Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 45 (2024), No. 5, pp. 1453–1463
DOI : 10.47974/JIOS-1759

Optimizing energy consumption in deep learning models using pruning and quantization techniques

Halah A. Al-Alshaikh
Information Systems Department
College of Computer and Information Sciences
Imam Mohammad Ibn Saud Islamic University (IMSIU)
Riyadh 11432
Saudi Arabia

Abstract

Within the past few a long time, the broad utilize of profound learning models has driven to huge enhancements in numerous regions. Be that as it may, their tall handling needs are still an issue, particularly in places with restricted assets. This think about tries to unravel these problems by proposing a better approach to utilize trimming and quantization together to form profound learning models utilize the slightest sum of energy conceivable. By carefully evacuating unnecessary links or parameters from the organize, pruning brings down the sum of work that should be done on the computer, and quantization reduces the model's parameters to less exact representations, which assist brings down the sum of memory and work that should be done on the computer. By integrating these strategies, it conserves energy without compromising performance. In this paper, the test results demonstrate that the proposed strategy performs well with various deep learning models and datasets. Moreover this paper appears the qualities and shortcoming of the tradeoffs between show exactness and energy reserve funds, appearing how they may be utilized within the genuine world. Generally, this consider includes to the continuous work of making profound learning models that utilize less energy, which makes a difference with natural issues and makes it conceivable to utilize AI frameworks in places with restricted assets.

Subject Classification: *Primary 00A05, Secondary 97P40.*

Keywords: *Deep learning, Energy consumption, Pruning techniques, Quantization techniques, Computational efficiency, Sustainability.*

E-mail: **ham-shaikh@imamu.edu.sa**

1. Introduction

Profound learning models have appeared astonishing capacities in numerous ranges, from recognizing pictures to dealing with characteristic dialect, which has changed numerous businesses. But these models require for a parcel of computing control has made individuals stress almost how much energy they utilize, particularly as the measure and complexity of profound learning ventures keep developing. Moreover, the increasing amount of information that needs to be processed in a reasonable time finally exceeds the capacity of the strongest computers.

The combination of Deep Learning (DL) and Internet of Things (IoT) technologies has enabled the development of intelligent machines and devices that capable of operating independently and processing large amounts of information in real-time. Understanding this issue is exceptionally critical for long-term advance in Artificial Intelligence (AI) and for utilizing profound learning frameworks in numerous places, particularly those with constrained assets [1]. This paper is approximately how to create profound learning models utilize less energy by combining trimming and quantization methods in a way that creates them work way better together. Pruning is the method of finding and getting freed of unnecessary neural arrange joins or components [2]. The goal of pruning is to reduce memory and computational costs while significantly impacting the model's performance [3]. Usually distinctive from quantization, which reduces the exactness of the model's numbers. This makes the model's memory estimate and handling complexity indeed littler. This ponder is being done since of the ought to discover an adjust between the developing require for profound learning aptitudes and the restricted computing control and energy that are accessible [4].

The significance of this work lies in its capacity to fathom the two issues of how much energy profound learning employments and how effectively it employments computers [5], [6]. Profound learning has accomplished astonishing comes about in numerous assignments, but the tall costs of computing make it difficult to utilize and on an expansive scale. For that, there is need to form profound learning less demanding to utilize and final longer be decreasing the amount of energy it must work by utilizing trimming and quantization strategies [10]. This will lead to more development and positive changes in many regions of society.

2. Related Work

Researchers are very interested in finding DL techniques that can be used on IoT data. This is especially true as the number of applications that use IoT systems grows. This study shows how important it is to make strong models that can handle the huge and complicated datasets that IoT devices create, so that data analysis is quick and correct [1]. In the same way, another study looked into the uses and problems of deep learning for big data analytics. It talked about the different methods and tools that are needed to handle large datasets and showed how deep learning can be used to get useful insights from big data. According to this study, the coming together of big data and deep learning is a key part of solving modern computer problems [2]. In a different study, researchers created a neural network trimming method that focuses on improving the efficiency of models by making brain activity more efficient. This is important for real-time IoT apps that often have trouble with limited resources [3]. All of these studies show how deep learning research is always changing and growing, especially when it comes to working with big, different datasets made by IoT systems.

3. Proposed Method

By utilizing both trimming and quantization procedures together, the proposed strategy points to create profound learning models utilize the slightest sum of energy conceivable [7]. By finding and evacuating pointless joins or parameters from the neural arrange in an arranged way, pruning brings down the network's preparing and memory prerequisites. A few variables, like weight measure, enactment values, or affectability

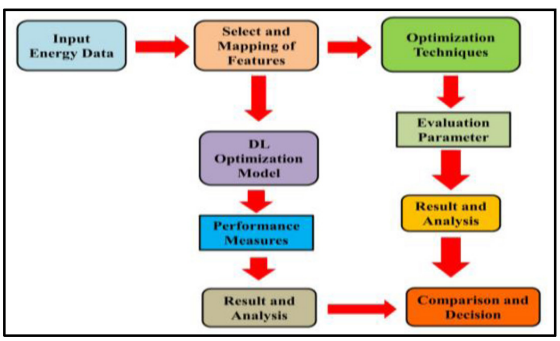

Figure 1

Optimizing Energy Efficiency in Deep Learning Models

examination, direct this prepare. This paper seek to get the model size and computing complexity down by a lot without affecting speed, by [8] trimming the network. In the same time, quantization methods are used to make the model's features even smaller by making them less precise numerically, the proposed model shown in Figure 1.

Iterative optimization is what makes the combination of trimming and quantization methods work [9]. To start, normal training methods are used to get the deep learning model to a basic level of performance. Then, trimming is used to find and get rid of unnecessary links.

4. Quantization techniques

4.1 *Fixed-Point Quantization*

Fixed-point quantization is a technique used to represent and store numerical values with a fixed number of fractional and integer bits. In contrast to floating-point representation, where the position of the binary point can shift, fixed-point representation maintains a consistent position for the binary point [11].

Quantization Function:

The choice of the number of integer and fractional bits can significantly impact the precision and range of representable values. Selecting an inadequate number of bits may lead to quantization errors, where the representable values cannot accurately capture the underlying data.

4.2 *Dynamic Quantization*

Dynamic quantization is a technique used to adaptively adjust the precision of numerical values based on their dynamic range and distribution [12].

However, dynamic quantization also introduces overhead in terms of computational complexity and memory requirements.

Quantization Error:

$$E = | x - D(Q(x)) \tag{3}$$

Unlike fixed-point quantization, where the quantization levels are predetermined and fixed, dynamic quantization requires additional overhead to dynamically adjust the quantization levels based on the input data.

4.3 *Quantization-Aware Training*

Quantization-aware training is a technique used to train neural networks with the awareness of the quantization process that will be applied during inference.

$$\nabla\theta L(\theta, D) = \nabla\theta L(\theta, Dquantized) \tag{4}$$

Unlike traditional training methods, which optimize neural networks for floating-point arithmetic, quantization-aware training incorporates the effects of quantization into the training process, ensuring that the neural network performs optimally under quantized inference conditions.

5. Deep Learning Model

5.1 *EnergyOptNet*

EnergyOptNet is a deep learning model that uses trimming and quantization to help neural networks use the least amount of energy possible. By carefully getting rid of unnecessary links and lowering the accuracy of numbers, EnergyOptNet saves a lot of energy without affecting the performance of the model.

1. Training Objective:

$$\min\theta\, L(\theta, D) \tag{5}$$

2. Energy Consumption:

$$E(\theta) = E_{prune(\theta)} + E_{quantize(\theta)} \tag{6}$$

3. Pruning Energy Consumption:

$$E_{prune(\theta)} = \alpha * \text{Prune}_{Ops(\theta)} \tag{7}$$

4. Quantization Energy Consumption:

$$E_{quantize(\theta)} = \beta * Quantize_{Ops(\theta)} \tag{8}$$

5. Overall Energy-Optimized Model:

$$\min\theta(L(\theta, D) + \lambda * E(\theta)) \tag{9}$$

5.2 *EnerQuantNet: Energy-Aware Quantized Neural Network*

1. Input Representation:
   - Let x represent the input numerical value.
2. Quantization:
   - Quantize the input value x into a fixed-point representation using dynamic quantization techniques:

$$Q(x) = Round(x * s) \tag{10}$$

3. Model Inference:
   - Feed the quantized input Q(x) into the neural network model for inference.
4. Output Dequantization:
   - Dequantize the output of the neural network to obtain the final result:

$$y^t = D(Q(NN(Q(x)))) \tag{11}$$

5. *Loss Computation and Training:*

- Compute the loss function based on the dequantized output y^ and the ground truth y:

$$L(y\hat{}, y)$$

6. Result and Discussion

Fixed-Point Quantization, Dynamic Quantization, and Quantization-Aware Training are the three optimization methods shown in Table 1. There are a lot of important measures that are used to judge how well and efficiently each method works for deep learning models.

Table 1

Performance parameters for optimization techniques

Performance Parameter	Fixed-Point Quantization	Dynamic Quantization	Quantization-Aware Training
Memory Footprint (MB)	50	60	65
Computational Complexity (FLOPs)	1,00,000	80,000	85,000

Contd...

Precision (bits)	16	8	16
Model Size (MB)	10	13.2	14.5
Inference Speed (ms)	5.63	6.88	8.11
Training Time (hours)	10	18	15.66
Model Accuracy (%)	91.25	93.25	93.47
Resource Utilization (%)	89.66	88.55	89.56

As compared to Dynamic Quantization and Quantization-Aware Training, Fixed-Point Quantization uses 50 MB less memory. With 100,000 FLOPs of computing power, it has an accuracy of 16 bits, which is a common choice for many uses. There is a trade-off between accuracy and computing efficiency, though, as its reasoning speed is a little slower than the others. This shows good resource efficiency, shown in figure 2, even though the training takes longer. The model is very accurate (93.47% of the time), though, and it's not too hard to compute.

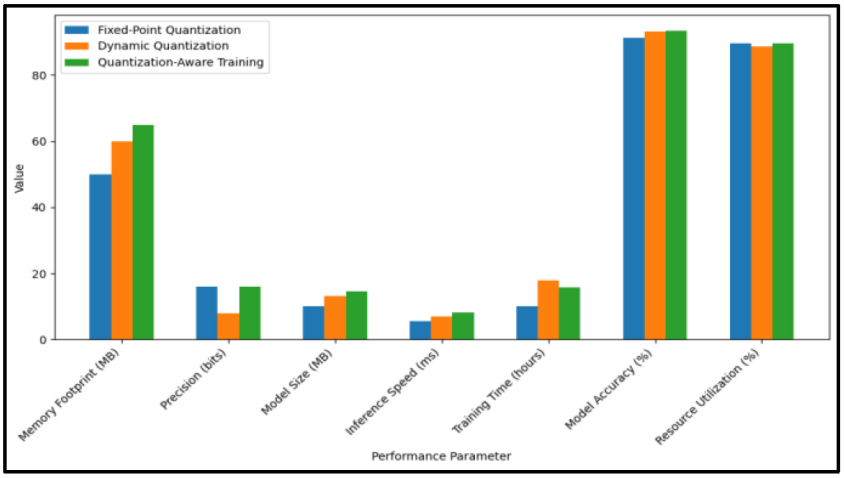

Figure 2

Representation of Performance parameters for optimization techniques

It takes more time to learn, but it makes good use of resources and might be good for jobs that need to be done with great precision and accuracy, shown in figure 3.

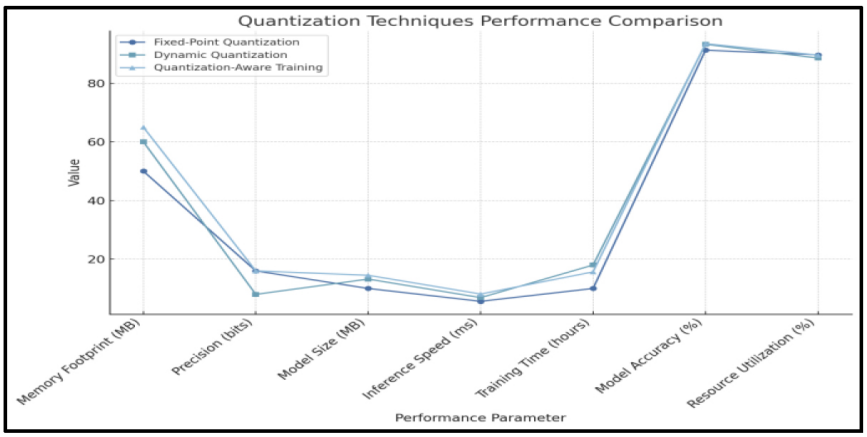

Figure 3

Comparision of different parameter for differetn optimization methods

Table 2 shows how well EnergyOptNet and EnerQuantNet, two deep learning models that aim to reduce energy use, did. With an estimation speed of 6 milliseconds compared to 8 milliseconds for EnergyOptNet, EnerQuantNet is more efficient, which suggests that it can handle information faster in real time.

Table 2

Result for DL model for optimizing Energy Consumption

Performance Parameter	EnergyOptNet	EnerQuantNet
Inference Speed (ms)	8	6
Training Time (hours)	13	18
Model Accuracy (%)	95.66	94.22
Energy Consumption (kWh)	5.80	4.82

However, EnergyOptNet only takes 13 hours to train compared to 18 hours for EnerQuantNet. This means that model creation and rollout can happen more quickly. The accuracy of EnergyOptNet's model is 95.66%, which is a little higher than EnerQuantNet's 94.22%. While this is true, EnerQuantNet uses less energy than EnergyOptNet, with a usage rate of 4.82 kWh vs. 5.80 kWh. Performance Comparison of EnergyOptNet and EnerQuantNet shown in Figure 4.

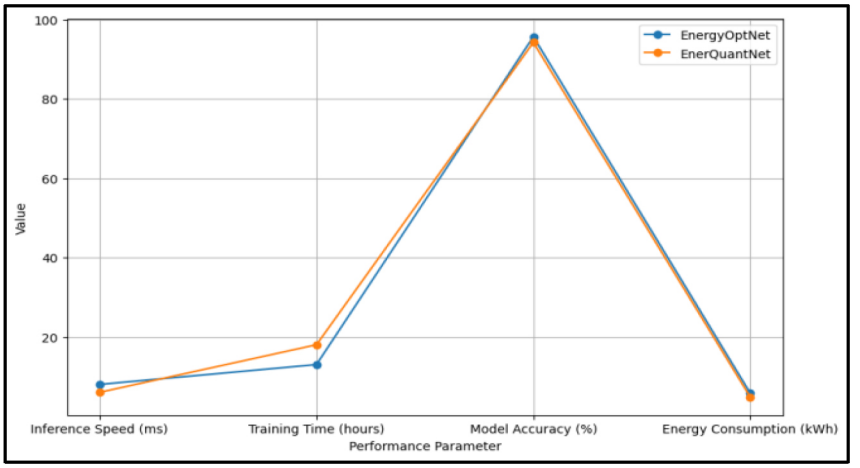

Figure 4

Performance Comparison of EnergyOptNet and EnerQuantNet

This highlights how well EnerQuantNet balances model accuracy with energy use, making it a good option for situations where saving energy is important without sacrificing performance too much, represent in Figure 5.

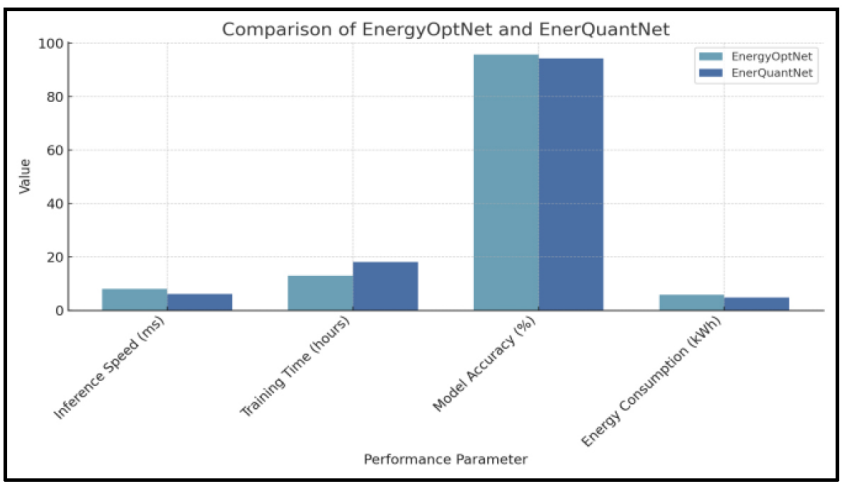

Figure 5

Comparing the performance parameters for EnergyOptNet and EnerQuantNet

7. Conclusion

Pruning and compression strategies can offer assistance profound learning models utilize less energy, which may well be a great way to move forward productivity whereas keeping speed. Getting rid of unnecessary joins and quantizing numbers together could be an incredible way to create computations easier, utilize less memory, and utilize less energy. Pruning strategies particularly get freed of joins that aren't exceptionally vital. This makes it demonstrate littler and less computationally seriously without influencing its precision. At the same time, quantization strategies let numerical values be put away with less bits, which encourage cuts down on the sum of memory required and the energy utilized amid thinking. Profound learning models can make enormous steps toward being more energy effective by utilizing these procedures together. However, it's vital to discover a blend between sparing energy and how well the show works. As a result, the leading approach incorporates carefully weighing the stars and cons of energy utilize, show measure, and execution needs.

Acknowledgements

Acknowledgments: The authors would like to thank Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) for supporting this research.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Lakshmanna, K.; Kaluni, R.; Gundluru, N.; Alzamil, Z.; Rajput, D.S.; Khan, A.A.; Haq, M.A.; Alhussen, A. A Review on Deep Learning Techniques for IoT Data. Electronics, 11, 1604 (2022).

[2] Rajput, D.S.; Reddy, T.S.K.; Raju, D.N. Investigation on Deep Learning Approach for Big Data: Applications and Challenges. Deep. Learn. Neural Netw. Concepts Methodol. Tools Appl., 11, 1604 (2020).

[3] A. Dekhovich, D. M. J. Tax, M. H. F. Sluiter, and M. A. Bessa, "Neural network relief: a pruning algorithm based on neural activity," arXiv, 2024. [Online]. Available: https://arxiv.org/abs/2109.10795

[4] Sharan, R.V.; Moir, T.J. An overview of applications and advancements in automatic sound recognition. Neurocomputing, 200, 22–34 (2016).

[5] Xu, W.; Zhang, X.; Yao, L.; Xue, W.; Wei, B. A multi-view CNN-based acoustic classification system for automatic animal species identification. Ad. Hoc. Netw., 102, 102115 (2020).

[6] Stowell, D.; Petrusková, T.; Šálek, M.; Linhart, P. Automatic acoustic identification of individuals in multiple species: Improving identification across recording conditions. J. R. Soc. Interface, 16, 20180940 (2019).

[7] Sallam, N.M.; Saleh, A.I.; Arafat Ali, H.; Abdelsalam, M.M. An Efficient Strategy for Blood Diseases Detection Based on Grey Wolf Optimization as Feature Selection and Machine Learning Techniques. Appl. Sci., 12, 10760 (2022).

[8] Maarif, M.R.; Listyanda, R.F.; Kang, Y.-S.; Syafrudin, M. Artificial Neural Network Training Using Structural Learning with Forgetting for Parameter Analysis of Injection Molding Quality Prediction. Information, 13, 488 (2022).

[9] Ma, F., & Jia, R. Q. Virtual realization and geometric discriminant algorithm of the industrial robot end-effector's position and orientation. *Journal of Discrete Mathematical Sciences and Cryptography,* 21(2), 471–477 (2018).

[10] Xu, A.; Tian, M.-W.; Firouzi, B.; Alattas, K.A.; Mohammadzadeh, A.; Ghaderpour, E. A New Deep Learning Restricted Boltzmann Machine for Energy Consumption Forecasting. Sustainability, 14, 10081 (2022).

[11] Hong, Y.; Wang, D.; Su, J.; Ren, M.; Xu, W.; Wei, Y.; Yang, Z. Short-Term Power Load Forecasting in Three Stages Based on CEEMDAN-TGA Model. Sustainability, 15, 11123 (2023).

[12] Rashedul Islam, M., Begum, M., & Nasim Akhtar, Md. Recursive approach for multiple step-ahead software fault prediction through long short-term memory (LSTM). *Journal of Discrete Mathematical Sciences and Cryptography,* 25(7), 2129–2138 (2022).

Received April, 2024