓	✓	x	x
Support for Adaptive Quantization	✓	x	x	x

IV. STRENGTH, LIMITATION, AND FUTURE SCOPE OF Q-NET COMPRESSOR

TABLE IV. STRENGTH, LIMITATION, AND FUTURE SCOPE

Aspect	Description
Strengths	● **Adaptability**: Q-Net stands out with its exceptional adaptability across diverse deep-learning architecture. This inherent versatility makes it an ideal solution for a myriad of applications that demand flexibility and efficiency simultaneously. ● **Precision Optimization:** One of Q-Net's key lies in its ability to dynamically optimize precision levels. Through an intricate process guided by parameters' importance, Q-Net strikes a delicate balance between achieving significant model size reduction and preserving predictive power
Limitations	● **Fine Tuning Requirement:** Q-Net may require additional fine-tuning for optimal performance on specific datasets, introducing an extra step in the deployment process. While fine-tuning is a common practice, minimizing its necessity could streamline the implementation of Q-Net in various use cases.
Future Scope	● **Large-Scale Training:** Exploring the implications and strategies for training Q-Net on large-scale datasets is crucial for ensuring its robustness and scalability in real-world applications

Q-NET Compressor exhibits several strengths that make it a valuable tool in the field of model compression. Its primary strength lies in its ability to significantly reduce the model size without compromising much on accuracy,

making it suitable for deployment in resource-constrained environments, such as edge devices and mobile applications.

V. Q-Net Compressor Applications

Q-NET Compressor is a versatile tool with applications across various domains, offering efficient solutions for model compression through quantization. In the field of computer vision, Q-Net Compressor proves valuable by reducing the memory footprint and computational requirements of deep neural networks, enabling seamless deployment on resource-constrained devices such as edge devices and mobile platforms.

Table V. Various Applications Of Q-Net Compressor

Domain	Application
Mobile and Edge Devices	● **Smartphones:** Compressing DNN models for image recognition and natural language processing on smartphones enables faster, on-device inference, reducing latency.
Healthcare	● **Medical Imaging:** Q-Net enables the deployment of compressed DNNs on medical imaging devices, enhancing the speed and accuracy of diagnoses for improved healthcare accessibility.
Natural Language Processing	● **Voice Assistants:** Compressed language models on smart speakers and voice assistants powered by Q-Net enable natural language understanding and generation while conserving device resources.
Industrial Automation	● **Robotics:** Compact DNN models enhanced by Q-Net contribute to the performance of robots by enabling real-time perception and decision-making in applications such as warehouse automation and logistics.
Autonomous Vehicles	● **Object Detection:** Q-Net is employed for compressing models used in object detection systems for autonomous vehicles, ensuring efficient and real-time processing of sensor data.

VI. Conclusion

In conclusion, this survey paper illuminates the evolving discourse on efficient deep learning deployment by navigating the dynamic intersection of precision, efficiency, and accuracy in neural network quantization. The critical trade-off between precision and accuracy is addressed, and strategies to maintain high accuracy post-quantization are explored.

As the field of deep learning continues to advance, the insights gleaned from the surveyed literature underscore the importance of quantization as a pivotal tool for adapting NNs to resource-constrained environments. Researchers, practitioners, and enthusiasts in the domain will find this survey a valuable resource, guiding them through the diverse landscape of quantization techniques and providing a foundation for further exploration and innovation in the realm of deep neural network deployment.

References

[1] Al-Qurabat, A. K. M. (2022). "A lightweight Huffman-based differential encoding lossless compression technique in IoT for smart agriculture." Int. J. Com. Dig. Sys 11(1).

[2] Bradley, J. N., et al. (1993). FBI wavelet/scalar quantization standard for gray-scale fingerprint image compression. Visual Information Processing II, SPIE.

[3] Cheng, Y., et al. (2018). "Model compression and acceleration for deep neural networks: The principles, progress, and challenges." IEEE Signal Processing Magazine 35(1): 126-136.

[4] Choudhary, T., et al. (2020). "A comprehensive survey on model compression and acceleration." Artificial Intelligence Review 53: 5113-5155.

[5] De Leon, J. D. and R. Atienza (2022). Depth pruning with auxiliary networks for tinyml. ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE.

[6] Deng, L., et al. (2021). "Comprehensive snn compression using admm optimization and activity regularization." IEEE Transactions on Neural Networks and Learning Systems.
[7] Finkelstein, A., et al. (2022). QFT: Post-training quantization via fast joint finetuning of all degrees of freedom. European Conference on Computer Vision, Springer.
[8] Hu, X. and H. Wen (2021). Research on Model Compression for Embedded Platform through Quantization and Pruning. Journal of Physics: Conference Series, IOP Publishing.
[9] Inés, A., et al. (2023). "Analysing semi-supervised learning for image classification using compact networks in the biomedical context." Soft Computing: 1-13.
[10] Khoram, S. and J. Li (2018). Adaptive quantization of neural networks. International Conference on Learning Representations.
[11] Kirtas, M., et al. (2022). "Quantization-aware training for low precision photonic neural networks." Neural Networks 155: 561-573.
[12] Kugunavar, S. and C. Prabhakar (2021). "Convolutional neural networks for the diagnosis and prognosis of the coronavirus disease pandemic." Visual computing for industry, biomedicine, and art 4(1): 12.
[13] Li, Z., et al. (2023). "Model Compression for Deep Neural Networks: A Survey." Computers 12(3): 60.
[14] Liu, Z., et al. (2022). Nonuniform-to-uniform quantization: Towards accurate quantization via generalized straight-through estimation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.
[15] Mary Shanthi Rani, M., et al. (2022). "DeepCompNet: A novel neural net model compression architecture." Computational Intelligence and Neuroscience 2022.
[16] Mishra, R. and H. Gupta (2023). "Transforming large-size to lightweight deep neural networks for iot applications." ACM Computing Surveys 55(11): 1-35.
[17] Msuya, H. and B. J. Maiseli (2023). "Deep Learning Model Compression Techniques: Advances, Opportunities, and Perspective." Tanzania Journal of Engineering and Technology 42(2): 65-83.
[18] Nori, M. K., et al. (2023). "On the Effectiveness of Activation Noise in Both Training and Inference for Generative Classifiers." IEEE Access.
[19] Oh, S., et al. (2022). Non-Uniform Step Size Quantization for Accurate Post-Training Quantization. European Conference on Computer Vision, Springer.
[20] Pei, Z., et al. (2023). "Quantization via Distillation and Contrastive Learning." IEEE Transactions on Neural Networks and Learning Systems.
[21] Peng, H., et al. (2021). "Deep network quantization via error compensation." IEEE Transactions on Neural Networks and Learning Systems 33(9): 4960-4970.
[22] Polino, A., et al. (2018). "Model compression via distillation and quantization." arXiv preprint arXiv:1802.05668.
[23] Rahebi, J. (2022). "Vector quantization using whale optimization algorithm for digital image compression." Multimedia Tools and Applications 81(14): 20077-20103.
[24] Sui, X., et al. (2023). "A hardware-friendly high-precision CNN pruning method and its FPGA implementation." Sensors 23(2): 824.
[25] Tao, C., et al. (2022). "Compression of generative pre-trained language models via quantization." arXiv preprint arXiv:2203.10705.
[26] Tian, Y., et al. (2023). An Empirical study on model pruning and quantization. International Conference on Broadband Communications, Networks and Systems, Springer.
[27] Xiao, G., et al. (2023). Smoothquant: Accurate and efficient post-training quantization for large language models. International Conference on Machine Learning, PMLR.
[28] Zhai, J., et al. (2023). "FPGA-based vehicle detection and tracking accelerator." Sensors 23(4): 2208.
[29] Zhang, Z. and A. Z. Kouzani (2021). "Resource-constrained FPGA/DNN co-design." Neural Computing and Applications 33(21): 14741-14751.
[30] Zhou, A., et al. (2017). "Incremental network quantization: Towards lossless cnns with low-precision weights." arXiv preprint arXiv:1702.03044.