ISSN 0002-306X. Proc. of the RA NAS and NPUA Ser. of tech. sc. 2023. V. LXXVI, N4

**UDC 621.382** **MICROELECTRONICS**

**DOI: 10.53297/0002306X-2023.v76.4-506**

**A.A. AVETISYAN**

# INVESTGATING POWER METRICS OF NEURAL NETWORKS AFTER PRUNING, QUANTIZATION, DPU IMPLEMENTATION: A COMPARATIVE ANALYSIS

The rapid proliferation of deep learning applications in various fields has highlighted the need for efficient neural network implementations, especially on resource-constrained edge devices. In response to this demand, pruning and quantization have emerged as essential techniques to reduce the computational and memory requirements of neural networks. Additionally, the deployment of dedicated hardware, such as Digital Processing Units (DPUs), has gained momentum for accelerating neural network inference.

This paper presents a comprehensive comparative analysis of the power metrics of neural networks after pruning and quantization, with a particular focus on their implementation on DPUs. The objective of this research is to investigate the energy efficiency and power consumption of pruned and quantized neural networks when executed on DPU platforms. The trade-offs between model size reduction and inference accuracy, as well as the power efficiency of different DPU architectures are researched.

The results reveal insights into the power efficiency of pruned and quantized neural networks on DPU platforms, offering a clear understanding of the benefits and trade-offs associated with these optimization techniques.

This research provides a valuable resource for researchers, developers, and practitioners interested in optimizing neural network implementations for power efficiency. The findings contribute to the ongoing effort to make deep learning more accessible and sustainable on edge devices and other power-constrained environments, ultimately enabling a wider range of applications with reduced energy consumption.

***Keywords:*** FPGA, DPU, pruning, quantization.

**Introduction**. The widespread adoption of deep learning has revolutionized the fields of artificial intelligence and machine learning, propelling remarkable advancements in various applications such as image recognition, natural language processing, autonomous systems, etc. However, this transformation has not come without its challenges, particularly when deploying deep neural networks on resource-constrained edge devices. As the demand for efficient, low-power artificial intelligence (AI) solutions continues to grow, researchers and engineers have turned their attention to techniques that reduce the computational and memory footprint of neural networks. Among these techniques, pruning [1] and quantization

[2] have emerged as pivotal strategies, allowing neural networks to maintain high performance while consuming fewer resources.

In parallel with the quest for resource-efficient neural networks, the development and utilization of dedicated hardware accelerators have gained momentum. Digital Processing Units (DPUs) [3] represent a class of specialized hardware designed to optimize neural network inference, offering both computational power and energy efficiency. This paper delves into the intersection of these two key domains: neural network optimization through pruning and quantization, and the deployment of these networks on DPU platforms.

The primary objective of this paper is to conduct a thorough and comparative analysis of the power metrics associated with neural networks after pruning and quantization, with a specific focus on their execution on DPUs. Pruning involves removing unimportant network connections, effectively reducing model size and, in some cases, enhancing inference speed. Quantization, on the other hand, reduces the precision of network weights and activations, thus leading to further compression of the model and resource-efficient execution. These techniques have the potential to unlock AI capabilities on edge devices, where computational and power constraints are paramount.

The findings presented in this paper hold relevance for researchers, developers, and practitioners engaged in the pursuit of efficient neural network implementations. This paper contributes to the ongoing effort to make deep learning more accessible and sustainable in power-constrained environments, fostering the proliferation of AI at the edge.

From autonomous vehicles and medical diagnostics to virtual assistants and industrial automation, DNNs have demonstrated their prowess in solving complex problems and enabling intelligent decision-making. Yet, as these networks have grown in size and complexity, so too have the challenges associated with deploying them efficiently on resource-constrained edge devices.

To address these challenges, researchers and engineers are trying to optimize the execution of DNNs. As a result, two main techniques were developed: pruning and quantization. Pruning focuses on reducing model size and computational demands by identifying and eliminating redundant network connections. Quantization, on the other hand, aims to reduce memory requirements and accelerate computation by decreasing the precision of network parameters. Together, these techniques have the potential to make deep learning feasible on devices with limited computational resources.

While neural network optimization, power efficiency, and hardware acceleration are widely studied topics, there are several key papers worth mentioning. Each of them gives a thorough description and together they make it possible to perform a comparative analysis.

In [4] the concept of efficient model scaling is introduced, demonstrating that model size and computational requirements can be balanced to achieve resource-efficient DNNs. EfficientNet serves as a foundational reference in the pursuit of optimizing DNNs for edge deployments. [5] presents a comprehensive study of quantization techniques, highlighting the advantages of reducing bit precision in network parameters. This work has inspired numerous studies on efficient DNN execution.

Pruning, with its potential to create compact neural networks, has attracted significant attention, as evidenced by paper [6]. This pioneering work laid the groundwork for understanding how network pruning can lead to a reduction in model size without sacrificing performance. Furthermore, the development of hardware accelerators tailored for neural network inference is described in [7]. It contains insights into specialized hardware designed to enhance the DNN execution efficiency.

In the following sections quantization, pruning and DPU implementation are researched, to understand the overall tradeoffs during neural network optimizations.

**Method.** The model is trained on the "Kaggle Dogs vs. Cats" dataset. After the training is finished, the model accuracy reaches 97.7%. The evaluation is performed on 1000 image dataset, and the power consumption is monitored via sensors on the GPU.

Then the trained model is pruned. Pruning is an iterative process in which the network is reduced by a certain amount (10% per iteration is used in this work) and then fine-tuned (retrained) to bring its performance back to the original. In this work, we are repeating the pruning-retraining sequence six times. The final model's accuracy reaches 92.94%. After the pruning is done, the model is evaluated similarly as it was done for the original float model and the power is measured again.

To further reduce the model's power consumption, quantization is applied to the pruned model. Quantization is a technique used to reduce the memory footprint and computational requirements of CNNs by representing and performing computations with lower precision data types. In a standard CNN, weights and activations are typically represented as 32-bit floating-point numbers (FP32), which consume significant memory and require higher computational resources. Quantization aims to replace these higher precision representations with lower precision data types, such as fixed-point or integer values, which have fewer bits.

Finally, the quantized CNN is taken and compiled for the proper DPU architecture.

The DPU is a soft-core IP whose only function is to accelerate the execution of CNNs. It acts as a co-processor to the host processor and has its own instruction

set: the Vitis AI compiler will convert and optimize, where possible, the quantized model to a set of micro-instructions and then output them to an xmodel file for the DPU.

The action sequence described in the previous paragraphs forms a standard flow which is shown in Fig. 1.

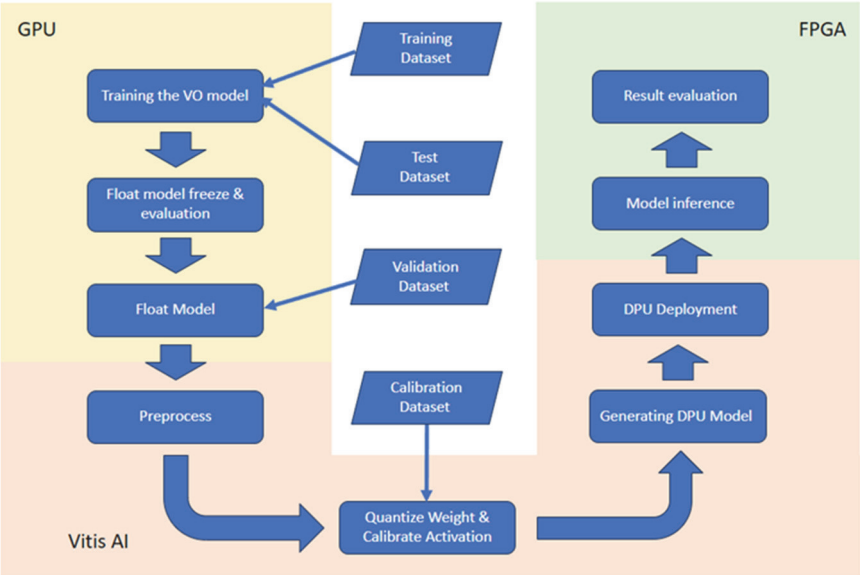


*Fig. 1. DPU deployment flow*

**Experimental results.** The experiments show that pruning and quantization are reducing overall power consumption. At the same time, network performance and accuracy do not suffer much. The simulation accuracy results and power measurements are presented in Table.

*Table*

*Object recognition neural network performance after pruning and quantization*

| | Power (Watts) | Accuracy (%) |
|---|---|---|
| Float | 26.810 | 97.7 |
| Prune1 | 23.403 | 96.4 |
| Prune2 | 23.648 | 96.3 |
| Prune3 | 20.446 | 94.1 |
| Prune4 | 19.101 | 93.8 |
| Prune5 | 17.252 | 93.4 |
| Prune6 | 15.383 | 92.9 |
| Quantize | 11.165 | 92.8 |

The DPU implementation is further reducing power consumption to 9.864 Watts. So, the tradeoff is 4.9% in accuracy vs 63.21% power reduction.

Although the implemented DPU (Fig. 2) has high utilization percent, the tested neural network is relatively small, meaning that the power efficiency can increase further for bigger neural networks.

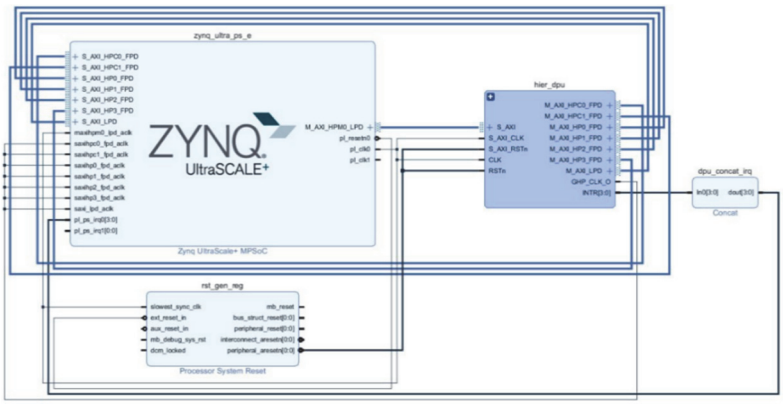


*Fig. 2. Synthesized DPU design*

After the pruning process is finished the network's size is reduced by 57.9%. Fig. 3 shows the visualization of the float and the pruned model. Some of the convolutions and SoftMax activation functions are pruned.

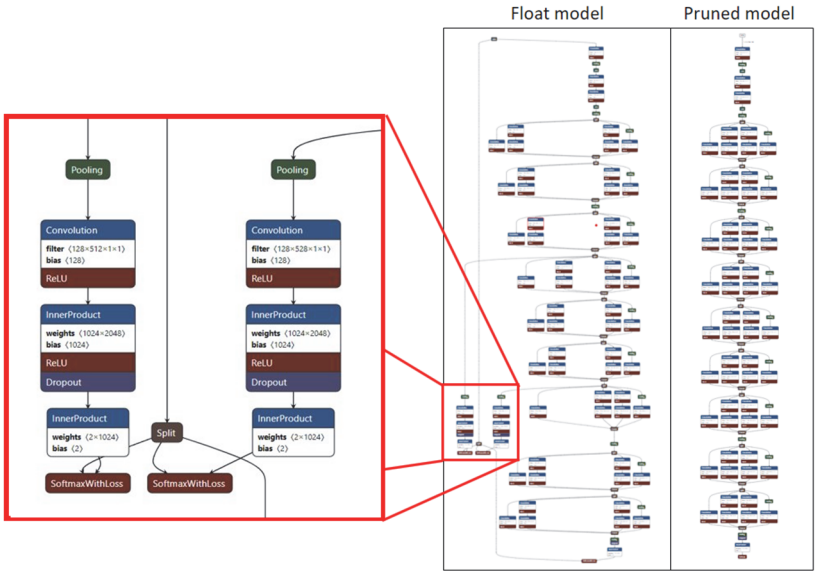


*Fig. 3. Visualization of the float original neural network and the pruned version of the same network*

The DPU implementation is further reducing power consumption to an extent of 9.864 Watts. So, the tradeoff is 4.9% in accuracy vs 63.21% power reduction.

Although the implemented DPU has high utilization percent, the tested neural network is relatively small, meaning that the power efficiency can increase further for bigger neural networks.

**Conclusion.** Based on the analysis of power metrics for neural networks after pruning and quantization and their implementation on DPUs, the research demonstrates that these optimization techniques reduce power consumption by 63% while maintaining high accuracy levels of 92.8%. This balance showcases the efficiency of DPUs in executing complex models with reduced energy requirements. The findings highlight the potential for wider application and sustainability of AI in power-constrained environments, pointing towards a future where deep learning is more accessible on edge devices.

## REFERENCES


1. **Nagel, M., Fournarakis, M., Amjad, R.A., Bondarenko, Y., van Baalen, M., & Blankevoort, T.** A White Paper on Neural Network Quantization. Qualcomm AI Research. – 2021.
2. **Hongrong Cheng, Miao Zhang,** Javen Qinfeng Shi A Survey on Deep Neural Network Pruning-Taxonomy, Comparison, Analysis, and Recommendations – ArXiv. – 2023.
3. https://www.xilinx.com/products/intellectual-property/dpu.html#documentation
4. **Tan, M., & Le, Q.V.** EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. – ArXiv. - 2019.
5. Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference/ **B. Jacob, B.S. Kligys, B. Chen, M. Zhu, et al,** Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). - 2018.
6. **Song Han, Jeff Pool, John Tran, William J. Dally** Learning both weights and connections for efficient neural network // Advances in Neural Information Processing Systems. - 2015.- P. 1135-1143.
7. **Chen Y.-H., Krishna T., Emer J.S. and Sze V.** Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks//IEEE Journal of Solid-State Circuits, 2017.- Vol. 52, no. 1.- P. 127-138.


National Polytechnic University of Armenia. The materials is received on 29.01.2024.

Ա.Ա. ԱՎԵՏԻՍՅԱՆ

# ՆԵՅՐՈՆԱՅԻՆ ՑԱՆՑԵՐԻ ԶԱՏՈՒՄԻՑ, ՔՎԱՆՏԱՑՈՒՄԻՑ, DPU-Ի ՎՐԱ ԻՐԱԿԱՆՈՒՑՄԻՑ ՀԵՏՈ ՀԶՈՐՈՒԹՅԱՆ ԱՆՋԱՏՄԱՆ ՈՒՍՈՒՄՆԱՍԻՐՈՒՄ: ՀԱՄԵՄԱՏԱԿԱՆ ՎԵՐԼՈՒԾՈՒԹՅՈՒՆ

Տարբեր ոլորտներում խոր ուսուցման կիրառությունների արագ տարածումը ցույց տվեց նեյրոնային ցանցերի արդյունավետ ներդրման անհրաժեշտությունը, հատկապես ռեսուրսներով սահմանափակ եզրային սարքերում: Ի պատասխան այս պահանջի՝ զատումը և քվանտացումը առաջացել են որպես նեյրոնային ցանցերի հաշվողական և հիշողության պահանջները էականորեն նվազեցնելու եղանակներ: Բացի այդ, մասնագիտացված սարքերի տեղակայումը, որոնցից են թվային մշակման միավորները (DPU), դարձել է նեյրոնային ցանցերի արագագործությունը մեծացնելու կիրառվող եղանակ:

Աշխատանքում ներկայացվել է նեյրոնային ցանցերի հզորության չափումների համապարփակ համեմատական վերլուծություն զատումից և քվանտացումից հետո՝ հատուկ ուշադրություն դարձնելով դրանց՝ DPU սարքերի վրա իրականացմանը: Հետազոտության նպատակն է ուսումնասիրել զատված և քվանտացված նեյրոնային ցանցերի էներգաարդյունավետությունը և հզորության սպառումը DPU հարթակներում գործելու ժամանակ: Ուսումնասիրված են մոդելի չափերի կրճատման և արդյունքի ճշգրտության փոխզիջումները, ինչպես նաև տարբեր DPU ճարտարապետությունների էներգաարդյունավետությունը:

Արդյունքները DPU հարթակներում զատված և քվանտացված նեյրոնային ցանցերի էներգաարդյունավետության մասին տեղեկություններ են տրամադրում՝ և տալիս հստակ պատկերացում լավարկման այս մեթոդի առավելությունների և փոխզիջումների մասին:

Հետազոտությունը էներգախնայողության համար նեյրոնային ցանցերի ներդրման լավարկմամբ հետաքրքրվող հետազոտողների, մշակողների և իրագործողների համար արժեքավոր ռեսուրս է տրամադրում: Գտածոները նպաստում են եզրային սարքերում և էներգիայի սահմանափակումով այլ միջավայրերում խոր ուսուցումն ավելի մատչելի և կայուն դարձնելու շարունակական ջանքերին՝ ի վերջո ապահովելով ցածր էներգասպառմամբ սարքերի ավելի լայն կիրառությունների հնարավորություն:

***Առանցքային բառեր.*** FPGA, DPU, զատում, քվանտացում:

**А.А. АВЕТИСЯН**

# ИССЛЕДОВАНИЕ ПОКАЗАТЕЛЕЙ ПОТРЕБЛЯЕМОЙ МОЩНОСТИ НЕЙРОННЫХ СЕТЕЙ ПОСЛЕ СОКРАЩЕНИЯ, КВАНТОВАНИЯ И РЕАЛИЗАЦИИ НА DPU. СРАВНИТЕЛЬНЫЙ АНАЛИЗ

Быстрое распространение приложений глубокого обучения в различных областях вызывает необходимость эффективных реализаций нейронных сетей, особенно на периферийных устройствах с ограниченными ресурсами. В ответ на этот спрос обрезка и квантование стали важными методами снижения вычислительных требований и требований к памяти нейронных сетей. Кроме того, развертывание специального оборудования, такого как цифровые процессоры (DPU), набирает обороты для ускорения вывода нейронных сетей.

Представлен всесторонний сравнительный анализ показателей потреблямой мощности нейронных сетей после обрезки и квантования с особым акцентом на их реализацию на DPU. Целью данного исследования является изучение энергоэффективности и энергопотребления сокращенных и квантованных нейронных сетей при их выполнении на платформах DPU. Исследуются компромиссы между уменьшением размера модели и точностью вывода, а также энергоэффективностью различных архитектур DPU.

Результаты дают представление об энергоэффективности сокращенных и квантованных нейронных сетей на платформах DPU, предлагая четкое понимание преимуществ и компромиссов, связанных с этими методами оптимизации.

Исследование предоставляет ценный ресурс для исследователей, разработчиков и практиков, заинтересованных в оптимизации реализации нейронных сетей для повышения энергоэффективности. Полученные результаты способствуют постоянным усилиям, направленным на то, чтобы сделать глубокое обучение более доступным и устойчивым на периферийных устройствах и в других средах с ограниченным энергопотреблением, что в конечном итоге позволит использовать более широкий спектр приложений с меньшим потреблением энергии.

***Ключевые слова:*** FPGA, DPU, обрезка, квантование.