2024 IEEE 6th International Conference on AI Circuits and Systems (AICAS 2024)

22-25 April 2024
Abu Dhabi, United Arab Emirates

IEEE Catalog Number: CFP24R18-POD
ISBN: 979-8-3503-8364-5

Copyright © 2024 by the Institute of Electrical and Electronics Engineers, Inc.
All Rights Reserved

Copyright and Reprint Permissions: Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limit of U.S. copyright law for private use of patrons those articles in this volume that carry a code at the bottom of the first page, provided the per-copy fee indicated in the code is paid through Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923.

For other copying, reprint or republication permission, write to IEEE Copyrights Manager, IEEE Service Center, 445 Hoes Lane, Piscataway, NJ 08854. All rights reserved.

****** This is a print representation of what appears in the IEEE Digital Library. Some format issues inherent in the e-media version may also appear in this print version.***

IEEE Catalog Number:	CFP24R18-POD
ISBN (Print-On-Demand):	979-8-3503-8364-5
ISBN (Online):	979-8-3503-8363-8
ISSN:	2834-9830

Additional Copies of This Publication Are Available From:

Curran Associates, Inc
57 Morehouse Lane
Red Hook, NY 12571 USA

Phone:	(845) 758-0400
Fax:	(845) 758-2633
E-mail:	curran@proceedings.com
Web:	www.proceedings.com

AICAS 2024 TABLE OF CONTENTS

A1L-A
AI and ML at the Edge

Date: Tuesday, April 23, 2024
Time: 9:00 - 10:30
Room: B 00055
Chair: Gianluca Setti, KAUST

Analog Features Extractor for Ultra-Low Power Embedded AI Listening and Keyword Spotting **1**

Sebastian Marzetti, Valentin Gies, Valentin Barchasz, Hervé Barthélemy, Hervé Glotin

Université de Toulon, France

On-Device Domain Learning for Keyword Spotting on Low-Power Extreme Edge Embedded Systems **6**

Cristian Cioflan[1], Lukas Cavigelli[2], Manuele Rusci[3], Miguel de Prado[4], Luca Benini[1,5]

[1]ETH Zürich, Switzerland; [2]Huawei Technologies, Switzerland; [3]Katholieke Universiteit Leuven, Belgium;
[4]Verese, Canada; [5]University of Bologna, Italy

Hardware-Friendly Lightweight Convolutional Neural Network Derivation at the Edge **11**

Li Zhang, Ahmed M. Eltawil, Khaled N. Salama

King Abdullah University of Science and Technology, Saudi Arabia

HeteroEML: Heterogeneous Design Methodology of Edge Machine Learning on CPU+FPGA Platform **16**

Yi-Ting Wu, Tzu-Yun Yen, Yu-Pei Lin, Bo-Cheng Lai

National Yang Ming Chiao Tung University, Taiwan

PF-Training: Parameter Freezing for Efficient On-Device Training of CNN-Based Object Detectors in Low-Resource Environments **21**

Dayoung Chun[1], Hyuk-Jae Lee[1], Hyun Kim[2]

[1]Seoul National University, Korea; [2]Seoul National University of Science and Technology, Korea

A1L-B
Spiking Neural Networks I

Date: Tuesday, April 23, 2024
Time: 9:00 - 10:30
Room: B 00107
Chair: Zihang Song, King's College London

SNN Architecture for Differential Time Encoding Using Decoupled Processing Time **26**

Daniel Windhager[1], Bernhard A. Moser[2], Michael Lunglmayr[2]

[1]Silicon Austria Labs, Austria; [2]Johannes Kepler University, Austria

Stochastic Spiking Attention: Accelerating Attention with Stochastic Computing in Spiking Networks **31**

Zihang Song, Prabodh Katti, Osvaldo Simeone, Bipin Rajendran

King's College London, United Kingdom

Neuromorphic Event-Based Line Detection on SpiNNaker **36**

Amélie Gruel[1], Adrien F. Vincent[1], Jean Martinet[2], Sylvain Saïghi[1,3]

[1]Université de Bordeaux, France; [2]Université Côte d'Azur, France; [3]CNRS@CREATE, Singapore

Active Dendrites Enable Efficient Continual Learning in Time-to-First-Spike Neural Networks **41**

Lorenzo Pes[1,2], Rick Luiken[1], Federico Corradi[1], Charlotte Frenkel[2]

[1]Eindhoven University of Technology, Netherlands; [2]Delft University of Technology, Netherlands

Neuromorphic Temporal Pattern Detection with Spiking Neural Networks Using Synaptic Delays **46**

Hugo Bulzomi[1,2], Yuta Nakano[2], Rémy Bendahan[2], Jean Martinet[1,2]

[1]Université Côte d'Azur, France; [2]IMRA Europe, France

A1L-C
ML Applications in Communications and ASR

Date: Tuesday, April 23, 2024
Time: 9:00 - 10:30
Room: KIC
Chair: Alex James, Digital University Kerala

RNN-Based Low Complexity High Speed Channel Estimation Architectures for Vehicular Networks **51**

Syed Asrar Ul Haq, Bhavesh Prasad Dangwal, Sumit Darak

Indraprastha Institute of Information Technology Delhi, India

Low-Complexity High Speed Residual Network-Augmented Channel Estimation for mmWave Massive MIMO **56**

Anurag Gulati, Syed Asrar Ul Haq, Sumit Darak, Varun Singh

Indraprastha Institute of Information Technology Delhi, India

A CNN-Based One-Shot Blind RX-Side-Only Equalization Scheme for High-Speed SerDes Links **61**

Yiheng Hui[1], Yuanlin Nong[1], He Ma[1], Jingjing Lv[1], Lei Chen[2], Li Du[1], Yuan Du[1]

[1]Nanjing University, China; [2]Beijing Microelectronics Technology Institute Beijing, China

AI for Antenna Design Re-engineering: Yes, Radiation Patterns Predict Antenna Structures! **66**

Anitha Gopi, Elizabeth George, Rahul RK, Alex James

Digital University Kerala, India

Enhancing ASR Performance through Relative Word Frequency in OCR and Normal Word Frequency Analysis **71**

Kyudan Jung[1], Seungmin Bae[2], Nam-Joon Kim[2], Jaepil Lim[2], Hyun Gon Ryu[3], Hyuk-Jae Lee[2]

[1]Chung-Ang University, Korea; [2]Seoul National University, Korea; [3]NVIDIA, Korea

A2L-A
Hardware Accelerator ICs

Date: Tuesday, April 23, 2024
Time: 13:10 - 14:40
Room: B 00055
Chair: Mahmoud Al-Qutayri, Khalifa University

An End-to-End RNS CNN Accelerator **75**

Vasilis Sakellariou[1], Vassilis Paliouras[2], Ioannis Kouretas[2], Hani Saleh[1], Thanos Stouraitis[1]

[1]Khalifa University, U.A.E.; [2]University of Patras, Greece

A 28nm 343.5fps/W Vision Transformer Accelerator with Integer-Only Quantized Attention Block **80**

Cheng-Chen Lin, Wei Lu, Po-Tsang Huang, Hung-Ming Chen

National Yang Ming Chiao Tung University, Taiwan

A 54.61-GOPS 96.35-mW Digital Logic Accelerator for Underwater Object Recognition DNN Using 40-nm CMOS Process **85**

Chua-Chin Wang[1], Shih-Heng Luo[1], Hsin-Che Wu[1], Ralph Gerard B. Sangalang[1,2], Chewn-Pu Jou[3], Harry Hsia[3], Lan-Chou Cho[3]

[1]National Sun Yat-Sen University, Taiwan; [2]Batangas State University, Philippines;
[3]Taiwan Semiconductor Manufacturing Company, Taiwan

A 266.7 TOPS/W Computing-In Memory Using Single-Ended 6T 4-Kb SRAM in 16-nm FinFET CMOS Process **90**

Cheng-Yao Lo[1], Lean Karlo Santos Tolentino[1,2], Jhih-Ying Ke[1], Jeffrey S. Walling[3], Yang Yi[3], Chua-Chin Wang[1,4]

[1]National Sun Yat-Sen University, Taiwan; [2]Technological University of the Philippines, Philippines; [3]Virginia Polytechnic Institute and State University, United States; [4]Vel Tech University, India

GA-Optimized 6.0-Gbps DDR5 SDRAM I/O Buffer Design for 16-nm FinFET CMOS Process **95**

Jhih-Ying Ke[1], Lean Karlo Santos Tolentino[1,2], Cheng-Yao Lo[1], Tzung-Je Lee[1], Chua-Chin Wang[1,3]

[1]National Sun Yat-Sen University, Taiwan; [2]Technological University of the Philippines, Taiwan;
[3]Vel Tech University, India

A2L-B
Spiking Neural Networks II

Date: Tuesday, April 23, 2024
Time: 13:10 - 14:40
Room: B 00107
Chair: Hani Saleh, Khalifa University (TBC)

DualNet: Efficient Integration of Artificial Neural Network and Spiking Neural Network with Equivalent Conversion 100

Seongyon Hong, Sangyeob Kim, Soyeon Kim, Hoi-Jun Yoo

Korea Advanced Institute of Science and Technology, Korea

Memory-Centric Computing for Image Classification Using SNN with RRAM 105

Nada AbuHamra, Baker Mohammad

Khalifa University, U.A.E.

SPVT: Spiked Pyramid Vision Transformer 110

Yazhuo Guo, Yuhan Qin, Song Chen, Yi Kang

University of Science and Technology of China, China

SNN-LIF Model for Glaucoma Classification 114

Harshiv Chandra[1], Akash Ghosh[2], Rahul Singh[3], M.B. Srinivas[4]

[1]Reckitt, U.A.E.; [2]Indian Institute of Technology Roorkee, India; [3]Birla Institute of Technology and Science, Pilani, U.A.E.; [4]Birla Institute of Technology and Science, Pilani, India

Self-Resetting Magnetic Tunnel Junction Neuron-Based Spiking Neural Networks 119

Aijaz H. Lone, Daniel N. Rahimi, Hossein Fariborzi, Gianluca Setti

King Abdullah University of Science and Technology, Saudi Arabia

A2L-C
Security, Safety and Trustworthiness

Date: Tuesday, April 23, 2024
Time: 13:10 - 14:40
Room: KIC
Chair: Abdulhadi Shoufan, Khalifa University

Driving towards Safety: Online PPG-Based Drowsiness Detection with TCNs 124

Pierangelo Maria Rapa[1,2], Mattia Orlandi[1], Andrea Amidei[2], Alessio Burrello[4], Roberto Rabbeni[5], Paolo Pavan[2], Luca Benini[1,3], Simone Benatti[1,2]

[1]University of Bologna, Italy; [2]University of Modena and Reggio Emilia, Italy; [3]ETH Zürich, Switzerland; [4]Polytechnic University of Turin, Italy; [5]Maserati, Italy

SVD-Based Peephole and Clustering to Enhance Trustworthiness in DNN Classifiers 129

Livia Manovi[1], Lorenzo Capelli[1], Alex Marchioni[1], Filippo Martinini[1], Gianluca Setti[1,2], Mauro Mangia[1], Riccardo Rovatti[1]

[1]Università di Bologna, Italy; [2]King Abdullah University of Science and Technology, Saudi Arabia

Strategic Deployment of Honeypots in Blockchain-Based IoT Systems 134

Daniel Commey[1], Sena Hounsinou[2], Garth Crosby[2]

[1]Texas A&M University, United States; [2]Metro State University, United States

Is Semantic Communication for Autonomous Driving Secured against Adversarial Attacks? 139

Soheyb Ribouh[1], Abdenour Hadid[2]

[1]Université Le Havre Normandie, France; [2]Sorbonne University Abu Dhabi, U.A.E.

Attacking a Joint Protection Scheme for Deep Neural Network Hardware Accelerators and Models 144

Simon Wilhelmstätter[1], Joschua Conrad[1], Devanshi Upadhyaya[2], Ilia Polian[2], Maurits Ortmanns[1]

[1]Universität Ulm, Germany; [2]Universität Stuttgart, Germany

A3L-A
Processors for AI

Date: Tuesday, April 23, 2024
Time: 15:00 - 16:30
Room: B 00055
Chair: Thanos Stouraitis, Khalifa University

TENG: A General-Purpose and Efficient Processor Architecture for Accelerating DNN **149**
Zekun Zhang[1,2], Yujie Cai[1], Tianjiao Liao[2], Chengyu Xu[2], Xin Jiao[2]
[1]Fudan University, China; [2]SenseTime Research, China

A Hybrid Heterogeneous Neural Network Accelerator Based on Systolic Array **154**
Zilin Wang[1], Yi Zhong[1], Guang Chen[1], Shuo Feng[1], Youming Yang[1], Xiaoxin Cui[1,4], Yuan Wang[1,4]
[1]Peking University, China; [2]Beijing Advanced Innovation Center for Integrated Circuits, China

Reusing Softmax Hardware Unit for GELU Computation in Transformers **159**
Christodoulos Peltekis, Kosmas Alexandridis, Giorgos Dimitrakopoulos
Democritus University of Thrace, Greece

A Low-Power 3D Point Clouds Matching Processor with 1D-CNN Prediction and CAM-Based In-Memory kNN Searching **164**
Jeongmin Shin, Hoichang Jeong, Seungbin Kim, Keonhee Park, Sangho Lee, Kyuho Lee
Ulsan National Institute of Science and Technology, Korea

Advancing Hardware Implementation of Hyperdimensional Computing for Edge Intelligence **169**
Eman Hassan, Meriem Bettayeb, Baker Mohammad
Khalifa University of Science and Technology, U.A.E.

A3L-B
In-memory Computing

Date: Tuesday, April 23, 2024
Time: 15:00 - 16:30
Room: B 00107
Chair: Khaled Nabil Salama, KAUST (TBC)

PyAIM: PYNQ-Based Scalable Analog In-Memory Computing Prototyping Platform 174

Minsang Yu, Minuk Hong, Sugil Lee, Seungsu Kim, Jongeun Lee

Ulsan National Institute of Science and Technology, Korea

ILP-Based Multi-branch CNNs Mapping on Processing-in-Memory Architecture 179

Haodong Han, Junpeng Wang, Bo Ding, Song Chen

University of Science and Technology of China, China

An Area and Energy-Efficient SRAM Based Time-Domain Compute-in-Memory Architecture for BNN 184

Subhradip Chakraborty[1], Dinesh Kushwaha[2], Anand Bulusu[2], Sudeb Dasgupta[2]

[1]Rajiv Gandhi Institute of Petroleum Technology, India; [2]Indian Institute of Technology Roorkee, India

TNN-CIM: An In-SRAM CMOS Implementation of TNN-Based Synaptic Arrays with STDP Learning 189

Harideep Nair, David Barajas-Jasso, Quinn Jacobson, John Paul Shen

Carnegie Mellon University, United States

A3L-C
Object Detection and Applications

Date: Tuesday, April 23, 2024
Time: 15:00 - 16:30
Room: KIC
Chair: Naoufel Werghi, Khalifa University

End-to-End Edge Neuromorphic Object Detection System **194**

D.A. Silva[1], A. Shymyrbay[1], K. Smagulova[1], A. Elsheikh[2], M.E. Fouda[3], A.M. Eltawil[1]

[1]King Abdullah University of Science and Technology, Saudi Arabia; [2]Cairo University, Saudi Arabia; [3]Rain Neuromorphics, Inc., United States

Fast Object Detection Algorithm Using Edge-Based Operation Skip Scheme with Viola-Jones Method **199**

Cheol-Ho Choi, Joonhwan Han, Jeongwoo Cha, Jungho Shin, Hyun Woo Oh

Hanwha Systems, Co., Ltd., Korea

An Efficient and Fast Filter Pruning Method for Object Detection in Embedded Systems **204**

Hyunjun Ko[1], Jin-Ku Kang[1], Yongwoo Kim[2]

[1]Inha University, Korea; [2]Korea National University of Education, Korea

LATTE: Low-Precision Approximate Attention with Head-Wise Trainable Threshold for Efficient Transformer **208**

Jiing-Ping Wang, Ming-Guang Lin, An-Yeu Wu

National Taiwan University, Taiwan

B1L-A
FPGA-based Accelerators

Date: Wednesday, April 24, 2024
Time: 9:00 - 10:30
Room: B 00055
Chair: Abdulhadi Shoufan, Khalifa University

Exploration for Efficient Depthwise Separable Convolution Networks Deployment on FPGA **213**

Zhijie Huang, Ao Qie, Chen Zhang, Jie Yang, Xin'An Wang

Peking University, China

Exploring Memory Access Techniques for Efficient FPGA Based 3D CNN Accelerator Design **218**

Fatima Hameed Khan, Muhammad Adeel Pasha, Shahid Masud

Lahore University of Management Sciences, Pakistan

Towards Automated FPGA Compilation of Spiking Neural Networks **223**

Ayan Shymyrbay[1], Mohammed E. Fouda[2], Ahmed Eltawil[1]

[1]King Abdullah University of Science and Technology, Saudi Arabia; [2]Rain Neuromorphics, Inc., United States

FPGA-Based CNN Acceleration Using Pattern-Aware Pruning **228**

Léo Pradels[1], Silviu-Ioan Filip[1], Olivier Sentieys[1], Daniel Chillet[1], Thibaut Le Calloch[2]

[1]Université de Rennes, Inria, CNRS, IRISA, France; [2]Safran Electronics & Defense, France

Efficient Hardware Implementation of Artificial Neural Networks on FPGA **233**

Kasem Khalil[1,3], Tamador Mohaidat[1], Mahmoud Darwich[2], Ashok Kumar[4], Magdy Bayoumi[4]

[1]University of Mississippi, United States; [2]University of Mount Union, United States; [3]Assiut University, Egypt;
[4]University of Louisiana at Lafayette, United States

B1L-B
Analog Accelerators

Date: Wednesday, April 24, 2024
Time: 9:00 - 10:30
Room: B 00107
Chair: Mihai Sanduleanu, Khalifa University

An Analog Neural Network for Estimating Sea State or Wave Height from Inertial Sensor Data **238**

Vassilis Alimisis, Andreas Papathanasiou, Georgios Georgousis, Paul P. Sotiriadis

National Technical University of Athens, Greece

Characterization of an Analog MAC Cell with Multi-bit Resolution for AI Inference Accelerators **243**

Raphael Nägele, Jakob Finkbeiner, Markus Grözing, Manfred Berroth, Georg Rademacher

Universität Stuttgart, Germany

Evaluating an Analog Main Memory Architecture for All-Analog In-Memory Computing Accelerators **248**

Kazybek Adam[1], Dipesh Monga[1], Omar Numan[1], Gaurav Singh[1], Kari Halonen[1], Martin Andraud[1,2]

[1]Aalto University, Finland; [2]Université catholique de Louvain, Belgium

Characterization of a Femtojoule Voltage-to-Time Converter with Rectified Linear Unit Characteristic for Analog Neural Network Inference Accelerators **253**

Jakob Finkbeiner, Raphael Nägele, Markus Grözing, Manfred Berroth, Georg Rademacher

Universität Stuttgart, Germany

Adaptive Mixed MLC-SLC FeFET Mapping for CIM AI Applications through Simulated Annealing **258**

Alptekin Vardar, Franz Müller, İpek Geçin, Nellie Laleni, Thomas Kämpfe

Fraunhofer IPMS, Germany

B1L-C
Anomaly Detection Techniques and Applications

Date:	Wednesday, April 24, 2024
Time:	9:00 - 10:30
Room:	KIC
Chair:	Khaled Nabil Salama, KAUST

Early-Exit with Class Exclusion for Efficient Inference of Neural Networks **263**

Jingcun Wang[1], Bing Li[2], Grace Li Zhang[1]

[1]Technische Universität Darmstadt, Germany; [2]Technische Universität München, Germany

An Efficient Anomalous Sound Detection by Robust Processing and Reformation of Objective **268**

Tsung-Lin Tsai, Yi-Cheng Lo, An-Yeu Wu

National Taiwan University, Taiwan

Enhancing Anomaly Detection with Entropy Regularization in Autoencoder-Based Lightweight Compression **273**

Andriy Enttsel[1], Alex Marchioni[1], Gianluca Setti[1,2], Mauro Mangia[1], Riccardo Rovatti[1]

[1]Università di Bologna, Italy; [2]King Abdullah University of Science and Technology, Saudi Arabia

Auditory Anomaly Detection Using Recurrent Spiking Neural Networks **278**

Shreya Sanjeev Kshirasagar[1,2], Benjamin Cramer[1], Andre Guntoro[1], Christian Mayr[2]

[1]Robert Bosch GmbH, Germany; [2]Technische Universität Dresden, Germany

Fine-Tuned Based Transfer Learning with Temporal Attention and Physics-Informed Loss for Bearing RUL Prediction **282**

Pavan Kumar MP[1], Zhe-Xiang Tu[1], Kun-Chih Chen[2]

[1]National Sun Yat-Sen University, Taiwan; [2]National Yang Ming Chiao Tung University, Taiwan

B2L-A
Flexible ML Architectures

Date: Wednesday, April 24, 2024
Time: 10:50 - 12:20
Room: B 00055
Chair: Terry Tao Ye, Southern University of Science and Technology

FlexBits: A Configurable Lightweight RISC-V Micro-architecture for Flexible Bit-Width Execution **287**

Zhiyuan Xu, Xinyu Kang, Xingbo Wang, Bingzhen Chen, Terry Tao Ye

Southern University of Science and Technology, China

RISCV-FNT: A Fast FNT-Based RISC-V Processor for CNN Acceleration **292**

Bingzhen Chen[1], Xingbo Wang[1], Yucong Huang[1,2], Zhiyuan Xu[1], Terry Tao Ye[1]

[1]Southern University of Science and Technology, China; [2]Hong Kong University of Science and Technology, China

RTPE: A High Energy Efficiency Inference Processor with RISC-V Based Transformation Mechanism **297**

Zhou Wang[1,2,3], Haochen Du[4], Baoyi Han[5], Yanqing Xu[6], Xiaonan Tang[7], Yang Zhou[1], Zhe Zheng[8], Wenpeng Cui[8], Yanwei Xiong[8], Shaojun Wei[1], Shushan Qiao[2,3], Shouyi Yin[1]

[1]Tsinghua University, China; [2]Institute of Microelectronics of Chinese Academy of Sciences, China; [3]University of Chinese Academy of Sciences, China; [4]Hong Kong University of Science and Technology, China; [5]China Telecom Co.,Ltd., Beijing, China

RCPE: An Excellent Performance Training Processor with RISC-V Based Compression Mechanism **302**

Zhou Wang[1,2,3], Haochen Du[4], Baoyi Han[5], Yanqing Xu[6], Xiaonan Tang[7], Yang Zhou[1], Zhe Zheng[8], Wenpeng Cui[8], Yanwei Xiong[8], Shaojun Wei[1], Shushan Qiao[2,3], Shouyi Yin[1]

*[1]Tsinghua University, China; [2]Institute of Microelectronics of Chinese Academy of Sciences, China;
[3]University of Chinese Academy of Sciences, China; [4]Hong Kong University of Science and Technology, China;
[5]China Telecom Co.,Ltd., Beijing, China*

A Low-Latency and Scalable Vector Engine with Operation Fusion for Transformers **307**

Mincheol Cha, Keehyuk Lee, Xuan Truong Nguyen, Hyuk-Jae Lee

Seoul National University, Korea

B2L-B
Neuromorphic Systems

Date: Wednesday, April 24, 2024
Time: 10:50 - 12:20
Room: B 00107
Chair: Mohamad Sawan, Westlake University

NeuroSORT: A Neuromorphic Accelerator for Spike-Based Online and Real-Time Tracking **312**

Ziyang Shen[1], Xiaoxu Xie[2], Chaoming Fang[1], Fengshi Tian[3], Shunli Ma[2], Jie Yang[1], Mohamad Sawan[1]

[1]Westlake University, China; [2]Fudan University, China; [3]Hong Kong University of Science and Technology, China

A 0.2-pJ/SOP Digital Spiking Neuromorphic Processor with Temporal Parallel Dataflow and Efficient Synapse Memory Compression **317**

Yu-Hsuan Lin, Kea-Tiong Tang

National Tsing Hua University, Taiwan

High-Density Digital Neuromorphic Processor with High-Precision Neural and Synaptic Dynamics and Temporal Acceleration **322**

Jongkil Park, YeonJoo Jeong, Jaewook Kim, Suyoun Lee, Joon Young Kwak, Jong-Keuk Park, Inho Kim

Korea Institute of Science and Technology, Korea

Device Feasibility Analysis of Multi-level FeFETs for Neuromorphic Computing **327**

Arnob Saha[1], Bibhas Manna[1], Sen Lu[1], Zhouhang Jiang[2], Kai Ni[2], Abhronil Sengupta[2]

[1]Pennsylvania State University, United States; [2]University of Notre Dame, United States

On-Chip Incremental Learning Based on Unsupervised STDP Implementation **332**

Guang Chen[1], Jian Cao[1], Shuo Feng[1], Zilin Wang[1], Yi Zhong[1], Qibin Li[1], Xiongbo Zhao[1,4], Xing Zhang[1], Yuan Wang[1]

[1]Peking University, China; [2]Beijing Aerospace Automatic Control Institute, China

B2L-C
Computer Vision Algorithms and Architectures

Date: Wednesday, April 24, 2024
Time: 10:50 - 12:20
Room: KIC
Chair: Naoufel Werghi, Khalifa University

Optimizing Vision Transformers: Leveraging Max and Min Operations for Efficient Pruning **337**

Philippe Bich[1], Chiara Boretti[1], Luciano Prono[1], Fabio Pareschi[1,2], Riccardo Rovatti[2], Gianluca Setti[2,3]

[1]Politecnico di Torino, Italy; [2]Università di Bologna, Italy; [3]King Abdullah University of Science and Technology, Saudi Arabia

Shuffle Vision Transformer: Lightweight, Fast and Efficient Recognition of Driver's Facial Expression **342**

Ibtissam Saadi[1,2], Douglas W. Cunningham[2], Abdelmalik Taleb-Ahmed[1], Abdenour Hadid[3], Yassin El Hillali[1]

[1]Université Polytechnique Hauts-de-France, France; [2]Brandenburg University of Technology Cottbus-Senftenberg, Germany; [3]Sorbonne University Abu Dhabi, U.A.E.

Adaptive Image Downscaling for Rate-Accuracy-Latency Optimization of Task-Target Image Compression **347**

Hangyul Choi[1], Seongmoon Jeong[1], Sangwoon Kwak[2], Soon-Heung Jung[1], Jong Hwan Ko[1]

[1]Sungkyunkwan University, Korea; [2]Electronics and Telecommunications Research Institute, Korea

Fast, Efficient and Lightweight Compressed Image Super-resolution Network for Edge Devices **352**

Jaemyung Kim[1], Jin-Ku Kang[1], Yongwoo Kim[2]

[1]Inha University, Korea; [2]Korea National University of Education, Korea

GaitSpike: Event-Based Gait Recognition with Spiking Neural Network **357**

Ying Tao[1], Chip-Hong Chang[1,2], Sylvain Saïghi[2,3], Shengyu Gao[1,2]

[1]Nanyang Technological University, Singapore; [2]CNRS@CREATE, Singapore; [3]Université de Bordeaux, France

B3L-A
Arithmetic for Machine Learning

Date: Wednesday, April 24, 2024
Time: 14:20 - 15:50
Room: B 00055
Chair: Shervin Vakili, INRS-EMT University

Approximate Adder Tree Design with Sparsity-Aware Encoding and In-Memory Swapping for SRAM-Based Digital Compute-in-Memory Macros **362**

Ming-Guang Lin, Jiing-Ping Wang, Cheng-Yang Chang, An-Yeu Wu

National Taiwan University, Taiwan

A Cost-Effective Baugh-Wooley Approximate Multiplier for FPGA-Based Machine Learning Computing **367**

Shervin Vakili

INRS-EMT University, Canada

OTFC-LSTM: An Efficient Design of LSTM Accelerator Based on On-the-Fly CORDIC **372**

Zhibin Luo, Junyi Mai, Enyi Yao

South China University of Technology, China

Progressive Variable Precision DNN with Bitwise Ternary Accumulation **377**

Junnosuke Suzuki, Mari Yasunaga, Kazushi Kawamura, Thiem Van Chu, Masato Motomura

Tokyo Institute of Technology, Japan

Energy-Efficient Ternary Multiplier **382**

Rishi Agrawal[1], Narayanabhatla Savyasachi Abhijith[1], Uppugunduru Anil Kumar[2], Sreehari Veeramachaneni[3], Syed Ershad Ahmed[1]

[1]Birla Institute of Technology and Science, Pilani, India; [2]ICFAI Foundation for Higher Education, India; [3]Gokaraju Rangaraju Institute of Engineering and Technology, India

B3L-B
Memristors for ML/AI

Date: Wednesday, April 24, 2024
Time: 14:20 - 15:50
Room: B 00107
Chair: Bhaskar Choubey, Siegen University

Weight Update Scheme for 1T1R Memristor Array Based Equilibrium Propagation **388**

Brady Taylor[1], Xiaoxuan Yang[1,2], Hai Li[1]

[1]Duke University, United States; [2]University of Virginia, United States

Dynamic Detection and Mitigation of Read-Disturb for Accurate Memristor-Based Neural Networks **393**

Sumit Diware[1], Mohammad Amin Yaldagard[1], Anteneh Gebregiorgis[1], Rajiv V. Joshi[2], Said Hamdioui[1], Rajendra Bishnoi[1]

[1]Delft University of Technology, Netherlands; [2]IBM Thomas J. Watson Research Centre, United States

The Ouroboros of Memristors: Neural Networks Facilitating Memristor Programming **398**

Zhenming Yu[1,2], Ming-Jay Yang[1,2], Jan Finkbeiner[1,2], Sebastian Siegel[2], John Paul Strachan[1,2], Emre Neftci[1,2]

[1]RWTH Aachen, Germany; [2]Forschungszentrum Jülich GmbH, Germany

Effect of Line Resistance of Passive Memristive Crossbars on Spiking Neural Network Performance **403**

Pierre Lewden[1], Adrien F. Vincent[2], Jean Tomas[2], Chip-Hong Chang[3], Sylvain Saïghi[1,2]

[1]CNRS@CREATE, Singapore; [2]Université de Bordeaux, France; [3]Nanyang Technological University, Singapore

Vibration May Break the Conductive Filament in Amorphous Germanium Based Memristor **408**

Zidu Li, Phil David Börner, Maurice Müller, Andreas Bablich, Peter Haring Bolívar, Bhaskar Choubey

Universität Siegen, Germany

B3L-C
Convolutional Neural Networks

Date: Wednesday, April 24, 2024
Time: 14:20 - 15:50
Room: KIC
Chair: Vasilis Sakellariou, Khalifa University

An Area-Efficient CNN Accelerator Supporting Global Average Pooling with Arbitrary Shapes **413**

Yichuan Bai[1], Xiaopeng Zhang[1], Qian Wang[1], Jingjing Lv[1], Lei Chen[2], Yuan Du[1], Li Du[1]

[1]Nanjing University, China; [2]Beijing Microelectronics Technology Institute Beijing, China

ConvMapSim: Modeling and Simulating Convolutional Weight Mapping for PIM Arrays **417**

Kang Eun Jeon[1], Wooram Seo[2,3], Johnny Rhe[1], Jong Hwan Ko[1]

[1]Sungkyunkwan University, Korea; [2]Samsung Electronics, Korea

A Feature Map Lossless Compression Framework for Convolutional Neural Network Accelerators **422**

Zekun Zhang[1,2], Xin Jiao[2], Chengyu Xu[2]

[1]Fudan University, China; [2]SenseTime Research, China

Design of Convolutional Neural Network Using a Real-Time Probability-Based Sigmoid Function Approximation .... **N/A**

Nour Elshahawy, Sandy A. Wasif, Maggie Mashaly, Eman Azab

German University in Cairo, Egypt

CNN Implementation of Bayesian Plasticity for Robust Learning **432**

Nicolas Ramos, Hai Li

Duke University, United States

B4L-A
Quantization in Machine Learning Systems

Date: Wednesday, April 24, 2024
Time: 16:10 - 17:40
Room: B 00055
Chair: Vassilis Paliouras, University of Patras

A Novel Number Representation and Its Hardware Support for Accurate Low-Bit Quantization on Large Recommender Systems **437**

Yu-Da Chu, Pei-Hsuan Kuo, Lyu-Ming Ho, Juinn-Dar Huang

National Yang Ming Chiao Tung University, Taiwan

AdaQAT: Adaptive Bit-Width Quantization-Aware Training **442**

Cédric Gernigon[1], Silviu-Ioan Filip[1], Olivier Sentieys[1], Clément Coggiola[2], Mickaël Bruno[2]

[1]Université de Rennes, Inria, CNRS, IRISA, France; [2]National Centre for Space Studies, France

HLC: A Hardware-Friendly Quantization and Cache-Based Accelerator for Transformer **447**

Xiangfeng Sun, Yuanting Zhang, Yunchang Jiang, Zheng Li, Bingjin Han, Junyi Mai, Zhibin Luo, Enyi Yao

South China University of Technology, China

PQDE: Comprehensive Progressive Quantization with Discretization Error for Ultra-Low Bitrate MobileNet towards Low-Resolution Imagery **452**

Zongcheng Yue[1], Ran Wu[2], Longyu Ma[1], Chong Fu[3], Chiu-Wing Sham[1]

[1]University of Auckland, New Zealand; [2]Harbin Institute of Technology, China; [3]Northeastern University, China

SRU-Q: Hardware-Friendly Stochastic Rounding Unit-Based Gradient Quantization for CNN Training **457**

Sangbeom Jeong, Dahun Choi, Hyun Kim

Seoul National University of Science and Technology, Korea

B4L-B
Implementation Techniques

Date: Wednesday, April 24, 2024
Time: 16:10 - 17:40
Room: B 00107
Chair: José M. de la Rosa, Institute of Microelectronics of Seville

Resistive Processing Unit-Based On-Chip ANN Training with Digital Memory **462**

Shreyas Deshmukh, Shubham Patil, Anmol Biswas, Vivek Saraswat, Abhishek Kadam, Ajay K. Singh, Laxmeesha Somappa, Maryam Shojaei Baghini, Udayan Ganguly

Indian Institute of Technology Bombay, India

The Multilayer Neural Network Implementation Using SRAM-Based Reconfigurable Cognitive Computation Matrices **467**

I-Chun Liu[1], Chun-Jui Chen[1], Xiu-Zhu Li[1], Yong-Qi Cheng[1], Chung-Wei Huang[1], Pin-Han Lin[1], Hsuan-Wei Pu[1], Sheng-Yu Peng[1], Yu Tsao[2]

[1]National Taiwan University of Science and Technology, Taiwan; [2]Academia Sinica, Taiwan

ACC: Adaptive Compression Framework for Efficient On-Device CNN Training **472**

Gilha Lee, Seungil Lee, Hyun Kim

Seoul National University of Science and Technology, Korea

Deploying Artificial Intelligence in Design Verification to Accelerate IP/SoC Sign-Off with Zero Escape **477**

Surajit Bhattacherjee, Daksh Shah, Dipankar Pal

Birla Institute of Technology and Science, Pilani, India

Variable Layout CMOS Pixels for End-to-End Learning in Task Specific Image Sensors **482**

Bhaskar Choubey, Hendrik Sommerhoff, Michael Moeller, Andreas Kolb

Universität Siegen, Germany

B4L-C
Large Language Models

Date: Wednesday, April 24, 2024
Time: 16:10 - 17:40
Room: KIC
Chair: Xuan-Truong Nguyen, Seoul National University

Accelerator Design Using 3D Stacked Capacitorless DRAM for Large Language Models 487

Janak Sharda, Po-Kai Hsu, Shimeng Yu

Georgia Institute of Technology, United States

Language Modeling on a SpiNNaker2 Neuromorphic Chip 492

Khaleelulla Khan Nazeer[1], Mark Schöne[1], Rishav Mukherji[2], Bernhard Vogginger[1], Christian Mayr[1], David Kappel[3], Anand Subramoney[4]

[1]Technische Universität Dresden, Germany; [2]Birla Institute of Technology and Science, Pilani, India; [3]Ruhr Universität Bochum, Germany; [4]University of London, United Kingdom

Real Post-training Quantization Framework for Resource-Optimized Multiplier in LLMs 497

Minseok Seo, Seongho Jeong, Hyuk-Jae Lee, Xuan-Truong Nguyen

Seoul National University, Korea

Accelerate Large Language Model Inference on Edge TPU with OpenVX Framework 502

You-En Wu[1], Hsin-I Wu[2], Kuo-Cheng Chin[2], Yi-Chun Yang[3], Ren-Song Tsay[3]

[1]Taipei Fuhsing Private School, Taiwan; [2]DeepMentor, Taiwan; [3]National Tsing Hua University, Taiwan

Kinship Verification from Text: Towards Discovering Subtitle Textual Features Shared by Family Members Using Large Language Models 507

Salah Eddine Bekhouche[1], Abdenour Hadid[2]

[1]University of the Basque Country UPV/EHU, Spain; [2]Sorbonne University Abu Dhabi, U.A.E.

C1L-A
Special Session: Intersection of Hardware and Software for Edge AI and TinyML

Date: Thursday, April 25, 2024
Time: 9:00 - 10:30
Room: B 00055
Chairs: Wei Mao, Xidian University
Juntao Guan, Xidian University

ILD-MPQ: Learning-Free Mixed-Precision Quantization with Inter-layer Dependency Awareness **512**

Ruge Xu[1], Qiang Duan[2], Qibin Chen[2], Xinfei Guo[1]

[1]Shanghai Jiao Tong University, China; [2]Inspur Academy of Science and Technology, China

Q-Segment: Segmenting Images In-Sensor for Vessel-Based Medical Diagnosis **517**

Pietro Bonazzi, Yawei Li, Sizhen Bian, Michele Magno

ETH Zürich, Switzerland

Microarchitecture Aware Neural Architecture Search for TinyML Devices **522**

Juntao Guan, Gufeng Liu, Fanhong Zeng, Rui Lai, Ruixue Ding, Zhangming Zhu

Xidian University, China

An Energy-Efficient Look-Up Table Framework for Super Resolution on FPGA **527**

Huanan Li, Shicheng Jia, Juntao Guan, Rui Lai, Shubin Liu, Zhangming Zhu

Xidian University, China

A Hardware-Efficient EMG Decoder with an Attractor-Based Neural Network for Next-Generation Hand Prostheses **532**

Mohammad Kalbasi[1,2], MohammadAli Shaeri[1], Vincent Alexandre Mendez[1], Solaiman Shokur[1], Silvestro Micera[1], Mahsa Shoaran[1]

[1]École Polytechnique Fédérale de Lausanne, Switzerland; [2]Sharif University of Technology, Iran

C1L-B
Design Techniques for ML Hardware

Date: Thursday, April 25, 2024
Time: 9:00 - 10:30
Room: B 00107
Chair: Hani Saleh, Khalifa University

A Calibratable Model for Fast Energy Estimation of MVM Operations on RRAM Crossbars **537**

José Cubero-Cascante, Arunkumar Vaidyanathan, Rebecca Pelke, Lorenzo Pfeifer, Rainer Leupers, Jan Moritz Joseph

RWTH Aachen University, Germany

TPC-NAS: Simple and Effective Neural Architecture Search Based on Total Path Count **542**

Ming-Shan Huang, Pi-Chuan Chen, Tzi-Dar Chiueh

National Taiwan University, Taiwan

Adapting Spatial Transformer Networks across Diverse Hardware Platforms: A Comprehensive Implementation Study **547**

Meriem Bettayeb[1], Eman Hassan[1], Muhammad Umair Khan[1], Yasmin Halawani[2], Hani Saleh[1], Baker Mohammad[1]

[1]Khalifa University, U.A.E.; [2]University of Dubai, U.A.E.

Error Checking for Sparse Systolic Tensor Arrays **552**

Christodoulos Peltekis, Dionysios Filippas, Giorgos Dimitrakopoulos

Democritus University of Thrace, Greece

Too-Hot-to-Handle: Insights into Temperature and Noise Hyperparameters for Differentiable Neural-Architecture-Searches **557**

Joschua Conrad[1], Simon Wilhelmstätter[1], Rohan Asthana[2], Vasileios Belagiannis[2], Maurits Ortmanns[1]

[1]Universität Ulm, Germany; [2]Friedrich-Alexander Universität Erlangen-Nürnberg, Germany

C1L-C
Health and Brain-Machine Applications

Date: Thursday, April 25, 2024
Time: 9:00 - 10:30
Room: KIC
Chair: Laxmeesha Somappa, Indian Institute of Technology Bombay

A 16-Channel GRU Based On-Chip Seizure Classifier for Closed Loop Neuromodulation 562

Lakshmi Iyer, Laxmeesha Somappa

Indian Institute of Technology Bombay, India

A Pulse Correlation Model for PPG Signal Quality Assessment and Key Pulse Extraction 567

Shijie Cheng[1,2], Ruifang Liu[1,2], Hao Wu[1,2], Keith Siu-Fung Sze[2,3], Qianjin Feng[2,3]

[1]Shenzhen University, China; [2]RingConn LLC, United States; [3]Shenzhen Ninenovo Technology Limited, China

An Efficient Ventricular Arrhythmias Detection on Microcontrollers with Optimized 1D CNN 572

Chanwook Hwang, Jaehyeon So, Johnny Rhe, Jiyoon Kim, Juhong Park, Kang Eun Jeon, Jong Hwan Ko

Sungkyunkwan University, Korea

An FPGA Accelerator for 3D Cone-Beam Sparse-View Computed Tomography Reconstruction 577

Yuhan Gu[1], Qing Wu[1], Zhechen Yuan[1], Xiangyu Zhang[1], Wenyan Su[1], Yuyao Zhang[1], Xin Lou[1,2]

[1]ShanghaiTech University, China; [2]Ministry of Education, China

CS-Net: An End-to-End Network for Motor Imagery Brain-Machine Interface with Adaptive Channel Selection and Compressed Sensing 582

Jie Lu, Guanghan Yu, Liyu Qian, Jiacheng Cao, Lirong Zheng, Zhuo Zou

Fudan University, China

C2L-A
Circuits and Systems for AI

Date: Thursday, April 25, 2024
Time: 10:50 - 12:20
Room: B 00055
Chair: Tzi-Dar Chiueh, National Taiwan University

Design and Implementation of an Easy-to-Deploy Energy-Efficient Inference Acceleration System for Multi-precision Neural Networks **587**

Pi-Chuan Chen[1], Yu-Tung Liu[2], Guo-Yang Zeng[2], Tzi-Dar Chiueh[1]

[1]National Taiwan University, Taiwan; [2]Mediatek Inc., Taiwan

MSCA: Model-Driven Search for Optimal Configuration for SpMM Accelerators **592**

Yuhan Qin, Yulong Meng, Haitao Du, Yazhuo Guo, Yi Kang

University of Science and Technology of China, China

OSAHS Detection Capabilities of RingConn Smart Ring: A Feasibility Study **597**

Hongbo Guo[1,2,3], Hao Wu[2,3], Jiahui Xia[4], Yibang Cheng[4], Qianhui Guo[4], Yi Chen[4], Tingyan Xu[4], Jiguang Wang[4], Guoxing Wang[1,2,3]

[1]Shanghai Jiao Tong University, China; [2]RingConn LLC, United States; [3]Shenzhen Ninenovo Technology Limited, China; [4]Shanghai Jiao Tong University School of Medicine, China

Optimizing with Phases: Design and Application Space Assessment for Networks of Phase-Locked Ring Oscillators **602**

A. Bazzi, E. Hardy, J. Ballester, F. Badets, L. Hutin

CEA-Leti, Université Grenoble Alpes, France

Echo State Networks for Accurate and Efficient Modeling of Dynamic Circuits **607**

Paulo Machado[1,2], Ruxandra Barbulescu[1], Luis Miguel Silveira[1,2]

[1]INESC-ID, Portugal; [2]Universidade de Lisboa, Portugal

Additional Paper:

AICAS Grand Challene 2024: Software and Hardware Co-Optimiztion for General Large Language Model Inference on CPU **612**

Junfeng Tan, Guosheng Yu, Jianing Li, Xiaohan Ma, Fang Bao, Evens Pan, David Bian, Yongfu Li, Yuan Du, Li Du, Bo Li, Wei Mao