NEW METHOD OF A LOW-COST NEURAL NETWORK: ECONOMIC AND MODELING ISSUES

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

Keywords: Neural MAGNet architecture, FPGA, ASIC embedded system, deep neural networks.

Economic significance. Deep Neural networks (DNNs) have been widely applied to finance and economic forecasting as a powerful modeling technique. Economic fundamentals are important in driving exchange rates, stock market index price and economic growth. Most neural network inputs for exchange rate prediction are univariate, while those for stock market index prices and economic growth predictions are multivariate in most cases. Prediction performance of neural networks can be improved by being integrated with other technologies. Nonlinear combining forecasting by neural networks also provides encouraging results. Finally, this method is compared and contrasted with standard (statistical) approach on real economic data to show the potential of using deep neural network in modelling economic variables. For all these reasons the low-cost neural networks important in economics.

Introduction. With the mass acceptance of deep neural networks (DNN) across a wide range of functional domains, there has been a significant increase in the usage of inference processors. However, because of the potentially severe consequences of the ASIC embedded system, it is not feasible. Custom accelerators for a given criterion can be challenging to develop. To bring it down. To reduce design costs, we offer MAGNet, a customizable accelerator producer for parallel processing. Neural networks are a type of computer network that uses neural connections to process information. MAGNet accepts access credentials that consist of one or more components as input and output, numerous neural networks, and technological limits that enable the creation of RTL for a neuromorphic ASIC. In addition to effective mappings for conducting the aim systems on the developed data interface [1].

Tuner, mapper, and designer are the three components of the neural MAGNet software package: the MAGNet Tuner is the third component [3]. In the Neural MAGNet architecture, a highly flexible architecture framework with several design-time variables creates DNN accelerators specifically customized to specific workloads and application scenarios. The MAGNet mapper handles many neural networks onto the electronics that have been developed and enables the refinement of mapping algorithms

while they are running, which is quite valuable. When conducting a rapid exploration of the design space, the MAGNet tuner employs Function optimization, a well-known tool for investigating diverse areas of interest. Several alternative accelerator executions may be swiftly prototyped using the RTL High-Level Synthesizer (HLS), power analysis, and logic manufacturing, enabling the most efficient design model validation for the desired DNN accuracy productivity and efficiency requirements. A DNN induction accelerator generator called neural MAGNet is presented in this research. It is constructed around a highly flexible Processing Element (PE) conceptual structure written in C++ and synthesized by HLS and C++ tools, and it is described in detail.

The figure below shows the overview of the research design:

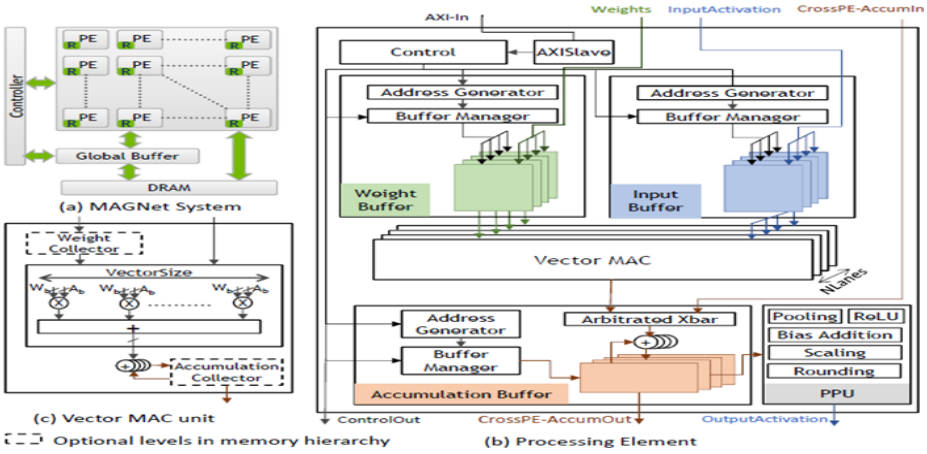

Figure 1. Architecture of MAGNet

Importance of MAGNet. There have been many research efforts conducted to improve the performance and reliability of DNN inference acceleration systems. As for the layout of the accelerator, early contemporary research looked at certain accelerator formulations with diverse memory topologies, dataflow patterns, and connectivity networks, as well as different dataflow patterns and connectivity networks [3]. A range of reuse patterns is employed in these designs to optimize the design for various target workloads. In contrast to previous attempts, MAGNet addresses the industrial automation features of specialty computational intelligence accelerators instead of preceding initiatives, which specify specific design instances. The use of neural MAG allows for the systematic analysis of design space and the founder of accelerator idea and run-time aspects, such as alternate segments and sub-parameters, data flows, patterning schemes, and quantization strategies things [4]. In earlier research endeavors, other researchers have pushed for FPGAs and the associated intelligent technologies to speed deep neural networks. Other researchers have supported this. The majority of these efforts are focused on increasing the speed of a particular neural network and producing the most efficient FPGA resource best practices possible [5].

Data Flows. Various dataflows are depicted in the form of loop nests in Figure 2. The looping constraints for every data flow are set by the MAGNet coder, which is a mathematical formula. The weight-stationary (WS) flow of information under which the external loops apply the weight dimensionality (C1; K1; S; R) to salvage weights spanning vector MAC computations with various parameter vectors, as shown in the middle loops. Because the loop nested is specified for one PE, the measurements C1; K1 are applied instead of C; K to indicate that every PE is responsible for a component of the all-inclusive spectrum rather than the complete range. The single-entry weight collectors and the omission of the aggregation collector in the vector MAC unit are used to construct a DNN accelerator with a WS dataflow using the MAGNet algorithm. Creating the proper order of address for the local buffers is accomplished by establishing the address synthesizer logic to provide the exact categorization of discourses for the local buffers [6]. The figure below illustrates the data flow:

```
for k1=[0:K1):
 for r=[0:R):
  for s=[0:S):
   for c1=[0:C1):
    // WS
    for p1=[0:P1):
     for q1=[0:Q1):

        Vector MACs
```

(a) WS dataflow

```
for k1=[0:K1):
 for p1=[0:P1):
  for q1=[0:Q1):
   // OS
   for r=[0:R):
    for s=[0:S):
     for c1=[0:C1):

        Vector MACs
```

(b) OS dataflow

```
for h1=[0:H1):
 for w1=[0:W1):
  for c1=[0:C1):
   // IS
   for k1=[0:K1):
    for r=[0:R):
     for s=[0:S):

        Vector MACs
```

(c) IS dataflow

```
for k1=[0:K1):
 for r=[0:R):
  for s=[0:S):
   for c1=[0:C1):
    // WS
    for p1=[0:P1):
     for q1=[0:Q1):
      // LOS
      for c0=[0:C0):
       Vector MACs
```

(d) WS-LOS dataflow

```
for k1=[0:K1):
 for p1=[0:P1):
  for q1=[0:Q1):
   // OS
   for r=[0:R):
    for s=[0:S):
     for c1=[0:C1):
      // LWS
      for q0=[0:Q0):
       Vector MACs
```

(e) OS-LWS dataflow

Figure 2. Data Flows

Results And Discussion. For all AlexNet layers integrated, the energy summary across specific elements of something like the PE for WS and OS ultimately increases as shown in Figure 3 above with VectorSize = 6, NLanes = 6, 6-bit sophistication for inputs and weights, and 16-bit exactness for restricted sums using the parameters VectorSize = 6, NLanes = 6, and NLanes = 6. In WS information flow, which seems tuned for the time-based regeneration of weights, the buffer weight provides only a tiny fraction of total energy expenditure. In contrast, the accumulated buffer represents a considerable proportion of total energy consumption. When using an OS dataflow, is from the other hand, you optimize the cumulative buffer performance at the expense of increasing the weight buffering energy. Because of the restricted write/read bandwidth and spatial re-application over multiple lanes of the PE, the relative importance from the input buffer is tiny in each scenario. As a result, the energy conservation from input quiescent data flow is also tiny. Non - linear and non-dataflows are proposed to handle the difficulty of maximizing both the weight and the buildup buffer energy simultaneously. The figures below show results for the study:

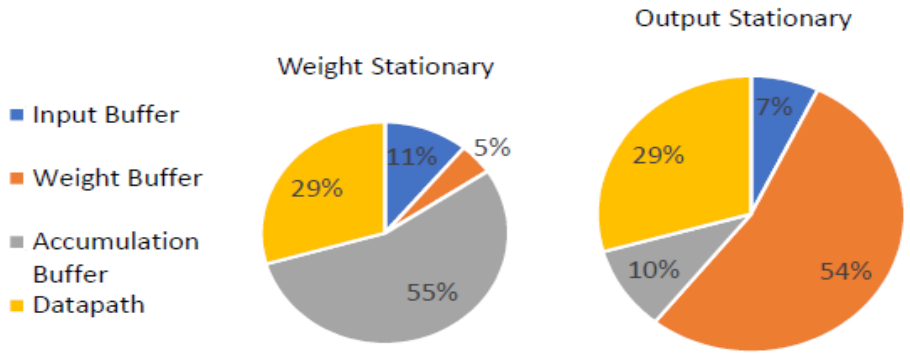

Figure 3. Energy Breakdown

The weighted reserve and the accretion buffer account for a significant fraction of the overall energy consumption incurred in these configurations. Due to a reduction in the number of accesses to ensure this storage, the OS-LWS dataflows can emit less pollution than the WS and OS dataflows. WS-LOS dataflows consume a large amount of energy compared to conventional dataflows. Interestingly, another observation reveals this trend: the unanticipated growth in energy consumption. Consider the mapping that the neural MAG mapper chose to gain a better understanding of this effect. When we acquired the PE's VectorSize and NLanes characteristics from the convolution loop-nest, we used the C and K measurements from the convolution kernel sequence to translate them to the PE's VectorSize and NLanes characteristics. WS-LOS also utilizes the C dimension to achieve the goal of temporal reuse within collectors. WS-LOS achieves low reuse for Resnet-50 layers upon layers with reduced C measurements as a function. Still, it incurs a considerable energy cost due to the installation of extra detectors in the process. OS-LWS, on the other hand, takes advantage of it. It is possible to accomplish temporal recovery along the Q dimension in the collectibles industry, resulting in increased recycling and waste reduction overall. Additionally, one of the Vector MAC elements' supply remains static during the OS-LWS information flow, resulting in lower noise levels and lower energy consumption overall. For example, according to a third perception, for setups with 4-bit accuracy and smaller vector sizes, the OS dataflow provides better environmental benefits than the OS-LWS, even if the stacking buffers consume a more significant share of the total energy consumed.

Conclusion. This study investigates the relationship between energy consumption and MAC application using a variety of neural MAGNet-developed setups. The figure 3 shows that neural MAG topologies can achieve approximately 90% MAC occupancy, which results in highly high-performance solutions for ResNet-50 networks. For a given application, an extensive range of designs with variable degrees in energy consumption are available, underscoring the significance of conducting rigorous inventor investigations. We provide one-of-a-kind dataflows that use extensive reuse across

weights and component sums to achieve excellent performance. The effectiveness of neural MAGNet in increasing the energy speed and reliability is proved by the development of real-world hardware counterparts, the study of post-synthesis discoveries, and the measurement of power on real-world workload, among other things. According to our findings, only a neural MAGnet-developed accelerator can achieve 2.6 TOPS/mm2 and 30 fJ/op using a 14nm FinFET technology, outperforming conventional DNN inference accelerators in both performance and efficiency. This result is accomplished by altering the neural network, Dataflow, and architecture simultaneously, in parallel. As a result, improved energy efficiency is achieved by using ASIC and RTL integrated neural architecture.

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”. 20th Int.'l Workshop on Microproc./SoC Test, Secur. & Verification (MTV), (2019).
[4] Lee, J., Kim, C., Kang, S., Shin, D., Kim, S., Yoo, H. An energy-efficient deep neural network accelerator with fully variable weight bit precision. IEEE Journal Solid-State Circuits 2018.
[5] Ma, Y., Cao, Y., Vrudhula, S., Seo, J. An automatic RTL compiler for high-throughput FPG implementation of diverse deep convolutional neural networks 27th Int.'l Conf. on FPLA (2017).
[6] Sun, Y., Huang, S., Oresko, J., Cheng, A. 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

Ցածր արժողությամբ նեյրոնային ցանցի կառուցման նոր մեթոդ. Տնտեսագիտական և մոդելավորման հարցեր

Բանալի բառեր. նեյրոնային MAGNet կառուցվածք, FPGA, ներդրված համակարգեր, խորը նեյրոնային ցանց

Այս հոդվածի նպատակն է ուսումնասիրել ցածր արժողությամբ նեյրոնային ցանցերի կառուցման հնարավորությունը: Չնայած նրան, որ բազմաթիվ հետազոտական աշխատանքներ են իրականացվել տարբեր նեյտրոնային ցանցերի նախագծերի վրա, շատ քչերն են մտածել այդ բարդ նեյրոնային ցանցի արժեքը նվազեցնելու մասին: Առաջարկվում է նեյրոնային ցանցի կառուցման նոր մեթոդ խորը նեյրոնային ցանցի ինդուկցիոն արագացուցիչի գեներատորով՝ նեյրոնային MAGNet, որը նվազեցնում է նեյրոնային ցանցերի արժեքը: Նեյրոնային MAGNet գեներատորը գրված է C++-ով լեզվով և սինթեզված է HLS և C++ գործիքներով: Այս մեթոդի կիրառմամբ կարելի է հասնել մոտավորապես 90% MAC-ի զբաղվածության, ինչը հանգեցնում է ResNet-50 ցանցերի համար բարձր արդյունավետ լուծումների ավելի ցածր արժեքով:

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

New method of a Low-Cost Neural Network: Economic and Modeling Issues

Key words: Neural MAGNet architecture, FPGA, ASIC embedded system, deep neural networks

The primary purpose of this paper is to study the possibility of constructing neural networks at a low cost. Even though numerous research studies worked on different neural network designs, very few researchers have considered lowering the cost of these complex neural designs. This paper offers a deep neural network induction accelerator generator called neural MAGNet constructed around a highly flexible Processing Element (PE) conceptual structure written in C++ and synthesized by HLS and C++ tools. Many research studies researchers have pushed for FPGAs and the associated intelligent technologies to speed deep neural networks. However, these neural network designs are pretty expensive. This paper seeks to reduce the cost of constructing neural networks by using MAGNet, which allows for the systematic analysis of design space and addresses the industrial automation features of specialty computational intelligence accelerators instead of preceding initiatives, which specify specific design instances.This study concludes that MAGNet design is cheaper because it can achieve approximately 90% MAC occupancy, which results in highly high-performance solutions for ResNet-50 networks at a lower cost.

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

Новый Метод Бюджетный Нейронной Сети: Экономические И Модельные Проблемы.

Ключевые слова: Нейронная структура MAGNet, FPGA, встроенные системы, глубокая нейронная сеть

Цель этой статьи - изучить возможность построения недорогих нейронных сетей. Несмотря на то, что было проведено много исследований различных конструкций нейтронных сетей, очень немногие задумывались о снижении стоимости этой сложной нейронной сети. В работе предлагается новый метод построения нейронной сети с использованием генератора индукционного ускорителя глубокой нейронной сети, называемого нейронной MAGNet. Новый метод направлен на снижение затрат на построение нейронных сетей. Генератор нейронной сети MAGNet написан на C++ և и синтезирован инструментами HLS և C++. Используя этот метод, можно достичь примерно 90% занятости MAC, что приведет к более экономичным решениям для сетей ResNet-50.