ACHIEVING HIGH SPEED PERFORMANCE AND LOW POWER UTILIZATION IN NEURAL NETWORKS

Sravanthi Pinninti[1], Radhika R[2], Jayasree Das[3], Vijaykumar J[4]
[2]Asst prof, BRECW, [3]Asso Prof, BVRIT

ABSTRACT: Deep learning to know is the subordinate of the gadget learning involved with algorithms stimulated by using the shape & technique of the intelligence referred as artificial neural networks. However, with the increasing accuracy necessities and complexity for the realistic applications, the scale of the neural networks will become explosively massive scale. Therefore, it poses large challenges to put into effect excessive overall performance deep learning networks with low strength fee, mainly for large-scale deep learning neural network models. So far, the brand new way for accelerating deep studying algorithms are field-programmable gate array (FPGA), application specific integrated circuit (ASIC), and Graphic processing unit (GPU). Compared with GPU acceleration, hardware accelerators like FPGA and ASIC can achieve at least slight performance with lower energy consumption. However, each FPGA and ASIC have fantastically confined computing assets, memory, and I/O bandwidths, therefore it is challenging to develop complex and large DNNs using hardware accelerators. To overcome these limitations, we are supplying a scalable deep gaining knowledge of accelerator unit named Deep Learning Accelerator Unit (DLAU) to hurry up the kernel computational elements of deep learning algorithms.

I. INTRODUCTION

Convolutional neural network (CNN), a famous deep learning architecture extended from artificial neural network, has been considerably adopted in various programs, which encompass video surveillance, cellular robot vision, photograph search engine in information centers, etc. Inspired by using the conduct of optic nerves in living creatures a CNN design tactics statistics with a couple of layers of neuron connections to obtain excessive accuracy in image reputation. Recently, rapid increase of current programs primarily based on deep gaining knowledge of algorithms has in addition stepped forward research on deep convolutional neural network. Due to the specific computation sample of CNN, preferred cause processors aren't green for CNN implementation and may hardly ever meet the overall performance requirement. Thus, numerous accelerators primarily based on FPGA, GPU, and even ASIC design has been proposed currently to enhance overall performance of CNN designs. Among these strategies, FPGA primarily based accelerators have attracted an increasing number of interest of researchers because they have advantages of excellent overall performance, excessive strength efficiency, rapid improvement spherical, and functionality of reconfiguration.

A. Deep Neural Networks

Deep neural networks, specifically Deep Belief Networks (DBN), have shown trendy consequences on various laptop imaginative and prescient and popularity responsibilities.

Deep Learning is a sort of Neural Network Algorithm that takes metadata as enter and manner the information thru some of layers of the non-linear transformation of the input records to compute the output. This set of rules has a unique characteristic i.e. computerized characteristic extraction. This means that this set of rules routinely grasps the relevant functions required for the answer of the hassle. This reduces the weight on the programmer to pick the features explicitly.

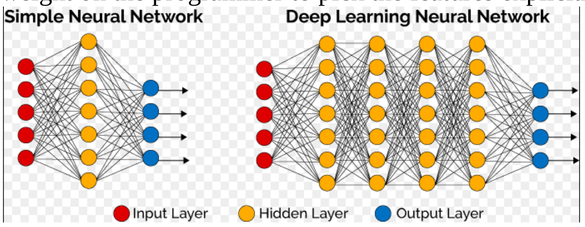


Fig1. Example for Deep Learning Neural network

This may be used to remedy supervised, unsupervised or semi-supervised form of issues. In Deep Learning Neural Network, each hidden layer is chargeable for training the specific set of functions based on the output of the preceding layer. As the wide variety of hidden layers increases, the complexity and abstraction of data also increase.

DBN may be shaped by using stacking Restricted Boltzmann Machine (RBM) on top of every different to assemble a deep community, as proven in Fig2.

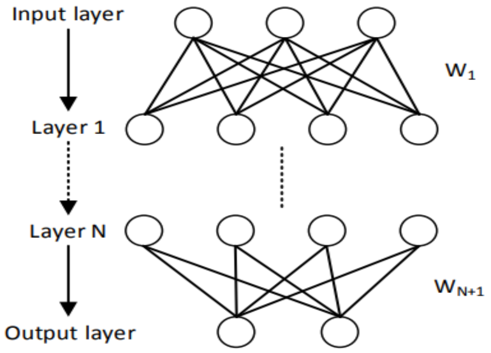


Fig2. N-Layer Deep Neural network

RBMs utilized in DBN are pretrained using Gradient-based Contrastive Divergence (GCD) algorithms, observed through gradient descent and returned propagation algorithms for category and satisfactory-tuning the outcomes.

DBNs are built of a couple of layers of RBMs and a class layer at the quit. The main computation kernel consists of masses of vector-matrix multiplications accompanied by means of non-linear features in every layer. Since multiplications are expensive to put in force in hardware, existing parallel or semi-parallel VLSI implementations of such a network suffer from excessive silicon region and strength intake.

II. RELATED WORK

Convolutional neural community (CNN) has been widely employed for image reputation due to the fact it may achieve excessive accuracy by means of emulating behavior of optic nerves in residing creatures. Recently, rapid growth of present day packages based totally on deep mastering algorithms has in addition stepped forward studies and implementations. Especially, numerous accelerators for deep CNN were proposed based totally on FPGA platform as it has advantages of excessive performance, reconfigurability, and rapid improvement spherical, etc. Although contemporary FPGA accelerators have established better overall performance over everyday processors, the accelerator layout space has now not been nicely exploited. One critical hassle is that the computation throughput may not nicely match the reminiscence bandwidth provided an FPGA platform. Consequently, existing approaches can not obtain great overall performance because of underutilization of either common sense resource or memory bandwidth. At the same time, the increasing complexity and scalability of deep gaining knowledge of programs aggravate this hassle. In order to conquer this trouble, C. Zhang et al proposed an analytical layout scheme using the roofline model. For any answer of a CNN layout, they quantitatively analyze its computing throughput and required memory bandwidth the usage of numerous optimization strategies, such as loop tiling and transformation. Then, with the assist of roofline model, they identified the solution with great overall performance and lowest FPGA resource requirement. As a case have a look at, they implemented a CNN accelerator on a VC707 FPGA board and evaluate it to preceding strategies. Their implementation performed a top overall performance of 61.62 GFLOPS much less than 100MHz operating frequency, which outperforms previous techniques considerably.

T. Chen et al centered on accelerators for machine-studying due to the large set of programs and the few key contemporary algorithms offer the uncommon opportunity to mix high performance and huge application scope. Since contemporary CNNs and DNNs mean very huge networks, they especially focused on the implementation of big-scale layers. By cautiously exploited the locality houses of such layers and through delivered garage structures customized to take advantage of those residences, we display that it's far possible to design a system-gaining knowledge of accelerator able to high overall performance in a very small location footprint. Their measurements aren't circumscribed to the accelerator material, they thing inside the overall performance and power overhead of important memory transfers; nonetheless, they proven that it's miles feasible to reap a speedup of 117.87x and an strength reduction of 21.08x over a 128-bit 2GHz SIMD middle with a normal cache hierarchy.

Deep Belief Nets are an rising system mastering device, which might be based totally on Restricted Boltzmann Machines. FPGAs can correctly make the most the inherent pleasant-grained parallelism in RBMs to reduce the computational bottleneck for huge scale DBN studies. As a prototype of building a fast DBN research machine S. K. Kim, L. C. McAfee, P. L. McMahon, and K. Olukotun carried out a excessive-pace, configurable RBM on a single FPGA. They established a 25X speedup of the RBM implementation on the FPGA in comparison to a unmarried precision software implementation jogging on an Intel Core 2 processor.

Q. Yu, C. Wang, X. Ma, X. Li, and X. Zhou proposed an FPGA-primarily based accelerator for LSTM-RNN. They optimized each computation performance and conversation requirements, and implement an accelerator on a Xilinx VC707 FPGA board. The experimental outcomes proven that their design done tremendous speedup over software program implementations & it outperforms previous LSTM-RNN accelerators as nicely; there are numerous opportunities for similarly research, which includes storing parameters in cautiously quantized fixedpoint facts to reduce resource utilization and improve typical performance. Additionally, they attempted to extend this acceleration framework to some different versions of LSTM-RNNs.

III. FRAMEWORK

A. System Architecture

DLAU system structure carries an embedded processor, a DDR3 memory controller, a DMA module, and the DLAU accelerator. The embedded processor is liable for offering programming interface to the customers and speaking with DLAU via JTAG-UART. In precise it transfers the enter information and the burden matrix to internal BRAM blocks, turns on the DLAU accelerator, and returns the outcomes to the person after execution. The DLAU is incorporated as a standalone unit which is flexible and adaptive to deal with exclusive programs with configurations.

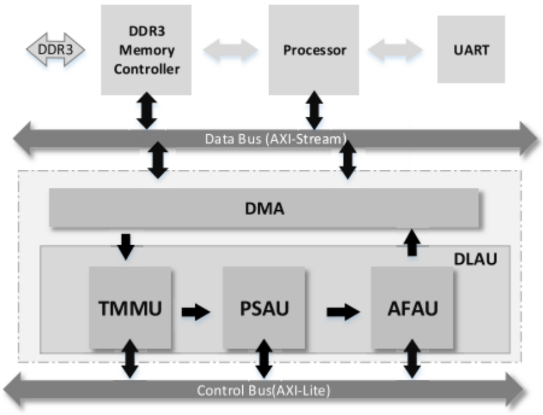


Fig3. Proposed System Architecture

The DLAU has three processing units prearranged in a pipeline approach:

- Tiled Matrix Multiplication Unit (TMMU)
- Part Sum Accumulation Unit (PSAU)
- Activation Function Acceleration Unit (AFAU)

For execution, DLAU reads the tiled facts from the reminiscence through DMA, computes with all the three processing devices in turn, after which writes the outcomes back to the memory.

B. Techniques Used in Proposed DLAU

1. FIFO Buffers

A FIFO buffer is a useful manner of storing information that arrives to a microcontroller peripheral asynchronously however cannot be study immediately. One example is storing bytes incoming on a UART. Buffering the bytes eases the real-time necessities for the embedded firmware. A FIFO buffer stored information on a first-in, first-out foundation. The storage structure is generally an array of contiguous memory. Data is written to the “head” of the buffer and read from the “tail”. When the top or tail reaches the cease of the memory array, it wraps around to the beginning. If the tail runs in to the pinnacle, the buffer is empty. But if the top runs in to the tail, the implementation have to outline if the oldest facts is discarded or the write does now not whole.

2. Pipelined Adder

A famous scheme for acquiring excessive throughput adders is a pipeline wherein each degree contains an array of half-adders acting a bring-shop addition.

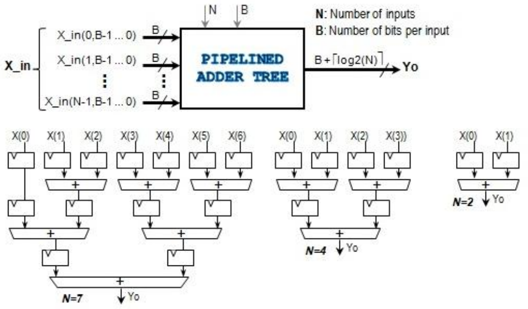
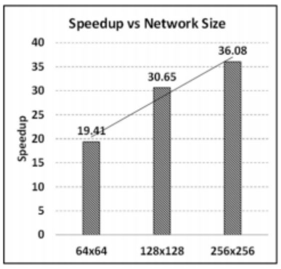


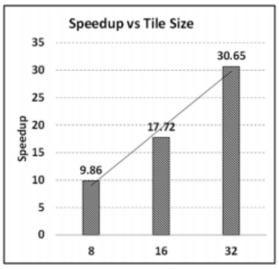
Fig4. Example for Pipelined Adder

In this paper we used pipelined binary adder tree structure to optimize the presentation. The pipeline takes benefit of time-sharing the coarse-grained accelerators.

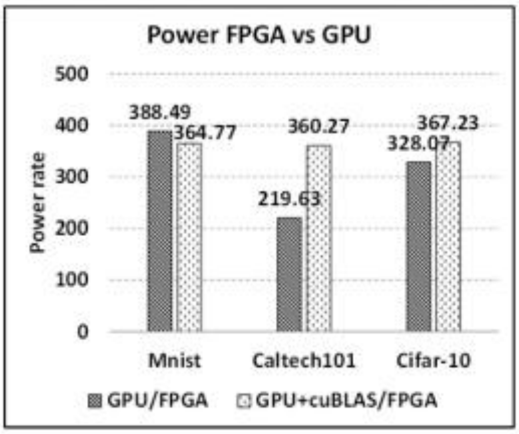
IV. EXPERIMENTAL RESULTS

Experimental results demonstrate that the DLAU is able to reap a reasonable ascendant speedup with the boom of neural networks sizes and additionally demonstrate that the DLAU framework is configurable and scalable with one-of-a-kind tile sizes. The speedup can be leveraged with hardware price to attain enjoyable tradeoffs.





In order to assess the power intake of accelerator, we use Xilinx Vivado tool set to reap electricity fee of every processing unit in DLAU and the DMA module. The results show that the DLAU is pretty power green in addition to surprisingly scalable in comparison to other accelerating strategies.



We can observe the comparison results between the energy and power of FPGA-based accelerator and GPU-based accelerators.

V. CONCLUSION

In this paper we carried out Deep Learning Accelerator Unit (DLAU), that is scalable accelerator structure for large-scale deep mastering networks using field-programmable gate array (FPGA) because the hardware prototype. The DLAU accelerator employs 3 pipelined processing devices to improve the throughput and makes use of tile techniques to discover locality for deep getting to know programs.From the experimental results we proved that the proposed DLAU accelerator can provide the high speed performance as well as low power consumption in neural networks.

REFERENCES

[1] Chao Wang, Lei Gong, Qi Yu, Xi Li, Yuan Xie and Xuehai Zhou, "DLAU: A Scalable Deep Learning Accelerator Unit on FPGA", IEEE Transactions On Computer-Aided Design Of Integrated Circuits And Systems, Vol. 36, No. 3, March 2017

[2] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[3] J. Hauswald et al., “DjiNN and Tonic: DNN as a service and its implications for future warehouse scale computers,” in Proc. ISCA, Portland, OR, USA, 2015, pp. 27–40.

[4] C. Zhang et al., “Optimizing FPGA-based accelerator design for deep convolutional neural networks,” in Proc. FPGA, Monterey, CA, USA, 2015, pp. 161–170.

[5] D. L. Ly and P. Chow, “A high-performance FPGA architecture for restricted Boltzmann machines,” in Proc. FPGA, Monterey, CA, USA, 2009, pp. 73–82.

[6] T. Chen et al., “DianNao: A small-footprint high-throughput accelerator for ubiquitous machine-learning,” in Proc. ASPLOS, Salt Lake City, UT, USA, 2014, pp. 269–284.

[7] S. K. Kim, L. C. McAfee, P. L. McMahon, and K. Olukotun, “A highly scalable restricted Boltzmann machine FPGA implementation,” in Proc. FPL, Prague, Czech Republic, 2009, pp. 367–372.

[8] Q. Yu, C. Wang, X. Ma, X. Li, and X. Zhou, “A deep learning prediction process accelerator based FPGA,” in Proc. CCGRID, Shenzhen, China, 2015, pp. 1159–1162.

[9] J. Qiu et al., “Going deeper with embedded FPGA platform for convolutional neural network,” in Proc. FPGA, Monterey, CA, USA, 2016, pp. 26–35.