IRSTI 28.23.37 ` https://doi.org/10.26577/jpcsit2024-02b03

B. Daribayev , N. Azatbekuly* , A. Mukhanbet

LTD DigitAlem, Almaty, Kazakhstan
*e-mail: nurtugang17@gmail.com

OPTIMIZATION OF NEURAL NETWORKS FOR PREDICTING OIL RECOVERY FACTOR USING QUANTIZATION TECHNIQUES

The optimization of computational efficiency in Artificial Neural Network (ANN) models plays a crucial role in enhancing predictions of oil recovery factors in reservoir engineering and Enhanced Oil Recovery (EOR). This study investigates the application of dynamic quantization to improve the efficiency of ANN models deployed in resource-constrained environments. Dynamic quantization, which converts model weights and activations to lower precision formats during inference, aims to reduce memory usage and accelerate computation without significant loss of predictive accuracy.

Using a synthetic dataset generated from the Buckley-Leverett model, encompassing parameters such as porosity, oil viscosity, permeability, classification, and time series data, we evaluated the impact of dynamic quantization on model size, inference time, and predictive performance. Experimental results demonstrate that dynamic quantization effectively reduces model size and speeds up inference, making it suitable for deployment on edge devices with limited computational resources.

This research contributes to advancing the practical implementation of dynamic quantization techniques in optimizing ANN models for complex predictive tasks in reservoir engineering and related fields. The findings underscore the potential of dynamic quantization in improving computational efficiency and facilitating the deployment of ANN models in real-world applications.

Keywords: ANN, Neural Network Quantization, Dynamic Quantization, Enhanced Oil Recovery (EOR), Computational efficiency, Optimization.

1. Introduction

The optimization of computational efficiency in machine learning models is a critical focus in the domain of reservoir engineering, especially for predicting oil recovery factors. As Enhanced Oil Recovery (EOR) techniques become more sophisticated, the demand for accurate, efficient, and scalable predictive models has increased significantly. Artificial Neural Networks (ANNs) have been widely used in this context due to their ability to model complex, non-linear relationships inherent in EOR processes. However, the deployment of these models in resource-constrained environments poses significant challenges due to their computational and memory demands.

Dynamic quantization has emerged as a promising technique to address these challenges. By converting model weights and activations from floating-point precision to lower precision formats, such as 8-bit integers, dynamic quantization reduces the memory footprint and accelerates inference times without significantly compromising predictive accuracy. This technique is particularly relevant for models deployed on edge devices or in real-time applications where computational resources are limited.

Previous studies have explored various aspects of ANN optimization and quantization in the context of EOR. In [1], it was demonstrated that ANNs have the fundamental capabilities to approximate complex functions, providing a theoretical basis for their application in EOR modeling. Research in [2], [3] expanded on this by applying ANNs to specific EOR scenarios, highlighting their effectiveness in predicting oil recovery under various conditions.

More recent advancements in neural network optimization techniques, including quantization, have further enhanced the applicability of ANNs in this field. A method for deep compression, which combines pruning, quantization, and Huffman coding to reduce the size of neural networks, was introduced in [4]. This approach was later refined in [5], focusing on the benefits of quantization for inference speed and memory usage in deep learning models.

In the realm of EOR, quantization techniques have been applied to improve the deployment of predictive models in practical settings. Studies in [6] investigated the use of quantized ANNs to predict oil recovery factors, demonstrating significant improvements in model performance and deployment efficiency. Moreover, quantization techniques have shown promising results in various fields such as image classification [7], [8], object detection [9], [10], and language transformer models [11], [12].

There are two main types of neural network quantization: dynamic and static. Dynamic quantization dynamically computes the clipping range for each activation during inference, often achieving higher model accuracy [13]. However, this approach requires additional computational resources due to the frequent calculation of signal ranges. In contrast, static quantization fixes the clipping range for all input data, simplifying the inference process and reducing computational complexity. While static quantization may lead to a slight decrease in accuracy due to the fixed range, it is often chosen for scenarios with limited computational resources. There are also more advanced quantization methods such as Post-Training Quantization (PTQ) and Quantization-Aware Training (QAT), which are particularly suited for deeper neural networks like convolutional or transformer models. These methods are crucial for reducing the computational costs of neural network inference, making them essential for integrating modern networks into edge devices with strict power and computational resource requirements [14].

In this study, we opted for dynamic quantization to optimize artificial neural networks designed for predicting oil recovery factors, thereby enhancing computational efficiency and prediction accuracy in resource-constrained environments. This approach allows for dynamically adapting quantization parameters based on input data distributions, building on the body of work that introduced dynamic quantization for efficient neural network inference. The model, trained on synthetic data generated through the Buckley-Leverett model, described in [15], captures critical parameters influencing oil recovery, such as porosity, oil viscosity, permeability, and time series data. The impact of quantization on model size, inference time, and predictive accuracy is systematically evaluated and compared to a non-quantized baseline.

The aim of this work is to analyze the impact of the dynamic quantization improving the performance of the model without the loss of the predictive accuracy utilizing the synthetic datasets created based on the Buckley-Leverett model and containing the parameters such as porosity, viscosity, absolute permeability, classification and time series. Therefore, this work aids in extending the understanding and effective implementation of optimized ANN models in the field of reservoir engineering and EOR. It shows the importance of dynamic quantization in furthering the advancement of these methods. Apart from the enhancements it introduces in calculating oil recovery factors, this research also paves the way for the use of dynamic quantization in enhancing neural networks in various engineering and scientific fields.

2. Methodology

In this section, we detail the experimental setup and methodologies employed in applying dynamic quantization to enhance the performance of ANN models in predicting oil recovery factors. The methodology encompasses model architecture, training specifics, and the evaluation metrics used to assess model performance.

2.1. Synthetic EOR Dataset Overview

The dataset utilized in this study is generated synthetically using the Buckley-Leverett model, which simulates oil displacement processes within porous media. This synthetic dataset has been curated to encompass key parameters that influence oil recovery factors, with a specific focus on predicting etta, the oil recovery factor. The parameters included in the dataset are:

- Porosity (poro): Represents the proportion of void space within the rock formation, influencing fluid flow dynamics during oil recovery processes.
- Viscosity of the oil phase (visc_oil): Characterizes the resistance of oil to flow through the porous rock matrix, impacting the efficiency of oil recovery operations.
- Absolute permeability of the rock (kviews): Measures the ability of the reservoir rock to transmit fluids, with higher permeability potentially enhancing oil recovery rates.
- Class categorizing etta results from 1 to 10 (class): Categorizes scenarios affecting oil recovery efficiency, providing insights into predictive variability.
- Time series (time): Captures temporal variations impacting oil recovery processes, including time-dependent factors affecting etta predictions.

This synthetic dataset serves as a controlled environment for evaluating and optimizing predictive models in reservoir engineering and enhanced oil recovery (EOR) scenarios. It provides insights into the relationships between input parameters and etta, facilitating the development of robust predictive models that account for the complexities of oil recovery processes.

The distribution of the oil recovery factor (etta) across the dataset is illustrated in Figure 1, providing insights into its variability and distribution. Additionally, Figure 2 presents a correlation heatmap showcasing the relationships between the input parameters and etta, crucial for understanding the interdependencies that influence oil recovery predictions.

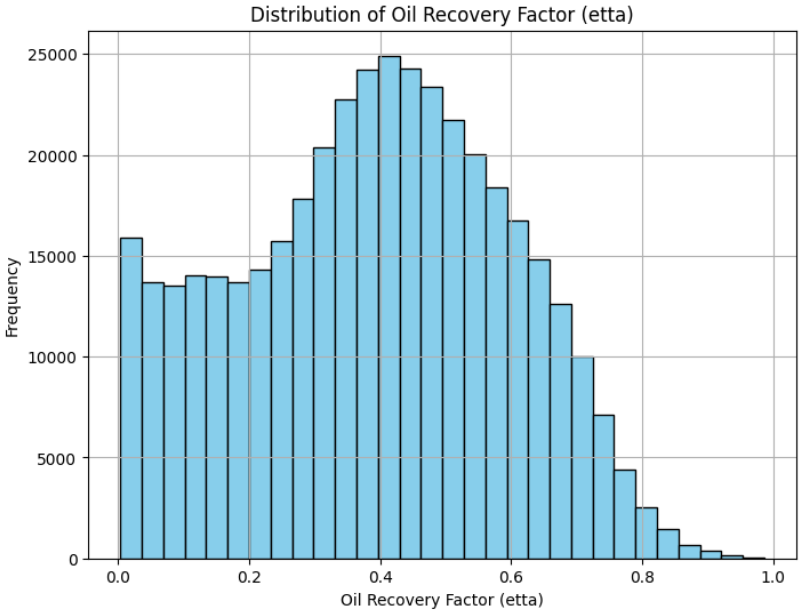

Figure 1 – Distribution of Oil Recovery Factor (etta)

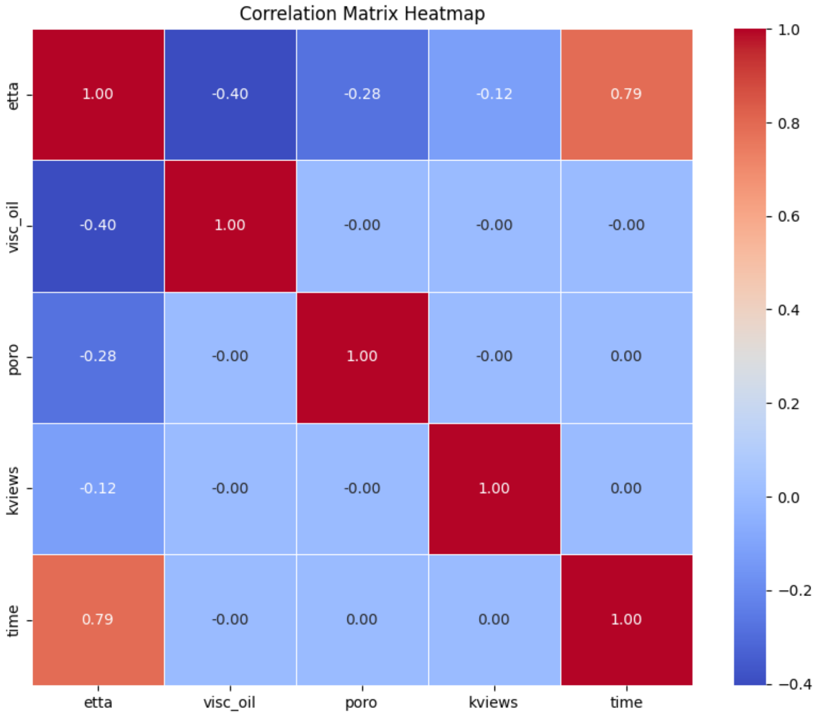

Figure 2 – Correlation Matrix Heatmap

2.2. Experimental Setup

The experimental setup of the dynamic quantization of the ANN models that were developed to predict the oil recovery factors are described by the following elements. The architecture of the employed ANN model consisted of an input layer that corresponds to the feature dimensions of the employed dataset, two hidden layers consisting of 32 neurons, and ReLU activation functions. The output layer included one neuron, which addressed the value of the oil recovery factor, etta.

In the training process, the Adam optimizer was employed with a learning rate of 0. 001 and Mean Squared Error (MSE) as the loss function which minimize the difference between the predicted values and the actual values of the stock price. The model was then trained for three epochs with a batch size of 32. Throughout the training of the model, frequent checks were made on the loss values as illustrated in figure 3. This figure depicts the training loss epochs of the classical ANN model, which depicts the optimization process of the model.

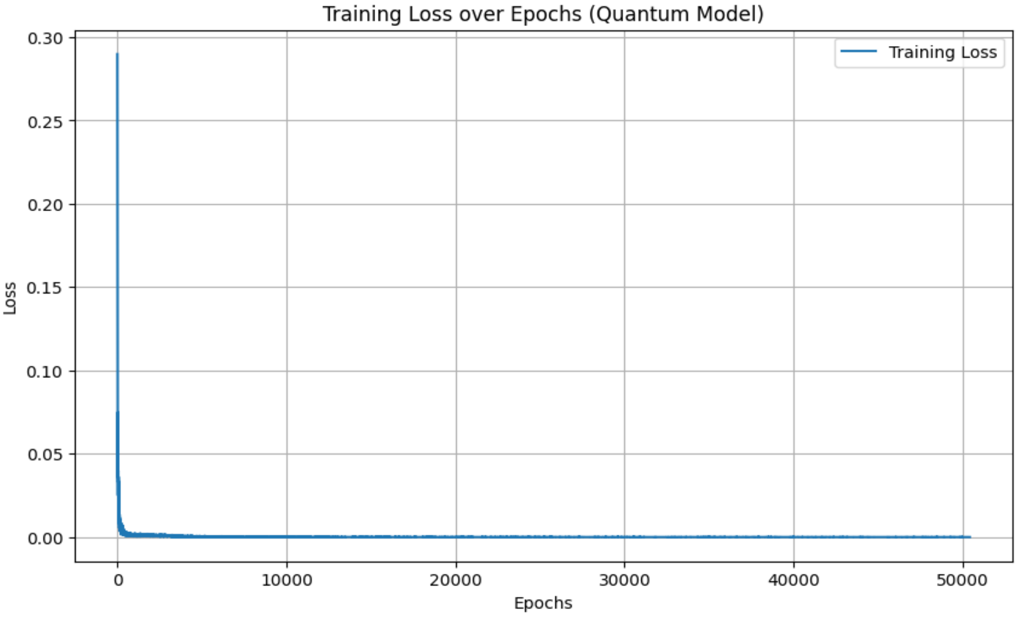

Figure 3 – Training Loss over epochs

Figures 4 shows the ANN architecture used in this research. It describes in detail how the data flows through the model and how each layer – the input layer, the hidden layer, and the output layer – is connected and what specific weight values are assigned to the connections between the layers. This diagram is necessary to help explain the basic architectural organization of the model so that the transformations of data within a neural network during the training and inference phases can be understood.

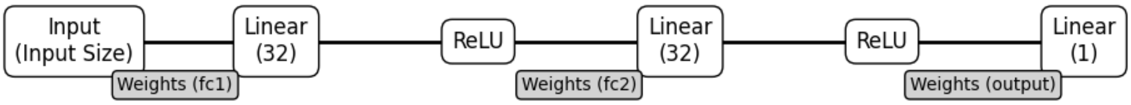

Figure 4 – ANN architecture

These, together, form a platform for experimental purposes which assures clearness and openness during the realization of dynamic quantization strategies whose main goal is to increase the computational efficiency of ANN models designed to forecast oil recovery ratios; the subsequent sections will explicitly discuss how dynamic quantization is applied and its impact on model performance.

2.3. Dynamic Quantization Approach

Dynamic quantization is a way hired on this examine to optimize the computational performance of neural community fashions used for predicting oil healing factors. To illustrate the effect of dynamic quantization on version weights, we applied a easy neural network with one linear layer (10 enter functions and five output functions). Initially, the weights of this version have been in floating-factor format, as proven in Figure five (left). Through dynamic quantization the use of PyTorch`s torch.quantization.quantize_dynamic function, those weights have been transformed to 8-bit integers (torch.qint8), optimizing the version for deployment on resource-restricted devices.

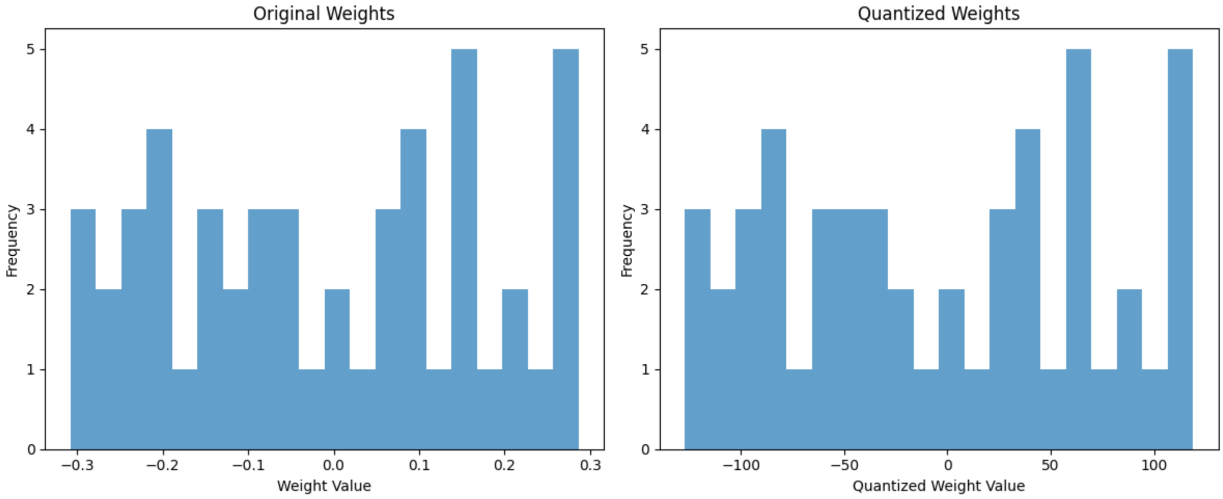

Figure 5 – Distribution of weights before and after dynamic quantization.

The histograms in Figure five reveal the transformation of weight distributions. The left histogram depicts the frequency distribution of authentic floating-factor weights, showcasing a variety of values. After making use of dynamic quantization, proven withinside the proper histogram, those weights are represented as quantized integers. Despite the discount in precision, the distribution stays comparable, making sure minimum effect at the version`s predictive accuracy.

Dynamic quantization has been hired on this observe to beautify the computational performance of Artificial Neural Network (ANN) fashions used for predicting oil restoration factors. This method entails optimizing version overall performance with the aid of using changing floating-factor version weights and activations to decrease precision formats, commonly 8-bit integers. By reducing the accuracy of these parameters, dynamic quantization can effectively reduce the memory footprint and speed up the inference process without significantly sacrificing prediction accuracy.

The process of dynamic quantization starts with loading an ANN model using the PyTorch framework to execute the implementation of dynamic quantization.

In the quantization process, the model weights and activations that are affected during inference are scaled in real-time so as to produce the required performance when operating on different data distribution. This flexibility enables quantized model to provide the same level of accuracy as that of its non-quantized counterpart.

Evaluation parameters deployed in the assessment of the quantized ANN models are R-squared (R^2), Mean Absolute Error (MAE) and Mean Squared Error (MSE). These metrics give a more complete way of measuring the models' performance and efficiency.

Overall, dynamic quantization of ANN models leads to smaller models and faster inference time, which advances the practical deployment of ANN models making the decision-making processes in oil recovery applications faster, hence enhancing their operational efficiency.

Figure 6 presents a quantized architecture of the ANN employed in this study and showcases the effects of dynamic quantization on the efficiency of the model.

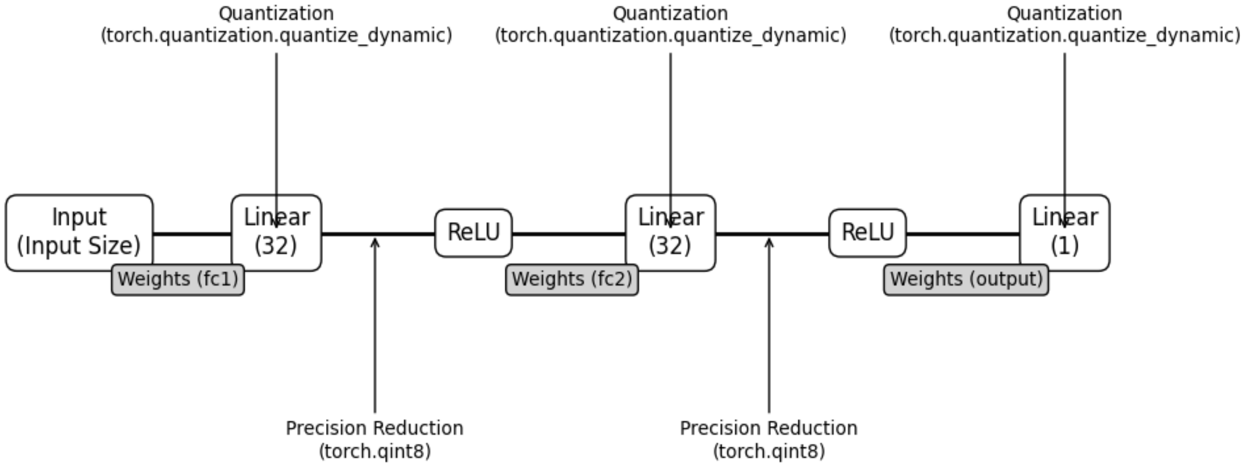

Figure 6 – The quantized architecture of the ANNModel

3. Performance evaluation

This section will describe the results obtained by the models during the evaluation of the effectiveness of dynamic quantization in optimizing Artificial Neural Network (ANN) models for predicting oil recovery factors.

3.1. Performance comparison

The primary metrics evaluated include Test Loss, R² Score, Mean Absolute Error (MAE), and Mean Squared Error (MSE) for both model variants. It is crucial to note that while these metrics traditionally gauge predictive accuracy, our study primarily assesses the impact of quantization on model efficiency rather than improving these accuracy measures.

Table 1 – Performance Comparison of Classic and Quantized Models

Metric	Classic Model	Quantized Model
Test Loss	3.43e-05	6.34e-05
R² Score	0.99917	0.99845
Mean Absolute Error	0.00460	0.00636
Mean Squared Error	0.00003	0.00006

It has been shown that due to quantization, the size of the ANN model in this study is much smaller and it consumes less time than before during inference. The quantized model takes only about 5. 99 KB, while the old model took 7. 50 KB, which clearly indicates less memory usage compared to other programs. This reduction is very important in model deployment, especially on limited devices where every kilobyte of memory matters.

Moreover, the output time for our quantized model is 0.00036 seconds per prediction, while the original model took 0.00062 seconds. This improvement shows the fact of increased computational speed by applying quantization, which allows the model to provide faster answers for immediate use or decision making.

These efficiency gains underscore the practical benefits of dynamic quantization in optimizing ANN models for deployment in diverse operational environments. Figure 9 provides a visual representation of the comparative model sizes and inference times, reaffirming the quantitative advantages observed in this study.

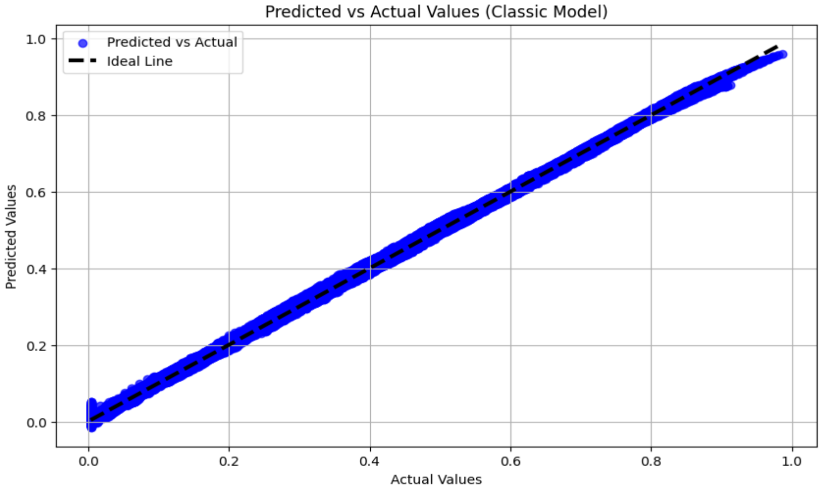

Figure 7 – Predicted vs Actual Values (Classical model)

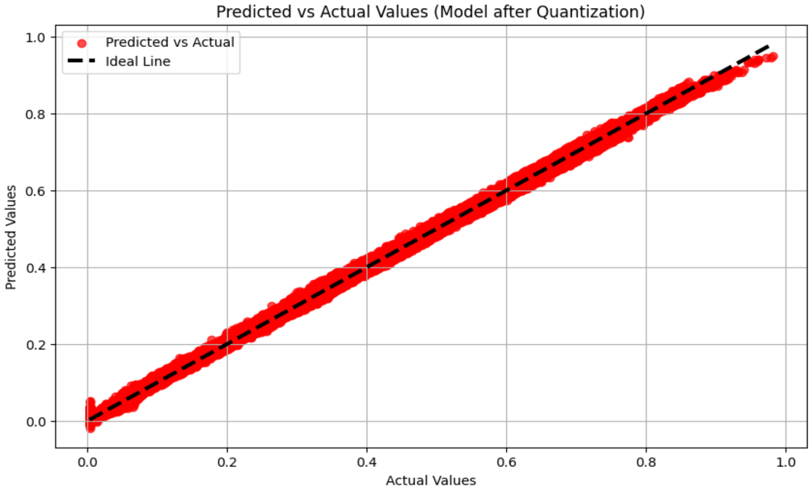

Figure 8 – Predicted vs Actual Values (Model after Quantization)

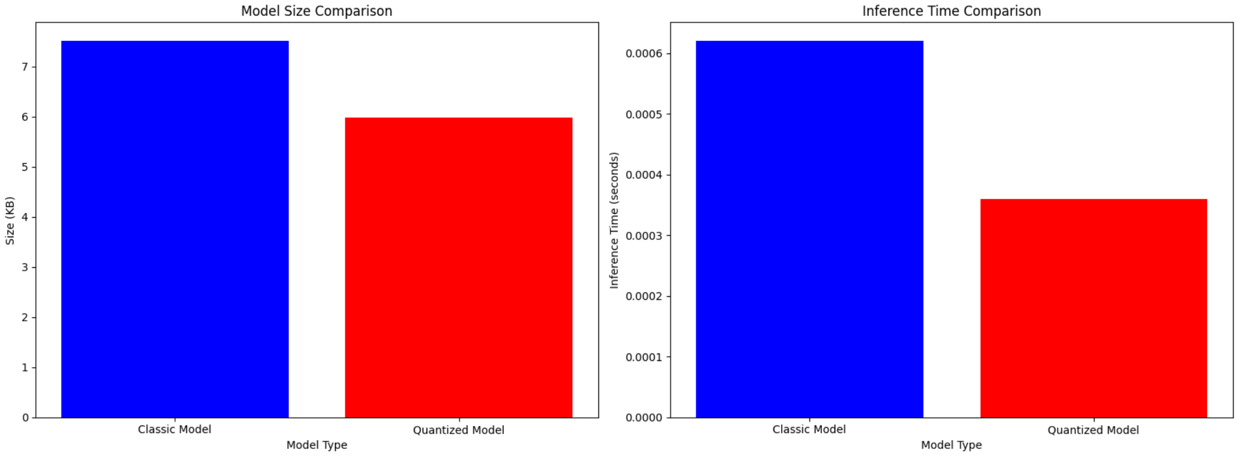

Figure 9 – Comparison of model sizes and inference time

The results not only validate the effectiveness of quantization in enhancing computational performance but also emphasize its role in facilitating streamlined and efficient deployment of machine learning models in practical applications.

4. Conclusion

In conclusion, this study underscores the efficacy of dynamic quantization in optimizing Artificial Neural Network (ANN) fashions for predicting oil restoration elements in reservoir engineering and superior oil restoration (EOR). By making use of dynamic quantization, we efficiently decreased the model size and extended inference times, improving computational performance without compromising predictive accuracy.

The experimental results illustrated significant benefits of dynamic quantization. The quantized ANN models exhibited a noticeable decrease in model size, consuming only 5.99 KB compared to the original model's 7.50 KB, which is crucial for deployment on resource-constrained edge devices. Moreover, the quantized models showed faster inference times, with computations completing 0.00036 seconds per prediction as opposed to 0.00062 seconds for the non-quantized models. These efficiency gains highlight the practical advantages of dynamic quantization in real-time decision-making scenarios and operational environments.

The reduction in model size and advanced inference pace located in quantized ANN fashions demonstrates their suitability for deployment on resource-limited gadgets and in real-time packages. This study contributes to advancing the realistic packages of dynamic quantization in optimizing neural community fashions throughout various engineering and clinical domains.

Future studies can in addition discover and refine more advanced quantization strategies to beautify computational performance in greater complicated predictive modeling obligations past oil restoration, paving the manner for broader packages in enterprise and academia alike.

Acknowledgments

This research was funded by the Grant No. AP19576314 "Development of high-performance hybrid algorithms for solving oil displacement problems based on classical (GPU) and quantum computing" of Ministry of Science and Higher Education of the Republic of Kazakhstan.

References

1. Baldi, Pierre, and Kurt Hornik. 1989. "Neural Networks and Principal Component Analysis: Learning from Examples without Local Minima." Neural Networks. Elsevier BV. https://doi.org/10.1016/0893-6080(89)90014-2
2. Vo Thanh, Hung, Yuichi Sugai, and Kyuro Sasaki. 2020. "Application of Artificial Neural Network for Predicting the Performance of CO2 Enhanced Oil Recovery and Storage in Residual Oil Zones." Scientific Reports. Springer Science and Business Media LLC. https://doi.org/10.1038/s41598-020-73931-2
3. Muradkhanli, Leyla. 2018. "Neural Networks for Prediction of Oil Production." IFAC-PapersOnLine. Elsevier BV. https://doi.org/10.1016/j.ifacol.2018.11.339
4. Han, Song, Huizi Mao, and William J. Dally. 2015. "Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding." arXiv. https://doi.org/10.48550/ARXIV.1510.00149
5. Jacob, Benoit, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. 2018. "Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference." 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE. https://doi.org/10.1109/cvpr.2018.00286
6. Liu, Zhiwei, Shaoqi Yan, Hangyu Zang, Peixuan Cui, Xincheng Cui, Yingge Li, and Dongxing Du. 2023. "Quantization of the Water Presence Effect on the Diffusion Coefficients of the CO2/Oil System with the Dynamic Pendant Drop Volume Analysis Technique." Chemical Engineering Science. Elsevier BV. https://doi.org/10.1016/j.ces.2023.119142
7. Choi, Jungwook, Zhuo Wang, Swagath Venkataramani, Pierce I-Jen Chuang, Vijayalakshmi Srinivasan, and Kailash Gopalakrishnan. 2018. "PACT: Parameterized Clipping Activation for Quantized Neural Networks." arXiv. https://doi.org/10.48550/ARXIV.1805.06085
8. Chen, Peng, Jing Liu, Bohan Zhuang, Mingkui Tan, and Chunhua Shen. 2020. "AQD: Towards Accurate Fully-Quantized Object Detection." arXiv. https://doi.org/10.48550/ARXIV.2007.06919
9. Li, Rundong, Yan Wang, Feng Liang, Hongwei Qin, Junjie Yan, and Rui Fan. 2019. "Fully Quantized Network for Object Detection." 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE. https://doi.org/10.1109/cvpr.2019.00292
10. Chen, Peng, Jing Liu, Bohan Zhuang, Mingkui Tan, and Chunhua Shen. 2020. "AQD: Towards Accurate Fully-Quantized Object Detection." arXiv. https://doi.org/10.48550/ARXIV.2007.06919

11. Zafrir, Ofir, Guy Boudoukh, Peter Izsak, and Moshe Wasserblat. 2019. "Q8BERT: Quantized 8Bit BERT." arXiv. https://doi.org/10.48550/ARXIV.1910.06188

12. Bhandare, Aishwarya, Vamsi Sripathi, Deepthi Karkada, Vivek Menon, Sun Choi, Kushal Datta, and Vikram Saletore. 2019. "Efficient 8-Bit Quantization of Transformer Neural Machine Language Translation Model." arXiv. https://doi.org/10.48550/ARXIV.1906.00532

13. Gholami, Amir, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W. Mahoney, and Kurt Keutzer. 2021. "A Survey of Quantization Methods for Efficient Neural Network Inference." arXiv. https://doi.org/10.48550/ARXIV.2103.13630

14. Nagel, Markus, Marios Fournarakis, Rana Ali Amjad, Yelysei Bondarenko, Mart van Baalen, and Tijmen Blankevoort. 2021. "A White Paper on Neural Network Quantization." arXiv. https://doi.org/10.48550/ARXIV.2106.08295

15. Daribayev, B., D. Akhmed-Zaki, T. Imankulov, Y. Nurakhov, and Y. Kenzhebek. 2020. "Using Machine Learning Methods for Oil Recovery Prediction." ECMOR XVII. European Association of Geoscientists & Engineers. https://doi.org/10.3997/2214-4609.202035233

Information about authors:

Beimbet S. Daribayev – PhD, researcher at DigitAlem LTD (Almaty, Kazakhstan, e-mail: beimbet.daribayev@gmail.com).

Nurtugan Azatbekuly (corresponding author) – Junior researcher at DigitAlem LTD (Almaty, Kazakhstan, e-mail: nurtugang17@gmail.com)

Aksultan A. Mukhanbet – Researcher at DigitAlem LTD (Almaty, Kazakhstan, e-mail: mukhanbetaksultan0414@gmail.com).

Submission received: 15 June, 2024.
Revised: 22 June, 2024.
Accepted: 22 June, 2024.