NEW METHOD FOR DESIGN NEURAL MODULAR ACCELERATOR GENERATOR: ECONOMIC AND MODELING ISSUES

Karo SAFARYAN
R&D Engineer, NPUA, PhD
Arman GALSTYAN, Armen DANIELYAN,
Arman MANUKYAN, Davit GABRIELYAN
A&MS Circuit Design Engineers, NPUA,
Chair of Microelectronic Circuits and Systems, MA Students

Key words: Deep neural networks, Modular Accelerator Generator

Economic significance. Neural networks are one of a variety of machine learning models which are beginning to be widely used in economic forecasting applications. Despite this, there is relatively little understanding of the conditions in which neural networks provide accurate forecasts, the uncertainty bounds which can be put on such forecasts, and the most suitable network types and parameters for forecasting in the relatively small-sample settings encountered within economics. There are mixed comparison results of forecasting performance between neural networks and other models. The reasons may be the difference of data, forecasting horizons, types of neural network models and so on. Prediction performance of neural networks can be improved by being integrated with other technologies.

Introduction. As a powerful technique to handle complicated issues across a wide range of application sectors, deep neural networks (DNNs) had recognized as a dominant strategy in several fields, including autonomous driving, health care, robotics, processing of natural language, and image recognition [1]. The expectations placed on DNN prediction applications can vary dramatically from one another. For neural networks, we introduce a Modular Accelerator Generator (MAG) to tackle these difficulties. Neural MAG receives the following information in addition: 1) a target application requirements composed of a range of neural networks; 2) physical limitations, such as a surface budget or a time objective; and 3) a design purpose in terms of production and energy consumption, among other variables. A valid mapper for executing DNN prediction on a specific system is produced as an export by MAG and synthesizable RTL in an ASIC for DNN acceleration [2]. There are three parts to the neural MAG software package: the MAG Designer, the MAG Mapper, and the MAG Tuner [3]. The Neural MAG architecture comprises a highly flexible architecture framework with many design-time options, which allows the production of DNN accelerators tailored to individual workloads and application scenarios. The MAG mapper manages multiple neural networks onto the created electronics and allows for refining mapping algorithms at run-time, which is very useful. For a quick exploration of the design space, the MAG tuner takes advantage of

Function optimization, a well-known methodology for searching multifaceted areas. Through RTL High-Level Synthesizer (HLS), power analysis, logic fabrication, several potential accelerator executions may be quickly prototyped, allowing for the most efficient design model validation for the required DNN accuracy productivity and efficiency targets. This paper offers neural MAG, a DNN induction accelerator generator built around a highly flexible Processing Element (PE) conceptual structure written in C++ and synthesizable by C++ and HLS tools.

Literature Review. Many research projects have been undertaken to enhance the performance and reliability of DNN inference acceleration systems. In terms of accelerator layout, early current studies investigated a variety of accelerator designs with varying memory topologies, dataflow patterns, and connectivity networks [4]. These designs use a variety of reuse patterns to optimize the design for a variety of target workloads. MAG tackles the industrial automation aspects regarding specialist computational intelligence accelerators in contrast to earlier attempts that described particular design cases. Neural MAG allows for systematic examination of design space and founder of accelerator concept and run-time characteristics, including alternative segments and sub-parameters, data flows, patterning schemes, and quantization strategies [5]. Other researchers have advocated using FPGAs and the accompanying intelligent technologies to accelerate deep neural networks in previous research initiatives. Most of these efforts concentrate on speeding a distinct neural network and on developing the most efficient FPGA resource best practices. Also of note is that they offer a fixed flow of information and do not consider run-time transfer schemes for performing various neural networks on different hardware configurations. It is targeted at ASIC multiprocessors and supports the structure of various design-time and run-time specifications, including equipment variables, dataflows, and designs, to enhance the prototype across a wide variety of neural networks [6].

Methodology: Data Sources. This study pruned allocated capacity by placing two user-determined criteria, emphasizing either energy consumption or effectiveness. From the first limitation, we have the recycling factor, which sets the modest amount of temporally reuse may achieve that for various data kinds. The reuse factor can be used to prune small tile sizes that are inefficient to improve energy due to a lack of buffer reuse. This property is the penultimate limitation since it defines the frequency of PEs actively used in the system. The usage of more PEs results in a bigger PE tile size, which usually results in greater energy efficiency inside the PE. However, research shows that PE may diminish the overall performance if only a few PEs are used. A higher capacity leads to enhanced software quality, but it may not be the most energy-efficient option. ResNet-50 projected to MAG produced technology with a WS dataflow, maps the using various pruning algorithms and different clipping heuristics. As illustrated in the image, using a

minimal recycle factor in conjunction with utilization limits results in a decrease in the size of the mapped space of about 90 per cent.

Research Design. The findings of the design space exploration are depicted in Table 1. A compromise between energy consumption (y-axis) and effectiveness per unit area (x-axis) is displayed in the histogram; the bottom corner portion of the graph represents the ideal tradeoff. There are four broad categories of MAG designs presented here, including one that corresponds to a specific pairing with weight and stimulation precision and recalls in the design. Each location on the graph represents the measurements for a composite design occurrence chosen from a set of variables possibilities listed in Table 1. Each point represents a different design occurrence. Every classification is optimized using three distinct neural networks and pruning the mapped space aggressively to get the most feasible mapping. Lastly, we present the findings of a MAG instance set up to be identical to NVDLA, a transparent DNN accelerator that is utilized in comercialized SoCs, to provide a baseline for comparability. NVDLA makes use of WS dataflow, which has precisions of 8 bits for both weighting and actuation. Two separate frequency settings are included in the parametric sweeps for this masterpiece.

Statistical Analysis. Our evaluation case study involves designing a DNN interprettation accelerator for object recognition utilizing three distinct computational models, DriveNet and ResNet-50, and the ImageNet groups of data to assess MAG's performance. Using these machine learning, we investigate the constraints between energy consumption and efficacy in each of their fully connected layers, as well as a handful of MAG-generated reasoning accelerator configurations in a 14nm FinFET device dimensions for these neural systems. The product designs and characteristics used in the examination are listed in Table 1.

Table 1. Experimental Setup

Benchmarks	
Networks	AlexNet, ResNet-50, DriveNet
Dataset	ImageNet
Design Tools	
HLS Compiler	Mentor Graphics Catapult HLS
Verilog simulator	Synopsys VCS
Logic synthesis	Synopsys Design Compiler Graphical
Place and Route	Synopsys ICC2
Power Analysis	Synopsys PT-PX
Design Space	
Dataflows	WS, OS, WS-LOS, OS-LWS
VectorSize/NLanes	4, 8, 16
Weight/Activation precision	4 bits, 8 bits
Accumulation precision	16 bits, 20 bits, 24 bits
Weight Collector Size	8B - 2KB
Accumulation Collector Size	8B - 384B
Input Buffer Size	2KB, 8KB, 16KB
Weight Buffer Size	4KB - 128KB
Accumulation Buffer Size	1KB - 6KB
Global Buffer Size	64KB
Target Frequencies	500 MHz, 1 GHz
Supply Voltage	0.6V

Results and Discussion. The figures below show results for this research design:

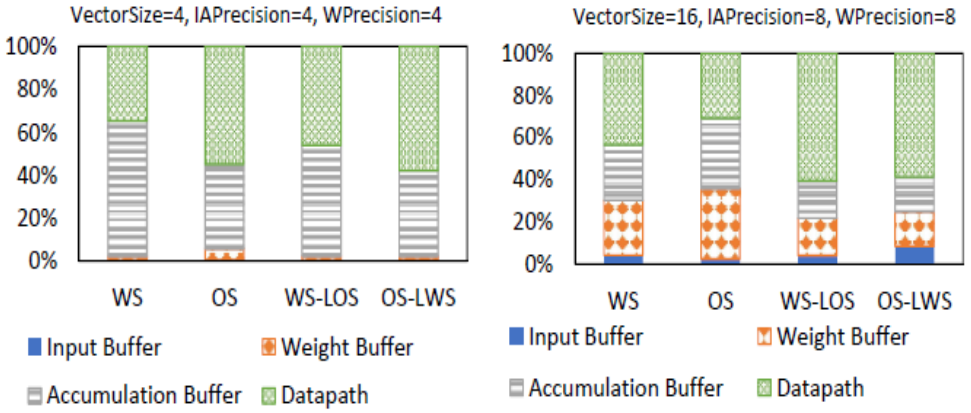

Figure 1. Effect of Parameters on Energy Efficiency

Figure 1 describes the correlation between the energy consumption of ResNet-50 and the precision, vector size and dataflow parameters. Appropriate Dataflow is defined as follows: Given the findings in Figure 1, we would like to integrate and vast important implications for dataflow selection. For starters, we discovered that the OS-LWS workflow provides the highest level of energy effectiveness in practically all cases. The benefits are most noticeable when 8-bit accuracy and a big VectorSize are used together.

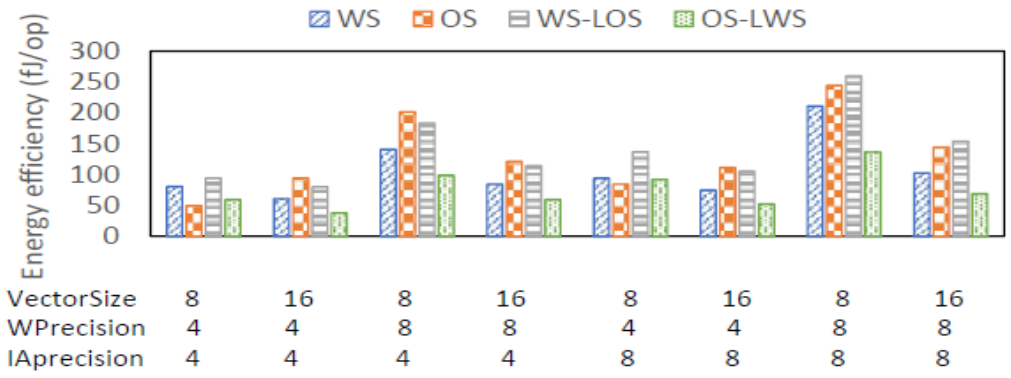

Figure 2. Neural MAG PE Energy Breakdown

As illustrated in Figure 2, the weighted reserve and the accretion buffer account for a large portion of the total energy usage in these arrangements. OS-LWS can produce lower emissions than the WS and OS dataflows by lowering the number of accesses to some of these stores. WS-LOS dataflows spend significantly more energy than standard dataflows. This trend is revealed in another insight: the unexpected increase in energy usage. Consider the mapping that the neural MAG mapper selected to understand this effect better. The mapper employed the C and K measurements of the convolution kernel

sequence to map to the PE's VectorSize and NLanes characteristics, which we obtained from the convolution loop-nest. WS-LOS also utilizes the C dimension for the objective of temporal reuse within collectors. For Resnet-50 layer upon layer with reduced C measurements as a function, WS-LOS accomplishes little reuse yet incurs a significant energy cost due to additional detectors. OS-LWS, on the other hand, capitalizes on. In the collectibles, temporal recovery along the Q dimension can be achieved, resulting in improved recycling and waste reduction overall. Aside from that, one of the Vector MAC elements' supplies remains static throughout the OS-LWS information flow, resulting in decreased noises and improved energy consumption. According to a third perception, for configurations with 4-bit accuracy and smaller vector sizes, the OS dataflow yields somewhat significant environmental benefits than the OS-LWS, even though a greater proportion of the energy is consumed by stacking buffers.

Increase in vector size: Except for maybe the OS dataflow, growing the vector size results in the decrease in energy usage due to greater retransmission of weighted sum for all other ultimately increases. For the operating system data flow, the involvement of abundance barrier vitality is trivial, and the stepwise improvement spectrum utilization from a significantly bigger vector size is negligible. In contrast, other overheads associated with a greater vector size, such as broader feedback flashbacks and catalogues, increase actual energy consumption.

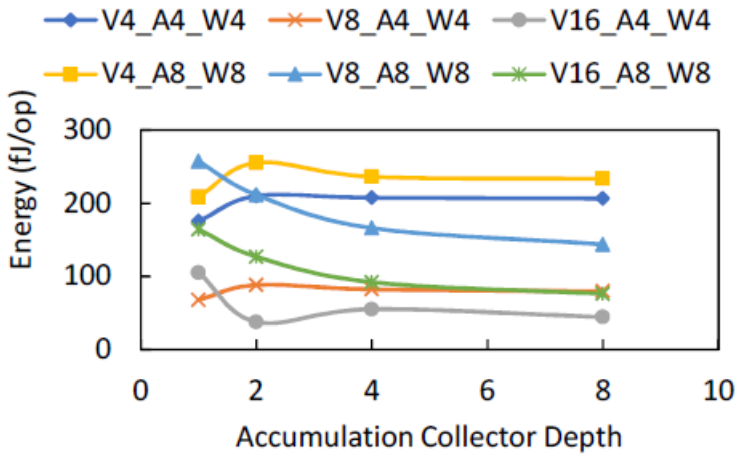

Figure 3. Accumulation Sensitivity

Collection collector depth: In the OS-LWS dataflow, Figure 3 illustrates the impact of raising the consolidation accumulator intricacy for diverse setups. When Vx Ay Wz is represented as a design, the vector size equals x, IAPrecision equals y, and precision is comparable to z. The size of the accumulator collector influences the amount of weight reclamation that we incorporated into a design. By increasing the capacity of the aggregation collector, more weight can be recycled, resulting in a reduction in

electricity consumption by the weight buffer. It does, however, increase the amount of energy consumed by the aggregate collector registries as a result of this. Because the mass pad contributes only a tiny percentage of the overall energy in configurations with reduced weighted accuracy and limited spatial recycle of partial aggregates, stimulating the growth of the accumulation collector results in an increase in electricity consumption in these configurations. When weight measurements are more significant, and vector sizes are more extensive, enhancing the aggregate collector capacity results in a reduction in electricity depletion leads to improved temporal recycling, as shown in the graph below.

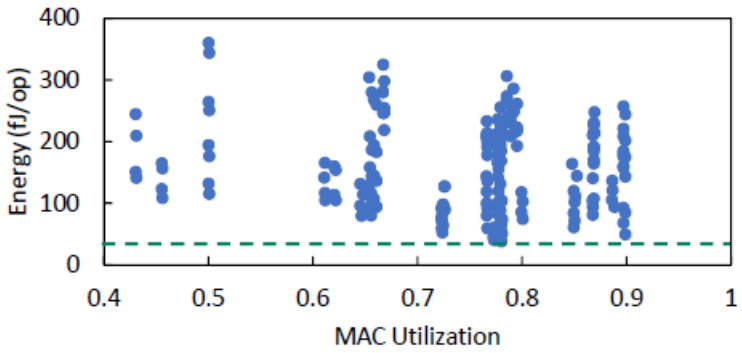

Figure 4. MAC Utilization

MAC Use: Figure 4 depicts the relationship between energy and MAC application for several neural MAG-developed configurations. As illustrated in the figure, neural MAG configurations can achieve approximately 90 per cent MAC occupancy, resulting in highly high-performance solutions for ResNet-50 networks. Furthermore, for a particular application, a diverse range of designs are available that provide varying degrees of energy consumption, indicating the importance of systematic inventor investigation. Utilization is valuable criterion for reducing design space since designs with very low utilization (less than 50 per cent) are poor in terms of quality and power consumption.

Conclusion. This study introduces neural MAG, a DNN information accelerator producer that accelerates assessment. It is proposed that a highly flexible architectural template, including configurable dataflows, be used with a mapper to map neural networks onto an embedded system. Offered unique dataflows that use reuse across weights and component sums to achieve high performance. Neural MAG's efficacy in enhancing the energy speed and reliability is demonstrated by the generation of real-world hardware equivalents, analysis of post-synthesis findings, and measurement of power on real-world workload. Data indicate that only a neural MAG-developed accelerator can reach 2.6 TOPS/mm2 and 30 fJ/op with a 14nm FinFET technology, beating jurisdiction DNN inference accelerators. This result is achieved by modifying the neural network,

Dataflow, and architecture simultaneously. Therefore, ASIC and RTL combined neural design yields improvements in energy efficiency.

References

[1] Wu, S., Zhong, S., & Liu, Y. “Deep residual learning for image steganalysis. Multimedia tools and applications”. (2018).
[2] Zhang, W., Li, X., & Ding, Q. “Deep residual learning-based fault diagnosis method for rotating machinery.” ISA transactions (2019).
[3] Jung, J., Park, J., & Kumar, A. “A verification Framework of Neural Processing Unit for super resolution”. In 20th International Workshop on Microprocessor/SoC Test, Security and Verification (MTV), (2019).
[4] Lee, J., Kim, C., Kang, S., Shin, D., Kim, S., & Yoo, H. J. “An energy-efficient deep neural network accelerator with fully variable weight bit precision”. IEEE Journal of Solid-State Circuits (2018).
[5] Ma, Y., Cao, Y., Vrudhula, S., & Seo, J. S. “An automatic RTL compiler for high-throughput FPGA implementation of diverse deep convolutional neural networks”. In 27th International Conference on Field Programmable Logic and Applications (FPL) (2017).
[6] Sun, Y., Huang, S., Oresko, J. J., & Cheng, A. C. (2010). “Programmable neural processing on a smartdust for brain-computer interfaces”. IEEE transactions on biomedical circuits and systems (2010).

Կարո ՍԱՖԱՐՅԱՆ, Արման ԳԱԼՍՏՅԱՆ, Արմեն ԴԱՆԻԵԼՅԱՆ, Արման ՄԱՆՈՒԿՅԱՆ, Դավիթ ԳԱԲՐԻԵԼՅԱՆ

Նեյրոնային մոդուլյար արագացուցիչի գեներատորի նախագծման նոր մեթոդ. Տնտեսագիտական և մոդելավորման հարցեր

Բանալի բառեր. խորը նեյրոնային ցանցեր, մոդուլյար արագացուցիչի գեներատոր

Աշխատանքի նպատակն է նեյրոնային ցանցերի մշակմանը ինտեգրել ASIC և RTL սխեմաներ: Բազմաթիվ հետազոտական նախագծեր են ձեռնարկվել DNN արագացման համակարգերի արդյունավետությունն ու հուսալիությունը բարձրացնելու համար: Այս հոդվածում առաջարկվում է նեյրոնային մոդուլյար արագացուցիչի գեներատորի (MAG) նախագծման նոր մեթոդ, որը կառուցված է ASIC-ի ճկուն պրոցեսորային տարրի (PE) կառուցվածքի շուրջ և սինթեզվում է RTL գործիքներով: Այս մեթոդի կիրարմամբ հնարավոր է կրճատել նեյտրոնային ցանցերի մշակման ծախսերը, քանի որ ASIC և RTL նախագծերն ավելի էժան են, շնորհիվ բարձր ճկունությամբ ճարտարապետական դիզայնի: ASIC և RTL նախագծերը կարող են օգտագործվել քարտեզագրողի հետ՝ ներկառուցված համակարգի վրա նեյրոնային ցանցերը քարտեզագրելու համար:

Karo SAFARYAN, Arman GALSTYAN, Armen DANIELYAN, Arman MANUKYAN, Davit GABRIELYAN

New method for design neural modular accelerator generator. Economic and modeling issues

Key words: Deep neural networks, Modular Accelerator Generator

The primary purpose of this paper is to integrate ASIC and RTL designs in the development of neural networks. Even though numerous research studies worked on different neural network designs using ASIC and RTL, very few researchers have worked on integrated ASIC-RTL neural network designs. This paper offers neural MAG, a DNN induction accelerator generator built around a highly flexible Processing Element (PE) conceptual structure of ASIC and synthesizable by RTL tools. Many research projects have been undertaken to enhance the performance and reliability of DNN inference acceleration systems. In terms of accelerator layout, early current studies investigated a variety of accelerator designs with varying memory topologies, dataflow patterns, and connectivity networks. However, ASIC stand-alone neural network designs are pretty expensive. This paper seeks to reduce the cost of constructing neural networks by integrating ASIC and RTL designs. This study concludes that ASIC and RTL designs are cheaper because they constitute a highly flexible architectural design, including configurable dataflows, and be used with a mapper to map neural networks onto an embedded system.

Каро САФАРЯН, Арман ГАЛСТЯН, Армен ДАНИЕЛЯН, Арман МАНУКЯН, Давид ГАБРИЕЛЯН

Новый метод проектирования нейронового модульного ускорительного генератора. Экономические и модельные проблемы.

Ключевые слова: Глубокие нейронные сети, Генератор модульного ускорителя

Целью данной статьи является интеграция проектов ASIC и RTL при разработке нейронных сетей. Было предпринято множество исследовательских проектов для повышения производительности и надежности систем ускорения вывода DNN. В этой статье предлагается нейронный MAG, генератор индукционного ускорителя DNN, построенный на очень гибкой концептуальной структуре элемента обработки ASIC и синтезируемый инструментами RTL. Используя этот метод, можно снизить стоимость разработки нейтронных сетей, поскольку проекты ASIC и RTL дешевле из-за очень гибкой архитектурной конструкции. Проекты ASIC и RTL могут использоваться с картографом для отображения нейронных сетей во встроенной системе.