Journal of Signal Processing Systems (2020) 92:1215–1217
https://doi.org/10.1007/s11265-020-01576-5

Guest Editorial: Special Issue on Architectures and Design Methods for Neural Networks

Ahmed Hemani [1] · Mohammed Shafique [2] · Kolin Paul [3]

Received: 29 June 2020 / Revised: 29 June 2020 / Accepted: 29 June 2020 / Published online: 8 July 2020

The key strength and differentiation of humans as species has been their ability to harness external sources to supplement their physical and mental faculties. For a greater part of history, humans have harnessed combination of tools, natural forces like gravity and wind, and animal muscles to supplement their own weak muscles. A big breakthrough in supplementing weak biological muscles came about when we learned to mechanize generation of power and achieved affordable 2–3 orders greater power density compared to biological power. This transformed the society from being primarily agricultural to industrial. Parallel to the history of our attempts to supplement our physical strength has been a trend to supplement our mental strength. Development of writing, paper, printing, storage and transmission of information and computation have been some of the key steps in supplementing our mental strength with external tools and machines.

In spite of this long and impressive history of progress in supplementing human abilities with external forces and tools, one aspect of humans has been considered as quintessentially human that cannot be mechanized or automated. This aspect is our ability to reason and create ideas – also known as intelligence. Human intelligence has been the key driver in our ability to supplement our physical and mental abilities. The question that has been discussed for more than a century is can human intelligence create artificial intelligence to supplement it and perhaps exceed that for certain tasks? The jury is still out on this question and is likely to remain so for some time. At this stage, we do not have an accepted definition of what constitutes *intelligence*, or its crux in form of the *cognition*. Our earlier attempts at Artificial Intelligence (AI) based on what is known as symbolic computation did not succeed well. This is how Marvin Minsky summed up the status of symbolic AI in 1977:

> *Our first foray into Artificial Intelligence was a program that did a credible job of solving problems in college calculus. Armed with that success, we tackled high school algebra; we found, to our surprise, that it was much harder. Attempts at grade school arithmetic, involving the concept of numbers, etc., provide problems of current research interest. An exploration of the child's world of blocks proved insurmountable, except under the most rigidly constrained circumstances. It finally dawned on us that the overwhelming majority of what we call intelligence is developed by the end of the first year of life.*

Another noted computer scientist Prof. Frederick P. Brooks Jr. deemed that that task of inventing algorithms as impossible. He made his case in an IEEE Computer article in April 1987. The title of his article was "No Silver Bullet. Essence and accidents of Software Engineering". Fred Brooks argued that all progress in computer science can only help solve the accidental problems like an array index out of bounds. The essential problem is the invention of the algorithms and for this, Fred decisively asserts, there can be no automation, whence the phrase "no silver bullet" in the title.

Humbled by the failure of Symbolic AI and warned by Fred Brooks that there are no solutions to automating the invention of algorithms, the AI research community has adopted the connectionist approach that to varying degrees is inspired by how a biological brain is organized and for this reason these structures are also known as artificial neural

✉ Ahmed Hemani
hemani@kth.se

Mohammed Shafique
muhammad.shafique@tuwien.ac.at

Kolin Paul
kolin@cse.iitd.ac.at.in

[1] Department of EE, School of EECS, KTH (Royal Institute of Technology), Stockholm, Sweden

[2] Institute of Computer Engineering, Vienna University of Technology, Vienna, Austria

[3] Department of CSE, Indian Institute of Technology, Delhi, India

networks or ANNs. Research in ANN has become a megatrend, especially in the recent times, due to the availability of large amount of training data as well as immense computing power to process that. The principal reason for this enthusiasm is that in a weak sense ANNs have proven Fred Brooks wrong. If, we have a sufficiently large ANN, training data and time, it is possible to approximate the behavior of any algorithm. In other words, as long as we know what outputs to expect for given inputs, we can train the ANN to approximate the outcome of an algorithm without having to invent it – the essential problem. That it cannot exactly replicate an algorithm is the reason why an ANN weakly falsifies Fred Brook's claims. John von Neumann predicted long back that if there is an algorithm, one can create a neural network to replace it. Given this, can we claim that what required human intelligence to invent an algorithm to transform a given set of inputs to a desired set of outputs has been replaced by the mechanical work of training an ANN? This is a tricky question, after all ANNs are also algorithms. So far, ANNs are being invented by humans.

ANNs superficially mimic how brains work. The learning algorithms that they adopt are very inefficient and nowhere close to what biological brain is capable of. To many neuroscientists, adopting Spiking Neural Networks (SNNs), where neurons more closely model the biological neurons is the way forward. However, due to the lack of efficient learning models, these SNNs still lie behind the Deep Neural Networks (DNNs) in certain tasks. However, since SNNs are more biological plausible, they bear a great potential to bridge the gap between human intelligence and AI.

Bulk of the applications of AI have focused on vision applications using feedforward DNNs. However, another challenging class of applications rely on sequence learning like natural language processing and are better served by Recurrent Neural Networks (RNNs). Implementations of these networks have received relatively less attention. Moreover, the concept of *Transformers,* which essentially is a neural network based on a self-attention mechanism has recently been enhanced through Universal Transformers (UTs) that provide a parallel-in-time self-attentive recurrent sequence model that promises much higher efficiency compared to RNNs for language understanding. In this special issue, the first two papers address the architectural challenges in implementing RNNs.

Next, we briefly summarize the seven papers included in this special issue:

1. While most of the work on ANNs/DNNs has focused on the convolutional neural networks (CNNs) that are feedforward type, RNNs like Long Short-Term Memory (LSTM) have attracted relatively less attention. The research from TU Kaiserslautern is noteworthy not just, as a research that restores balance and increases diversity, but it also reports key innovations and strong results. The paper shows proper understanding of LSTM data flow and matching it with a custom micro-architecture crafted on an FPGA that can significantly outperform very capable GPUs like K80. Secondly, this paper presents for the first time a hardware implementation of multi-dimensional LSTM. Finally, this paper also presents an innovative idea of doing Processing-In-Memory (PIM) for 1D-LSTM, the memory in this case is a classical commodity DRAM device. This PIM implementation of 1D-LSTM is 10X more efficient compared to FPGA implementations and the logic needed to implement LSTM as PIM adds an overhead of just 18% compared to unmodified commodity DRAMs.
2. Natural language processing often calls for long term dependencies on history. To address this, a new class of network called Memory Augmented Neural Network or MANN has emerged. MANNs are conventional RNNs with external memory for attention mechanism. The paper from Penn State address architectural challenges in accelerating MANN. Unlike conventional ANNs like LSTM and CNN that are dominated by MAC operations, MANN has a greater diversity. Besides MACs, MANN also has operations for similarity measure, sorting, weighted memory access, pair-wise arithmetic etc. For this reason, MANNs do not lend so easily to acceleration using the rich set of architectural solutions that have been developed for the MAC dominated ANNs. This paper proposes end-to-end acceleration solutions and applies them to variants of MANNs like Neural Turing Machine, Differential Neural Computer and Meta-learning model. These acceleration techniques offer 1–2 orders improvement over CPUs and GPUs. To further optimize the designs, this paper also proposes using PIM for MACs and similarity metric operations to minimize data movement.
3. In narrow domains like image recognition and classification, CNNs are already outperforming humans. GANs have beaten humans in the challenging game of GO. The demands for what these networks can do is increasing exponentially and with it are the size of the network. Simultaneously, there is ever greater need to deploy these increasingly capable and complex ANNs in field in power constrained products. This has spawned an entire sub-discipline to ANN called approximate computing. In this paper, the authors exploit the well-known method to transform weights into log domain to reduce the number of bits required and the area and energy required for computation. However, this simplistic policy would result in unacceptably high loss of accuracy. The trick is to use base 2 for high values and base sqrt (2) for low values. This in principle solves the problem but there are additional challenges, the weight distribution in each layer is different. Ideally, the range that should be compressed

with base 2 and the range that should be compressed sqrt (2) should be adjusted according to the weight distribution of each layer. To solve this conundrum, the authors propose, implement and validate the benefits of a Base Reconfigurable Log Unit.

4. Analog VLSI is in many respects the most natural and elegant way to implement and emulate neurons and structure composed in terms of neurons. Carver Mead wrote the classic book Analog VLSI and Neural Systems and inspired many scientists who are active still today and training the next generation scientists that carry on this craft. In the paper from the Heidelberg group they present the second generation of the well-known BrainScaleS architecture which extends these ideas and complements an accelerated analog VLSI model with high-speed digital communication and control logic. While this new design has significant new features in terms of neuron and synapse models, configurability, and digitally assisted learning rules, the main contribution of this work is that it presents a methodology to design and verify such complex mixed-signal designs. Several semi-custom methods for integrating large analog macro blocks with wide high-speed digital interfaces into a digital top-level design are highlighted. In the verification context, the use of Monte Carlo simulation to tune parameters for hypothetical silicon design instances with their manufacturing related variations as a way to explore the design space to tune parameters and pre-optimize designs before tapeout is noteworthy.
5. Online learning that involves not just synaptic plasticity, but also structural plasticity poses a challenge. This challenge stems from the fact that structural plasticity can change the structure of the SNN, some old synapses might be broken, while some new ones created. Consequently, the optimality of mapping an SNN to a fixed neuromorphic architecture is a moving target and cannot be decided statically at design time. Dynamically remapping with changing structural plasticity is desirable but such remapping has to be agile enough to find a new mapping before the structural plasticity changes substantially again making the remapping decision stale. The paper from IMEC and Drexel addresses this problem. It presents an innovative runtime methodology that is almost three orders faster than a design static mapping methodology making it practical to dynamically remap SNN in response to changing structural plasticity. The cost of agility in terms of optimality of mapping is acceptably low at less than 7% compared to a static mapping algorithm.
6. This paper is unique in being a hybrid of network of neuronal-cells in a culture connected to a synthetic mixed-signal neuromorphic design. For such a hybrid construction to work, the electronic neuromorphic system's behavior, especially its bursting characteristics must be predictable and finely tunable. To achieve this the researchers in Dresden have adopted a switch-capacitance based design to avoid the perils variations due to manufacturing in conventional analog circuits. Moreover, they use a mean-field approach to tune the parameters to match the behavior of the biological neural network in culture.
7. The paper from KTH, Sweden, presents a detailed ASIC implementation of Bayesian Confidence Propagation Neural Network (BCPNN). BCPNN is a biologically plausible model of cortex with a Bayesian learning rule as an alternative to the spike timing dependent plasticity. The eBrain-2 paper details a tile architecture that exploits custom 3D integrated DRAM and custom data organization in it to balance the access efficiency of row wise and column wise access of the synaptic storage. The paper presents a detailed methodology on how the multi-dimensional space in terms of storage, interconnect and computation have been systematically explored to decide the dimensions of the architecture. This eBrain-2 is benchmarked against GPU and SpiNNaker and shows three orders of magnitude improvement in energy-delay cost.

In conclusion, we are now in an era where Neural Networks and Neuromorphic systems that hold great promise for many emerging engineering and scientific applications. These systems are essentially based on non-conventional architectures that mimic neuro-biological architectures processes, The goal of this special issue is to provide a selection of high quality papers for researchers and practitioners from academia, government and industry with interests in the area of neuromorphic computing.