Institut für Theoretische Elektrotechnik und Mikroelektronik

Universität Bremen

PhD proposal

Research project:
System-on-Chip architectures for real-time machine-learning algorithms in industrial Internet-of-Things applications

Candidate name:
Yarib Israel Nevárez Esparza

March 19, 2021

1 General information

1.1 Personal information

First name: Yarib Israel
Last name: Nevárez Esparza
Date of birth: May 26, 1986
Phone: +49 (1) 788 936794
E-mail: nevarez@item.uni-bremen.de

Yarib Israel Nevárez Esparza

1.2 Supervisor

Title: Prof. Dr.-Ing.
First name: Alberto
Last name: García-Ortiz
Department: Fachbereich 1 - Physik / Elektrotechnik, Institut für theoretische Elektrotechnik und Mikroelektronik
Office address: Building NW 1, Otto-Hahn Allee 1, 28359 Bremen
Phone: +49 (0) 421 218 62533
E-mail: agarcia@item.uni-bremen.de

Prof. Dr.-Ing. Alberto García-Ortiz

1.3 Specialization and working direction

Specialization: Electrical and computer engineering
Working direction: Integrated digital systems

1.4 Expected total duration

The planned duration is 3 years. Start date: May 27, 2019. End date: May 27, 2022.

Abstract

In the emerging era of Industry 4.0, Machine Learning (ML) algorithms yield the power of Artificial Intelligence (AI) to the ubiquitous Internet of Things (IoT) devices. Applications in this field become smarter and more profitable as the availability of big data increases, driving the evolution of many aspects of daily life, science, and industry. However, state-of-the-art ML algorithms, specially Spiking Neural Networks (SNNs) and deep Convolutional Neural Networks (CNNs), are highly compute and data intensive. Therefore, the substantial demand for power and hardware resources of these algorithms represents a restriction for IoT devices in the scope of embedded systems.

Energy, performance, and resource utilization are the key design concerns in computer systems. Considering the intrinsic error resilience of ML algorithms, paradigms such as approximate computing come to the rescue by offering promising efficiency gains to overcome the aforementioned concerns. Approximation techniques are widely used in ML algorithms at the model-structure as well as at the hardware processing level. However, state-of-the-art approximate computing methodologies do not sufficiently address accelerator designs for Deep Neural Networks (DNN) as power- and resource-demanding algorithms.

To sustain the continuous expansion of ML applications on resource-constrained devices, approximate computing will gradually transform from a design alternative to an essential prerequisite. This PhD proposal focuses on the investigation of approximate computing techniques to exploit the intrinsic error resilience of ML algorithms to optimize computing embedded systems. The goal of this research is to contribute to state-of-the-art knowledge with formal methodologies to address hardware design for neural network accelerators based on approximate computing.

Furthermore, the expected outcome of this PhD is to develop high-efficiency neural network accelerator architectures with a common design methodology: (1) SNN accelerator for fundamental research; and (2) deep CNN accelerator for industrial computer vision applications (e.g., real-time multiple object detection and classification). Finally, the motivation of this work is to support the growing demand of processing capabilities of ML algorithms in the scope of embedded systems, and to contribute to the rise of a sustainable power-efficient next generation of neural network accelerators based on approximate computing.

2 Introduction

Industry is the piece of an economy that produces material goods which are highly mechanized and automatized. Since the beginning of industrialization, technological leaps have led to paradigm shifts that are now called "industrial revolutions": from mechanization, electrification, and later, digitalization (the so-called 3rd industrial revolution). Based on the advanced digitalization within factories, the combination of Internet technologies and future-oriented technologies in the field of "smart" things

(machines and products) seems to result in a new fundamental paradigm shift in industrial production. Emerging from this future expectation, the term "Industry 4.0" was established for an expected "4th industrial revolution" [1].

To build the emerging environment of Industry 4.0, disruptive technologies are required to handle autonomous communications between all industrial devices throughout the factory and the Internet. Such technologies offer the potential to transform the industry along the entire production chain and stimulate productivity and overall economic growth [2]. These technologies include cloud computing, big data, and specially a new generation of IoT devices fused with Cyber-Physical systems (CPS), augmented reality, ML analytics, and Artificial Intelligence (AI) in general [3].

2.1 ML Algorithms in Embedded Systems

The continuous evolution of ML/AI algorithms and IoT devices has not only made various ML/AI applications the major workload running on these embedded devices, but ML/AI has become the main approach for industrial solutions, especially in the rise of Industry 4.0 [3]. In fact, there is a clear motivation to run ML/AI algorithms on IoT devices because of [4]: (1) feasibility of mission-critical real-time processing and inference; (2) privacy and security of data; (3) offline operation capability; and (4) stressed communication robustness. Hence, the traditional term of IoT has also been redefined as AI of Things (AIoT) to emphasize the impact of ML/AI on this technology [5].

2.2 Constraints

The problem lies in the fact that state-of-the-art ML/AI algorithms, particularly DNNs, are highly compute and data intensive. This represents significant computational challenges across the spectrum of computing hardware, specially in the scope of embedded systems [6]. One of the most deployed applications is computer vision using CNNs. Compared to the conventional image processing approaches, the CNN accuracy has improved significantly that by 2015, a human can no longer beat a computer in image classification [4]. The early development of CNNs before 2016 mainly focused on accuracy improvement without considering computational costs. While accuracy of deep CNN for image classification improved 24% between 2012 and 2016, the demand on hardware resources increased more than 10×. Starting from 2017, significant attention was paid to improve hardware efficiency in terms of compute power, memory bandwidth, and power consumption, while maintaining accuracy at a similar level to human perception [6]. Nonetheless, the state-of-the-art of CNN-based algorithms, such as multiple object detection models (e.g., SPP-net [7], SSD [8], Faster R-CNN [9], and YOLOv4 [10]), are still unsuitable for the resource-limited nature of embedded systems [11, 12].

Consequently, the recent breakthroughs in ML applications have brought significant

advancements in neural network processors [13]. These rapid evolution, however, came at the cost of an important demand for computational power. Hence, to bring the inference speed to an acceptable level, custom ASIC Neural Processing Units (NPUs) are becoming ubiquitous in both embedded and general purpose computing. NPUs perform several tera operations per second in a confined area. Therefore, they become subject to elevated on-chip power densities that rapidly result in excessive on-chip temperatures during operation [14]. These design efforts focused on power-hungry parallel computing techniques, yet unsustainable for resource-constrained devices. As a result, radical changes to conventional computing approaches are required in order to sustain and improve performance while satisfying mandatory energy and temperature constraints [15].

2.3 Alternatives

To overcome the problem, based on the error-resilience of ML algorithms, an evident solution is approximate computing. This computing paradigm has been used in a wide range of applications to increase the hardware computational efficiency[16]. For neural network applications, two main approximation strategies are used, namely network compression and classical approximate computing[17].

2.3.1 Network Compression and Quantization

Researchers focusing on embedded applications started lowering the precision of weights and activation maps to shrink the memory footprint of the large number of parameters representing DNNs, a method known as network quantization. In this manner, reduced bit precision causes a small accuracy loss [18, 19, 20, 21]. In addition to quantization, network pruning reduces the model size by removing structural portions of the parameters and its associated computations [22, 23]. This method has been identified as an effective technique to improve the efficiency of CNN for applications with limited computational budget[24, 25, 26]. These techniques leverage the intrinsic error-tolerance of neural networks, as well as their ability to recover from accuracy degradation while training.

2.3.2 Approximate Computing

Approximate computing introduces quality loss as a new design metric to be traded off for energy, performance, and/or resource utilization. Data redundancy of neural networks incorporate a certain degree of robustness against random external and internal perturbations, for instance, processing quality loss. This property can be exploited in a cross-layer resilience approach [27]: by leveraging error-resilience at algorithmic-level, it can be allowed a certain degree of inaccuracies at the computing-level. This approach consists of designing processing elements that approximate their computation by employing cleverly modified algorithmic logic units [16].

Approximate computing techniques allow substantial enhancement in processing efficiency with moderated accuracy degradation. Some research papers have shown the feasibility of applying approximate computing to the inference stage of neural networks [28, 16, 29, 30, 31, 32]. Such techniques usually demonstrated small inference accuracy degradation, but significant enhancement in computational performance, resource utilization, and energy consumption. Hence, by taking advantage of the intrinsic error-tolerance of neural networks, approximate computing is positioned as a promising approach for inference on resource-limited devices. Nonetheless, the complex state-of-the-art of CNN-based algorithms has not been sufficiently addressed with approximate computing techniques.

2.4 Problem Statement

While the state-of-the-art approximate computing techniques have presented highly-efficient adders and multipliers, they do not sufficiently address accelerator designs for ML algorithms, specifically for the new generation of neural networks.

2.5 Research Objective

Considering the broad range of ML algorithms, te research objective for this PhD proposal is the following: *Investigating formal design methodologies for high-efficiency neural network hardware accelerator based on approximate computing in the scope of embedded systems.*

2.5.1 Research questions

- ◇ How to analyze neural networks for error resilience?
- ◇ How to exploit intrinsic error resilience of neural networks effectively?
- ◇ How to design neural network accelerators based on approximate computing?
- ◇ Considering the case study of SNNs applied for fundamental research, how do the proposed approximate computing methodology affects the quality and efficiency of the processing?
- ◇ Considering the case study of CNNs applied for industrial computer vision, how do the proposed approximate computing methodology affects the quality and efficiency of the processing?
- ◇ What are the possibilities and challenges to embrace approximate computing for neural network accelerators?

2.6 Motivations

2.6.1 Fundamental

1. Derive formal methodologies to address hardware design for neural network accelerators based on approximate computing

2. Promote the next generation of neural network accelerators based on approximate computing

2.6.2 Practical

1. Support the growing demand of processing capabilities of ML algorithms in the scope of embedded systems

2. Contribute with System-on-Chip architectures to extend the use of AIoT for scenarios not possible today for real-time machine learning algorithms (e.g. industrial computer vision, signal recognition, feature filtering, machine translation, material inspection, etc.)

To sustain the continuous expansion of ML applications on resource-constrained devices, approximate computing will gradually transform from a design alternative to an essential prerequisite. This PhD proposal focuses on the investigation of approximate computing techniques to exploit the intrinsic error resilience of ML algorithms to optimize computing embedded systems at the hardware architecture and circuit-level to achieve efficiency gains. The goal of this research is to contribute to state-of-the-art knowledge of this domain with formal methodologies to address accelerator designs for SNNs and deep CNNs. Furthermore, the expected outcome of this work is to develop high-efficiency accelerator architectures for SNN and deep CNN algorithms for computer vision applications (e.g., real-time multiple object detection and classification). Finally, the motivation of this work is to support the growing demand of processing capabilities of ML algorithms in the scope of embedded systems and to contribute to the rise of a sustainable power-efficient next generation of neural network accelerators based on approximate computing.

SNNs offer advantageous robustness and the potential to achieve a power efficiency closer to that of the human brain. SNNs emulate the real behavior of neurons in different levels of detail. The more detailed the biological part is emulated, the greater the computational complexity [36, 37]. Most of today's SNNs use a very detailed model. In contrast, Spike-By-Spike (SbS) neural networks are on the less realistic side of the biological realism scale [38, 34]. In spite of that, SbS still uses stochastic spikes as a means of transmitting information between populations of neurons, and thus retains the robustness advantages of SNNs. Correspondingly, the hardware complexity of the approach is greatly reduced [33, 39].

Moreover, since approximations and noise have qualitatively the same effect[40], we apply noise tolerance plots as an intuitive visual measure to provide insights into the quality degradation of SbS networks under approximate processing effects.

Our main contributions are as follows:

Driven by this high potential for power reduction, designing approximate circuits has attracted significant research interest. At the custom hardware level, approximate computing targets mainly arithmetic units [41, 42, 43, 44] (e.g., adders and multipliers) since they form the core components of all computations and a vast number of error-tolerant applications. Specifically, in NN inference the majority of the energy is consumed in the multiplication operations. Recent research showed that employing approximate multipliers in NN inference can deliver significant energy savings for a minimal loss in accuracy [44, 45, 46]. However, designing approximate circuits under quality constraints heavily increases the design time cycle since the designer has to verify both functionality and optimality as well as operating within error bounds [47]. This task becomes even more challenging as the circuits complexity increases. To this end, several research activities, such as approximate high-level synthesis (AHLS) [48], focus on automating the generation of approximate circuits. Approximate HLS estimates error propagations and distributes the available error budget to the different approximate sub-components of a larger accelerator, such as convolution operators and generic matrix multiply units. As a result, AHLS enables generating complex approximate micro-architectures that satisfy given quality requirements.

Moreover, approximate computing is further subdivided into static and dynamically reconfigurable approximation techniques. The latter, leveraging that error-tolerance and the induced errors are context- and input-dependent, aim to improve accuracy by providing a fine grain quality control and/or to further boost (power, energy, and/or delay) gains by applying more aggressive approximation on less-sensitive inputs. Finally, reconfigurable approximation was also recently applied to address thermal constraints [14]. Instead of addressing thermal emergencies by reducing performance, by reducing the accuracy and hence dynamic power in the same area, the circuits power density decreases, resulting in lower temperatures.

In this paper, we study state-of-the art approaches in each of the aforementioned categories and analyze their application in machine learning and neural network domains. In Section 2, we first evaluate approximate multipliers [44] at the component level and in Section 3, we then focus on AHLS [48] approaches targeting approximate design automation at the complete processor or accelerator level. Section 4 further examines neural network specific runtime reconfigurable approximation techniques that target energy and/or temperature optimization. Finally, in Section 5, we discuss the challenges, limitations, and open issues of approximate computing applications in the machine learning domain.

3 State-of-the-art

3.1 Dedicated hardware architectures on embedded FPGA

3.1.1 SbS networks for image classification

3.1.2 CNN-based object detection algorithms

4 Research goals

The goal of this research is to contribute to state-of-the-art knowledge with formal methodologies to address hardware design for neural network accelerators based on approximate computing.

4.1 Outcomes

The expected outcome of this PhD is to derive formal methodologies for neural network accelerators based on approximate computing. As a demonstration, it is expected to develop two high-efficiency neural network accelerator architectures:

- ◇ SNN accelerator applied for fundamental research
- ◇ Deep CNN accelerator for industrial computer vision applications (e.g., real-time multiple object detection and classification)

5 Project plan

The research project will be divided into three phases, the prospective milestones schedule is shown in **Tab.** 1. The total timeframe is expected to be 3 years.

5.1 Phase 1

In-depth research of existing work. Existing findings should be incorporated into the own work effectively. For this purpose, a trial and evaluation of existing analysis and design tools. The goal of Phase 1 is to develop first FPGA-based prototypes compatible with the requirements of Industry 4.0 devices.

Outcome: First, it will result in a library of IP-blocks and hardware architectures consisting on prior state of the art. Second, it will result in a quantitative evaluation of existing techniques as well as the identification of the most relevant bottlenecks in previous approaches. The library of IP-block is expected to be open-sourced and the quantitative analysis reported in a journal paper.

5.2 Phase 2

Selection of suitable design tools as well as a strategy for the further improvements in performance and real-time characteristics. In particular stochastic and approximation techniques in hardware acceleration will be considered. Investigations on feature extraction, ML approximations, and acceleration approaches in configurable logic and hardware. Development of a second prototype based on the automated design flow. The second prototype has a higher computing capabilities.

Outcome: Improved FPGA prototype that demonstrates in real Industry 4.0 applications the advantages of the proposed architecture.

5.3 Phase 3

Selection of final approach, strategies, and hardware architectures for the final project phase. Development of an optimized System-on-Chip architecture with resource constrained embedded devices that efficiently execute real-time machine leaning algorithms in IIoT applications. The research will be documented and published in written form.

Outcome: Development of a SoC in a nanometric technology using the architecture proposed in phase 2.

Milestone	Date	Description
M1	September, 2019	Completion of literature search
M2	January, 2020	Understand the key design considerations for efficient ML processing; understand trade-offs between various hardware architectures and platforms; learn about micro-architectural knobs such as precision, data reuse, and parallelism to architect ML accelerators given target area-power-performance metrics; evaluate the utility of various ML dataflow techniques for efficient processing; and understand future trends and opportunities from ML algorithms on Industry 4.0
M3	April, 2020	Outcome: Development of a library of IP-blocks and hardware architectures consisting of prior state of the art
M4	July, 2020	Outcome: Quantitative evaluation of existing techniques as well as the identification of the most relevant bottlenecks in previous approaches, report in a journal paper
M5	October, 2020	Selection of tools, strategies, techniques for further improvements in performance and real-time characteristics. Stochastic and approximation techniques in hardware acceleration

M6	January, 2021	Investigations on feature extraction, ML approximations, and acceleration approaches in configurable logic and hardware
M7	April, 2021	Outcome: Improved FPGA prototype that demonstrates in real Industry 4.0 applications the advantages of the proposed architecture
M8	July, 2021	Development of SoC architecture efficiently executing real-time machine leaning algorithms in IIoT is completed
M9	October, 2021	Development of a SoC in a nanometric technology is completed
M10	Jun, 2022	Written elaboration is completed

Table 1: Milestone schedule.

References

[1] H. Lasi, P. Fettke, H.-G. Kemper, T. Feld, and M. Hoffmann, "Industry 4.0," *Business & information systems engineering*, vol. 6, no. 4, pp. 239–242, 2014.

[2] H. Espinoza, G. Kling, F. McGroarty, M. O'Mahony, and X. Ziouvelou, "Estimating the impact of the internet of things on productivity in europe," *Heliyon*, vol. 6, no. 5, p. e03935, 2020.

[3] V. Alcácer and V. Cruz-Machado, "Scanning the industry 4.0: A literature review on technologies for manufacturing systems," *Engineering science and technology, an international journal*, vol. 22, no. 3, pp. 899–919, 2019.

[4] K.-H. L. Loh, "1.2 fertilizing aiot from roots to leaves," in *2020 IEEE International Solid-State Circuits Conference-(ISSCC)*. IEEE, 2020, pp. 15–21.

[5] J. Zhang and D. Tao, "Empowering things with intelligence: A survey of the progress, challenges, and opportunities in artificial intelligence of things," *IEEE Internet of Things Journal*, 2020.

[6] S. Venkataramani, K. Roy, and A. Raghunathan, "Efficient embedded learning for iot devices," in *2016 21st Asia and South Pacific Design Automation Conference (ASP-DAC)*. IEEE, 2016, pp. 308–311.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[8] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[9] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[10] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[11] I. Ahmad, S. Shahabuddin, T. Kumar, E. Harjula, M. Meisel, M. Juntti, T. Sauter, and M. Ylianttila, "Challenges of ai in wireless networks for iot," *arXiv preprint arXiv:2007.04705*, 2020.

[12] F. Al-Turjman, *Artificial intelligence in IoT*. Springer, 2019.

[13] N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers *et al.*, "In-datacenter performance analysis of a tensor processing unit," in *Proceedings of the 44th annual international symposium on computer architecture*, 2017, pp. 1–12.

[14] H. Amrouch, G. Zervakis, S. Salamin, H. Kattan, I. Anagnostopoulos, and J. Henkel, "Npu thermal management," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 39, no. 11, pp. 3842–3855, 2020.

[15] S. G. A. Gillani, "Exploiting error resilience for hardware efficiency: targeting iterative and accumulation based algorithms," 2020.

[16] J. Han and M. Orshansky, "Approximate computing: An emerging paradigm for energy-efficient design," in *2013 18th IEEE European Test Symposium (ETS)*. IEEE, 2013, pp. 1–6.

[17] M. Bouvier, A. Valentian, T. Mesquida, F. Rummens, M. Reyboz, E. Vianello, and E. Beigne, "Spiking neural networks hardware implementations and challenges: A survey," *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, vol. 15, no. 2, pp. 1–35, 2019.

[18] M. Courbariaux, Y. Bengio, and J.-P. David, "Binaryconnect: Training deep neural networks with binary weights during propagations," in *Advances in neural information processing systems*, 2015, pp. 3123–3131.

[19] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," *arXiv preprint arXiv:1510.00149*, 2015.

[20] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, and Y. Bengio, "Quantized neural networks: Training neural networks with low precision weights and activations," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6869–6898, 2017.

[21] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, "Xnor-net: Imagenet classification using binary convolutional neural networks," in *European conference on computer vision*. Springer, 2016, pp. 525–542.

[22] Y. LeCun, J. Denker, and S. Solla, "Optimal brain damage," *Advances in neural information processing systems*, vol. 2, pp. 598–605, 1989.

[23] B. Hassibi and D. Stork, "Second order derivatives for network pruning: Optimal brain surgeon," *Advances in neural information processing systems*, vol. 5, pp. 164–171, 1992.

[24] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz, "Pruning convolutional neural networks for resource efficient inference," *arXiv preprint arXiv:1611.06440*, 2016.

[25] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf, "Pruning filters for efficient convnets," *arXiv preprint arXiv:1608.08710*, 2016.

[26] Z. Liu, M. Sun, T. Zhou, G. Huang, and T. Darrell, "Rethinking the value of network pruning," *arXiv preprint arXiv:1810.05270*, 2018.

[27] N. P. Carter, H. Naeimi, and D. S. Gardner, "Design techniques for cross-layer resilience," in *2010 Design, Automation & Test in Europe Conference & Exhibition (DATE 2010)*. IEEE, 2010, pp. 1023–1028.

[28] U. Lotrič and P. Bulić, "Applicability of approximate multipliers in hardware neural networks," *Neurocomputing*, vol. 96, pp. 57–65, 2012.

[29] Z. Du, K. Palem, A. Lingamneni, O. Temam, Y. Chen, and C. Wu, "Leveraging the error resilience of machine-learning applications for designing highly energy efficient accelerators," in *2014 19th Asia and South Pacific design automation conference (ASP-DAC)*. IEEE, 2014, pp. 201–206.

[30] V. Mrazek, S. S. Sarwar, L. Sekanina, Z. Vasicek, and K. Roy, "Design of power-efficient approximate multipliers for approximate artificial neural networks," in *Proceedings of the 35th International Conference on Computer-Aided Design*, 2016, pp. 1–7.

[31] S. S. Sarwar, S. Venkataramani, A. Raghunathan, and K. Roy, "Multiplier-less artificial neurons exploiting error resiliency for energy-efficient neural computing," in *2016 Design, Automation & Test in Europe Conference & Exhibition (DATE)*. IEEE, 2016, pp. 145–150.

[32] G. Zervakis, H. Saadat, H. Amrouch, A. Gerstlauer, S. Parameswaran, and J. Henkel, "Approximate computing for ml: State-of-the-art, challenges and visions," in *Proceedings of the 26th Asia and South Pacific Design Automation Conference*, 2021, pp. 189–196.

[33] Y. Nevarez, A. Garcia-Ortiz, D. Rotermund, and K. R. Pawelzik, "Accelerator framework of spike-by-spike neural networks for inference and incremental learning in embedded systems," in *2020 9th International Conference on Modern Circuits and Systems Technologies (MOCAST)*. IEEE, 2020, pp. 1–5.

[34] U. Ernst, D. Rotermund, and K. Pawelzik, "Efficient computation based on stochastic spikes," *Neural computation*, vol. 19, no. 5, pp. 1313–1343, 2007.

[35] D. Rotermund and K. R. Pawelzik, "Biologically plausible learning in a deep recurrent spiking network," *bioRxiv*, 2019.

[36] E. M. Izhikevich, "Which model to use for cortical spiking neurons?" *IEEE transactions on neural networks*, vol. 15, no. 5, pp. 1063–1070, 2004.

[37] K. Amunts, A. C. Knoll, T. Lippert, C. M. Pennartz, P. Ryvlin, A. Destexhe, V. K. Jirsa, E. DAngelo, and J. G. Bjaalie, "The human brain project – synergy between neuroscience, computing, informatics, and brain-inspired technologies," *PLoS biology*, vol. 17, no. 7, p. e3000344, 2019.

[38] D. Rotermund and K. R. Pawelzik, "Back-propagation learning in deep spike-by-spike networks," *Frontiers in Computational Neuroscience*, vol. 13, p. 55, 2019.

[39] ——, "Massively parallel FPGA hardware for spike-by-spike networks," *bioRxiv*, 2019.

[40] S. Venkataramani, S. T. Chakradhar, K. Roy, and A. Raghunathan, "Approximate computing and the quest for computing efficiency," in *2015 52nd ACM/EDAC/IEEE Design Automation Conference (DAC)*. IEEE, 2015, pp. 1–6.

[41] J. Miao, K. He, A. Gerstlauer, and M. Orshansky, "Modeling and synthesis of quality-energy optimal approximate adders," in *Proceedings of the International Conference on Computer-Aided Design*, 2012, pp. 728–735.

[42] M. Shafique, W. Ahmad, R. Hafiz, and J. Henkel, "A low latency generic accuracy configurable adder," in *2015 52nd ACM/EDAC/IEEE Design Automation Conference (DAC)*. IEEE, 2015, pp. 1–6.

[43] G. Zervakis, K. Koliogeorgi, D. Anagnostos, N. Zompakis, and K. Siozios, "Vader: Voltage-driven netlist pruning for cross-layer approximate arithmetic circuits," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 27, no. 6, pp. 1460–1464, 2019.

[44] H. Saadat, H. Bokhari, and S. Parameswaran, "Minimally biased multipliers for approximate integer and floating-point multiplication," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 37, no. 11, pp. 2623–2635, 2018.

[45] S. S. Sarwar, S. Venkataramani, A. Ankit, A. Raghunathan, and K. Roy, "Energy-efficient neural computing with approximate multipliers," *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, vol. 14, no. 2, pp. 1–23, 2018.

[46] Z.-G. Tasoulas, G. Zervakis, I. Anagnostopoulos, H. Amrouch, and J. Henkel, "Weight-oriented approximation for energy-efficient neural network inference accelerators," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 67, no. 12, pp. 4670–4683, 2020.

[47] G. Zervakis, S. Xydis, D. Soudris, and K. Pekmestzi, "Multi-level approximate accelerator synthesis under voltage island constraints," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 66, no. 4, pp. 607–611, 2018.

[48] S. Lee, L. K. John, and A. Gerstlauer, "High-level synthesis of approximate hardware under joint precision and voltage scaling," in *Design, Automation & Test in Europe Conference & Exhibition (DATE), 2017*. IEEE, 2017, pp. 187–192.