Energy Efficient Neural Networks for Big Data Analytics

Yu Wang[1], Boxun Li[1], Rong Luo[1], Yiran Chen[2], Ningyi Xu[3], Huazhong Yang[1]
[1]Dept. of E.E., Tsinghua National Laboratory for Information Science and Technology,
Tsinghua University, Beijing, China
[2]Dept. of E.C.E., University of Pittsburgh, Pittsburgh, USA
[3]Microsoft Research Asia, Beijing, China
Email: yu-wang@mail.tsinghua.edu.cn

***Abstract*—The world is experiencing a data revolution to discover knowledge in big data. Large scale neural networks are one of the mainstream tools of big data analytics. Processing big data with large scale neural networks includes two phases: the training phase and the operation phase. Huge computing power is required to support the training phase. And the energy efficiency (power efficiency) is one of the major considerations of the operation phase. We first explore the computing power of GPUs for big data analytics and demonstrate an efficient GPU implementation of the training phase of large scale recurrent neural networks (RNNs). We then introduce a promising ultra-high energy efficient implementation of neural networks' operation phase by taking advantage of the emerging memristor technique. Experiment results show that the proposed GPU implementation of RNNs is able to achieve $2 \sim 11\times$ speed-up compared with the basic CPU implementation. And the scaled-up recurrent neural network trained with GPUs realizes an accuracy of 47% on the Microsoft Research Sentence Completion Challenge, the best result achieved by a single RNN on the same dataset. In addition, the proposed memristor-based implementation of neural networks demonstrates power efficiency of > 400 GFLOPS/W and achieves energy savings of $22\times$ on the HMAX model compared with its pure digital implementation counterpart.**

I. Introduction

The amount of data in our world is exploding at an astounding rate. We have entered the 'Era of Big Data'. Big data contain huge amount of intelligence and absolutely have the potential to change the way governments, organizations, and academic institutions conduct business and make discoveries [1].

Unstructured data are data that do not follow a specified format, including text, speech, image, video, and more. Unstructured data are the major components of big data [2]. As a new, relatively untapped source of intelligence, unstructured data analytics is able to reveal important interrelationships that were previously difficult or impossible to determine [3]. However, the management of unstructured data is a large problem. Unstructured data must be converted to structured data for better analysis and management. In other words, we need to put a label on everything and establish the structure of big data.

Large scale neural networks, also known as the deep neural networks (DNNs) or deep learning, have demonstrated a great promise in processing unstructured data. State-of-the-art performance has been reported in many unstructured data processing tasks, ranging from visual object classification, speech recognition, to nature language processing and information retrieval [4]. Nowadays, The DNN has become one of the most popular tools to process big data [5].

Processing big data with large scale neural networks includes two phases: the training phase and the operation phase. Huge computing power is required to support the training phase. Training phase is the critical operation to obtain a neural network for a specific task. It usually demands a large volume of memory and computing resources, and the time consumption can range from a few seconds to hundreds of hours, depending on the scale of the networks [6]. In addition, recent research has demonstrated that the performance of DNNs, such as the classification accuracy, can be drastically improved by increasing the scale of the network [7]. Therefore, in order to better process big data with large scale neural networks, we need to explore more computing power.

For the operation phase, the energy efficiency (power efficiency) is one of the major concerns. In order to obtain more computing power with less cost, we need to improve the energy efficiency of the computing system. The latest estimation indicated that power efficiency of ~ 100 GFLOPS/W is necessary to support more powerful information processing systems [8]. While, the existing contemporary CPU or GPU systems fall far behind (~ 10 GFLOPS/W) [9], [10]. A revolutionary architecture is demanded to satisfy the continuously growing requirement of energy efficiency for large scale neural networks and big data analytics.

We explore the computing power of GPUs for big data analytics by demonstrating an efficient GPU implementation of the training phase of large scale recurrent neural networks (RNNs). And we introduce a promising hardware solution to the ultra-high energy efficient implementation of the operation phase: the memristor-based neural networks. Finally, we discuss several challenges of the work and show some future directions.

II. Large Scale Recurrent Neural Networks on GPUs

The recurrent neural network (RNN) is a special type of neural network equipped with additional recurrent connections. Those unique recurrent connections enable the RNN to store the processed information and capture the long-range dependencies between input data. Therefore, the RNN has been regarded as an expressive model to process nonlinear sequential processing tasks, such as speech recognition, nature language processing and even video indexing [14]. However, the large computation complexity makes it difficult to effectively train a RNN and therefore significantly limits the research on RNNs in the last 20 years.

In recent years, the use of graphics processing units (GPUs) has been a significant advance to accelerate the training process and scale up neural network models [7]. In our work, we focus on accelerating the training process of RNNs with GPUs and further scaling up the RNN model to improve its performance.

We first explore the potential parallelism of the recurrent neural network and propose a fine-grained two-stage pipeline implementation. The experiment results are illustrated in Table I. The results show that the proposed GPU implementation can achieve $2 \sim 11\times$ speed-up compared with the basic CPU implementation with the Intel Math Kernel Library. We then use the proposed GPU implementation

This work was supported by the Microsoft Research Asia (MSRA), National Natural Science Foundation of China (No. 61373026, No. 61261160501, No. 61028006), 973 project 2013CB329000, National Science and Technology Major Project (2013ZX03003013-003), Youth Talent Development Plan of Beijing (YETP0099), and NSF CAREER CNS-1253424. And we gratefully acknowledge the support of NVIDIA Corporation with the donation of the GPUs used for this research.

978-3-9815370-2-4/DATE14/©2014 EDAA

TABLE I

PERFORMANCE OF THE LARGE SCALE RECURRENT NEURAL NETWORK ON GPUS

(a) Time consumption of different implementations of the RNN[1]

Hidde Size	BPTT[2]=2			BPTT=5			BPTT=10		
	CPU	GPU	Speed-up	CPU	GPU	Speed-up	CPU	GPU	Speed-up
128	96.63	39.25	2.46	150.59	77.31	1.95	264.07	121.81	2.17
256	128.42	49.89	2.57	240.56	95.14	2.53	389.44	143.39	2.72
512	429.09	73.78	5.82	871.02	137.52	6.33	1490.79	199.24	7.48
1024	1238.60	130.45	9.50	2385.18	239.78	9.95	3796.22	338.92	11.20

(b) Performance on the MRSC Challenge[3]

Method	Accuracy
Chance (Baseline)	20
Smoothed 4-gram in [11]	39
RNN-100 with 100 classes[4]	40
Proposed RNN-1000	47
vLBL+NCE5 in [12][5]	60.8

[1] The size of the RNN is $10,000 \times HiddenSize \times 10,000$. The RNN model is the same as the one described in [13]

[2] 'BPTT' represents the step (training depth) of training RNNs with the Backpropagation-Through-Time algorithm.

[3] We use the RNN as a language model and we test the model performance on the Microsoft Research Sentence Completion Challenge [11].

[4] The model is trained with the RNNLM Toolkit in [13]. The network nodes are divided into classes to reduce the computation complex at the cost of performance reduction.

[5] The best result we have seen so far.

to scale up the recurrent neural network and improve its performance. The experiment results of the Microsoft Research Sentence Completion Challenge demonstrate that the large scale recurrent network trained with GPUs is able to beat the traditional modest-size recurrent network and achieve an accuracy of 47%, the best result achieved by a single recurrent neural network on the same dataset.

III. MEMRISTOR-BASED ENERGY EFFICIENT IMPLEMENTATION OF NEURAL NETWORKS

Energy efficiency, or power efficiency, has become one of the most crucial considerations in computing system design. Unfortunately, the cessation of Moore's Law has limited further improvements in energy efficiency. In recent years, the physical realization of the memristor has demonstrated a promising solution to ultra-integrated hardware realization of neural networks, which can be leveraged for better performance and energy efficiency gains.

Memristor was first physically realized by HP Labs in 2008. Afterwards, the memristor attracts significant research interest as it is the first memory technology with enough power efficiency and density to rival biological computation [15]. For example, the memristor crossbar structure provides an incredible execution efficiency of the matrix-vector multiplication, which is one of the most significant operations of artificial neural networks [16]. And as illustrated in Fig. 1, the memristor crossbar-based neural network can be used to realize a low power approximated computation unit cooperating with CPUs, achieving power efficiency of > 400 GFLOPS/W and energy savings of $22\times$ with quality loss of at most 1.87% [17].

IV. FURTHER WORK AND DISCUSSION

Large scale neural networks are efficient tools for big data analytics. We leverage the computing power of GPUs for the training phase. And we introduce an ultra-high energy efficient implementation of the operation phase with the emerging memristor technique. However, many challenges still lie ahead: First, more computing power is demanded to support large scale neural networks and big data analytics. CPUs, GPUs, FPGAs and even mixed-signal systems should cooperate more efficiently to reach this target. Secondly, the training algorithm, especially the most widely used stochastic gradient descent (SGD) algorithm, is a sequential iterative process and hard to get parallelized. How to realize a parallel implementation of the SGD is a major challenge of further accelerating the training process and scaling up the neural networks. Thirdly, the present fabrication technology limits the scale of memristor crossbar arrays to modest size. We need efficient methods of combining many modest-size memristor crossbar arrays to realize large scale memristor-based neural networks. Finally, the memristor only enables an efficient implementation of the operation phase of neural networks, while the training phase is still confined to the traditional digital systems. Although there have been several works trying to realize the self-training of memristor-based neural networks with mixed-signal systems [18], these methods achieve the energy gains at the cost of training quality loss. How to realize better training results with less time and energy consumption is another major problem of processing big data with neural networks.

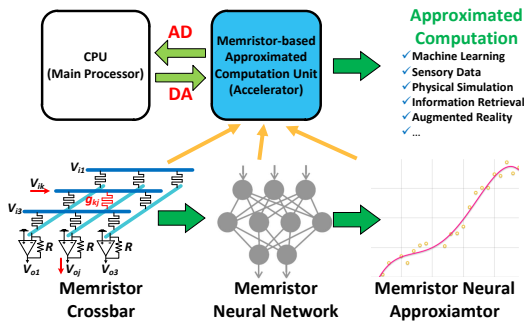

Fig. 1. Proposed Memristor-based Approximated Computation Framework

REFERENCES

[1] Microsoft. The big bang: How the big data explosion is changing the world.

[2] EMC Corporation IDC IView. Extracting value from chaos.

[3] Intel. Big data 101: Unstructured data analytics.

[4] Jeffrey Dean et al. Large scale distributed deep networks. In *Advances in Neural Information Processing Systems 2012*.

[5] Kai Yu. Large-scale deep learning at baidu. In *International Conference on Information and Knowledge Management 2013*.

[6] Jiquan Ngiam et al. On optimization methods for deep learning. In *International Conference on Machine Learning 2011*.

[7] Adam Coates et al. Deep learning with cots hpc systems. In *International Conference on Machine Learning 2013*.

[8] DARPA. Power efficiency revolution for embedded computing technologies.

[9] NVIDIA TESLA K-SERIES DATASHEET. Kepler family product overview, 2012.

[10] Intel. Intel microprocessor export compliance metrics, 2013.

[11] Geoffrey Zweig et al. A challenge set for advancing language modeling. In *NAACL-HLT 2012 Workshop*.

[12] Andriy Mnih et al. Learning word embeddings efficiently with noise-contrastive estimation. In *Advances in Neural Information Processing Systems 2013*.

[13] Tomas Mikolov et al. Rnnlm-recurrent neural network language modeling toolkit. In *2011 IEEE workshop on Automatic Speech Recognition and Understanding*.

[14] Ilya Sutskever et al. Generating text with recurrent neural networks. In *International Conference on Machine Learning 2011*.

[15] Massimiliano Versace and Ben Chandler. Moneta: a mind made from memristors. *IEEE Spectrum*, 2010.

[16] Miao Hu, Hai Li, Qing Wu, and G.S. Rose. Hardware realization of bsb recall function using memristor crossbar arrays. In *DAC 2012*.

[17] Boxun Li, Yi Shan, Miao Hu, Yu Wang, Yiran Chen, and Huazhong Yang. Memristor-based approximated computation. In *ISLPED 2013*.

[18] Boxun Li, Yuzhi Wang, Yu Wang, Yiran Chen, and Huazhong Yang. Training itself: Mixed-signal training acceleration for memristor-based neural network. In *ASPDAC 2014*.