Special Interest Group on Design Automation
ACM/SIGDA E-NEWSLETTER, Vol. 54, No. 6

SIGDA - The Resource for EDA Professionals

This newsletter is a free service for current SIGDA members and is added automatically with a new SIGDA membership.
Circulation: 2,700

Online archive: https://www.sigda.org/publications/newsletter

SIGDA News

1. Edge AI to Help RISC-V to Take 25% of Processor Market

According to Omdia, AI and automotive applications will help the RISC-V open-standard instruction set architecture (ISA) take nearly 25 percent of the processor market by 2030.

2. ARM to Compete with AI Customers in 2025, Says Report

Processor IP licensor ARM Holdings plc is planning to set up an AI chip division and build a prototype AI processor by spring 2025 and have it on sale in fall 2025, according to Nikkei.

3. Samsung Denies HBM Chips Have Failed Nvidia Tests

Samsung Electronics has denied reports that its high-bandwidth memory (HBM) products have failed to meet Nvidia quality standards, according to Business Korea.

4. Samsung Denies HBM Chips Have Failed Nvidia Tests

Samsung Electronics has denied reports that its high-bandwidth memory (HBM) products have failed to meet Nvidia quality standards, according to Business Korea.

5. CHIPS Act is Raising US Share of Global Production, Says Report

US wafer fab manufacturing capacity is projected to triple between 2022 and 2032, the largest percentage rise in the world, according to a report prepared by Boston Consulting Group.

Messages from the EiCs

Dear ACM/SIGDA members,

We are excited to present to you June E-Newsletter. We encourage you to invite your students and colleagues to be a part of the SIGDA newsletter.

The newsletter covers a wide range of information from the upcoming conferences to technical news and activities of our community. Get involved and contact us if you want to contribute articles or announcements.

The newsletter is evolving. Please let us know what you think.

Happy reading!

Debjit Sinha, Keni Qiu,
Editors-in-Chief,
SIGDA E-News

[6. US Includes Semiconductors in Sweeping China Tariff Increases]
The US government unveiled a broad set of tariffs on Chinese goods Tuesday (May 14), including doubling the tariff on semiconductor imports to 50 percent.

[7. China Doubles Down on Big Fund with US$47.5 Billion Third Phase]
China has launched its third and largest semiconductor investment fund with 344 billion yuan (about US$47.5 billion), according to reports.

What is

Contributing author: Shaoyi Huang <shaoyi.huang@uconn.edu>

AE: Xun Jiao <xun.jiao.30@gmail.com>

What is Sparse Training?

Shaoyi Huang
Assistant Professor (Incoming),
Department of Computer Science,
Stevens Institute of Technology

Increasing deep neural networks (DNNs) model size has shown superior prediction accuracy in a variety of real-world scenarios [1]. However, as model sizes continue to scale, a large amount of computation and heavy memory requirements prohibit the DNN training on resource-limited devices, as well as being environmentally unfriendly. A Google study showed that GPT-3 consumed 1,287 MWh of electricity during training and produced 552 tons of carbon emissions, equivalent to the emissions of a car for 120 years [2]. Fortunately, **sparse training** could significantly mitigate the training costs by using a fixed and small number of nonzero weights in each iteration, while preserving the prediction accuracy for downstream tasks [3].

Recently, three representative training methods with sparsity have garnered significant attention, including the train-prune-retrain approach [4], iterative pruning [5], and sparse training. Among these, sparse training stands out for its ability to achieve the highest accuracy, highest sparsity,

and lowest energy consumption [6]. Two research trends in sparse training have gained enormous popularity. One involves static mask-based methods, where sparsification begins at initialization before training, and the sparse mask remains fixed thereafter. However, this limited flexibility in subnetwork or mask selection often results in sub-optimal subnetworks with poor accuracy. To address this limitation, dynamic mask training has been proposed [3][6], where the sparse mask is periodically updated through drop-and-grow mechanisms to search for better subnetworks with high accuracy.

However, these methods primarily employ either random-based or greedy-based growth strategies. The former typically leads to lower accuracy, while the latter greedily searches for sparse masks with a local minimum over a short distance [7], resulting in limited coverage of weights and thus a sub-optimal sparse model [3]. To better preserve these non-active but important weights, dynamic sparse training, characterized by weight exploitation and coverage exploration, has been proposed to update the sparse mask and search for the 'best possible' subnetwork. Additionally, on GPUs, an algorithm-hardware structured pruning framework has been proposed. It utilizes a highly efficient matrix multiplication kernel, pruneSpMM [9], to exploit parallelism within the SIMD units of streaming multiprocessors, unlocking the acceleration potential of structured sparsity during GNN training. This framework achieves an average speedup of 2× over state-of-the-art GNN training solutions on the NVIDIA A100 GPU.

In summary, dynamic sparse training offers a promising solution for optimizing DNNs training, balancing computational efficiency, memory usage, and prediction accuracy. Researchers aim to refine sparse masks through growth strategies and to design hardware accelerators to unlock the full potential of sparse training, addressing resource constraints and environmental concerns in practical applications.

References

[1] Huang, Shaoyi, Dongkuan Xu, Ian Yen, Yijue Wang, Sung-En Chang, Bingbing Li, Shiyang Chen et al. "Sparse Progressive Distillation: Resolving Overfitting under Pretrain-and-Finetune Paradigm." In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 190-200. 2022.

[2] Patterson, David, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. "Carbon emissions and large neural network training." arXiv preprint arXiv:2104.10350 (2021).

[3] Huang, Shaoyi, Bowen Lei, Dongkuan Xu, Hongwu Peng, Yue Sun, Mimi Xie, and Caiwen Ding. "Dynamic sparse training via balancing the exploration-exploitation

Xun Jiao,
AE for What is

Muhammad Shafique,
AE for What is

Rajsaktish Sankaranarayanan,
AE for Researcher spotlight

Xin Zhao,
AE for Paper submission

Ying Wang,
AE for Technical activities

Jiaqi Zhang,
AE for Technical activities

Paper Deadlines

HiPC'24 – IEEE Int'l Conference on High Performance Computing, Data, And Analytics
Bengaluru, India
Deadline: June 26, 2024 (Abstracts due: June 19, 2024)
Dec. 18-21, 2024
http://www.hipc.org

ASP-DAC'25 - Asia and South Pacific Design Automation Conference
Tokyo Odaiba Miraikan, Japan
Deadline: July 12, 2024 (Abstracts due: July 5, 2024)
Jan. 20-23, 2025
http://www.aspdac.com

ISED'24 – Int'l Conference on Intelligent Systems and Embedded Design
NIT Rourkela, Odisha
Deadline: Aug. 15, 2024
Dec. 20-22, 2024
http://isedconf.org

trade-off." In 2023 60th ACM/IEEE Design Automation Conference (DAC), pp. 1-6. IEEE, 2023.
[4] Chen, Shiyang, Shaoyi Huang, Santosh Pandey, Bingbing Li, Guang R. Gao, Long Zheng, Caiwen Ding, and Hang Liu. "Et: re-thinking self-attention for transformer models on gpus." In Proceedings of the international conference for high performance computing, networking, storage and analysis, pp. 1-18. 2021.
[5] Frankle, Jonathan, and Michael Carbin. "The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks." In International Conference on Learning Representations. 2018.
[6] Huang, Shaoyi, Haowen Fang, Kaleel Mahmood, Bowen Lei, Nuo Xu, Bin Lei, Yue Sun, Dongkuan Xu, Wujie Wen, and Caiwen Ding. "Neurogenesis dynamics-inspired spiking neural network training acceleration." In 2023 60th ACM/IEEE Design Automation Conference (DAC), pp. 1-6. IEEE, 2023.
[7] He, Zheng, Zeke Xie, Quanzhi Zhu, and Zengchang Qin. "Sparse double descent: Where network pruning aggravates overfitting." In International Conference on Machine Learning, pp. 8635-8659. PMLR, 2022.
[8] Li, Zhiyong, Weiyou Wang, Yanyan Yan, and Zheng Li. "PS–ABC: A hybrid algorithm based on particle swarm and artificial bee colony for high-dimensional optimization problems." Expert Systems with Applications 42, no. 22 (2015): 8881-8895.
[9] Gurevin, Deniz, Mohsin Shan, Shaoyi Huang, MD Amit Hasan, Caiwen Ding, and Omer Khan. "PruneGNN: Algorithm-Architecture Pruning Framework for Graph Neural Network Acceleration." In 2024 IEEE International Symposium on High-Performance Computer Architecture (HPCA), pp. 108-123. IEEE, 2024.

SIGDA Partner Journal

ACM Transactions on Design Automation of Electronic Systems, TODAES, publishes innovative work documenting significant research and development advances on the specification, design, analysis, simulation, testing, and evaluation of electronic systems, emphasizing a computer science/engineering orientation. Design automation for machine learning/AI and machine learning/AI for design automation are very much welcomed.

If you are an active researcher in the design and design automation field and would like to be part of the TODAES review board, please fill out the following reviewer form. TODAES recognizes those reviewers that provide timely and high-quality reviews through the Distinguished Review Board. TODAES also recognizes papers and outstanding junior researchers through best paper and rookie of the year award. Authors can send their paper submissions on the manuscript portal.

Upcoming Conferences

IWLS'24 - International Workshop on Logic & Synthesis
ETH Zurich, Zurich, Switzerland
June 6-7, 2024
https://www.iwls.org

GLSVLSI'24 – ACM Great Lakes Symposium on VLSI
Tampa Bay Area, FL
June 12-14, 2024
http://www.glsvlsi.org

DAC'24 – Design Automation Conference
San Francisco, CA
June 23-27, 2024
http://www.dac.com/

OSCAR'24 - Second Workshop on Open-Source Computer Architecture Research
Buenos Aires, Argentina (co-located with ISCA 2024)
June 29, 2024
https://oscar-workshop.github.io/

ISVLSI'24 – IEEE Computer Society Annual Symposium on VLSI
Knoxville, TN
July 1-3, 2024
http://www.ieee-isvlsi.org

ICECET'24 - IEEE International Conference on Electrical, Computer and Energy Technologies
Sydney, Australia
July 25-27, 2024
www.icecet.com

TODAES welcomes special issue proposals from leading researchers/practitioners. Such proposals should be emailed to Joerg Henkel, Senior Associate Editor, at joerg.henkel@kit.edu.

Technical Activities

1. Microchip Rad-Hard FPGAs Offer Low-Power, Zero Configuration Upsets for Space Applications

Microchip's RT PolarFire SoC FPGA is the first real-time Linux capable, RISC-V-based microprocessor subsystem on a flight-proven RT PolarFire FPGA fabric…

2. Arm Brings Transformers to IoT Devices

The next generation of Arm's Ethos micro-NPU, Ethos-U85, is designed to support transformer operations, bringing generative AI models to IoT devices. The IP giant is seeing demand for transformer workloads at the edge, according to Paul Williamson, senior VP and general manager for Arm's IoT line of business, though in much smaller forms than their bigger brothers, large language models (LLMs). For example, Arm has ported vision transformer ViT-Tiny and generative language model TinyLlama-1.1B to the Ethos-U85 so far…

3. AMD Updates AI Engine In New Versal Series

AMD has updated the AI Engine in its second-generation Versal AI Edge Series Gen 2, and has added post-processing capabilities to put an application's entire data path on a single piece of silicon…

4. SpiNNaker-Based Supercomputer Launches in Dresden

A new neuromorphic supercomputer is claiming the title of world's biggest. University of Dresden spinout SpiNNcloud, formed to commercialize technology based on a second generation of Steve Furber's SpiNNaker neuromorphic architecture, is now offering a five-billion neuron supercomputer in the cloud, as well as smaller commercial systems for on-prem use. Among the startup's first customers are Sandia National Labs, Technische Universität München and Universität Göttingen…

ISLPED'24 – ACM/IEEE Int'l Symposium on Low Power Electronics and Design
Newport Beach, CA
Deadline: Mar. 11, 2024 (Abstracts due: Mar. 4, 2024)
Aug. 5-7, 2024
http://www.islped.org

RTCSA'24 - The 30th IEEE International Conference on Embedded and Real-Time Computing Systems and Applications
Sokcho, South Korea
Aug. 21-23, 2024
https://rtcsa2024.github.io/

MLCAD'24 - ACM/IEEE Workshop on Machine Learning for CAD
Snowbird, Utah
Sep. 9-11, 2024
https://mlcad-workshop.org/

ESWEEK'24 - Embedded Systems Week
Raleigh, NC
Sept. 29 - Oct. 4, 2024
http://www.esweek.org

VLSI-SoC'24 – IFIP/IEEE Int'l Conference on Very Large Scale Integration
Tanger, Morocco
Oct. 6-9, 2024
http://www.vlsi-soc.com

PACT'24 - Int'l Conference on Parallel Architectures and Compilation Techniques
Long Beach, CA
Oct. 13-16, 2024
http://www.pactconf.org

ICCAD'24 – IEEE/ACM Int'l Conference on Computer-Aided Design
New Jersey
Oct 27-31, 2024
https://iccad.com/

Job Positions

1. Stockholm University, Sweden

Job Title: Associate Senior Lecturer in Computer and Systems Sciences

Description: Conceptual modeling and especially its focus to enterprise modeling is a research area that contributes to developing and adapting the use of information systems to the operational needs of the organization by developing methods and tools for eliciting business requirements and designing information system solutions. Research in conceptual modeling deals with theories, methods and tools for knowledge management, representation and analysis for creating requirements specification for organizational and system design. Conceptual modeling is at the core of many successful research and application fields, such as requirements engineering, systems analysis, enterprise architecture, and capability management. The main duty is to conduct research in the area of the employment. The duties also include teaching, course development and to have course responsibility at all levels within the area of the announcement, as well as participate in administrative work. Examples of content in the courses are requirements engineering, object-oriented system analysis and design, model-based development tools, and IT in organizations. The applicant is expected to supervise bachelor's and master's theses in the subject area and participate in administrative work. The applicant will be involved in the development of the department's collaboration with business community and other research institutions in the field. For more information, please refer to https://su.varbi.com/en/what:job/jobID:731350/.

2. Hong Kong University of Science and Technology (Guangzhou), China

Job Title: Assistant Professor/Associate Professor/Professor in Internet of Things Thrust

Description: The Internet of Things (IoT) Thrust is an academic department in the Information Hub of HKUST(GZ). It adopts an inter-disciplinary pedagogy to prepare students with integrated knowledge to pursue new innovations and to explore new research frontiers in forming IoT-enabled future smart cities and the digital society. You can find the list of our existing faculty at https://iott.hkust-gz.edu.cn/faculty. We have multiple tenure-track or tenured positions at the ranks of Assistant Professor, Associate Professor, and Professor. We are interested in candidates who can create a multi-disciplinary IoT curriculum and pursue high impact research in a science, technology and engineering-oriented environment. Candidates should hold a Ph.D. degree and work in one or more areas related to the

ICCD'24 – IEEE Int'l Conference on Computer Design
Milan, Italy
Nov. , 2024
http://www.iccd-conf.com

MICRO'24 – IEEE/ACM Int'l Symposium on Microarchitecture
Austin, Texas
Nov. 2-6, 2024
http://www.microarch.org/micro57

iSES'24 – IEEE Int'l Symposium on Smart Electronic Systems
Ahmedabad, India
Dec. 16-18, 2024
http://www.ieee-ises.org

Thrust as follows: Sensors; embedded systems; IoT devices; Artificial intelligence; data science; machine learning; optimization; Wireless communications; computer networking; distributed computation and systems; Security and privacy-enhancing technologies; Metaverse technologies; human-computer interaction; IoT applications. For more information, please refer to https://facultyvacancies.com/assistant-professorassociate-professorprofessor-in-internet-of-things-thrust,i38241.html.

3. University of California Irvine, US

Job Title: Postdoc Position in Electrical Engineering

Description: The postdoctoral researcher is responsible for conducting state-of-the-art research in millimeter-wave (mm-Wave)/Terahertz(THz) integrated circuits for applications including but not limited to 5G/6G joint sensing/communication, imaging, and radar. The researcher will focus on developing new circuits and systems for next generation of wireless systems, designing high-performance RF/mm-wave/THz integrated circuits, carrying out comprehensive simulations and verifications [circuit level and electromagnetic characterizations], implementation and tape out of the chip, measurement and full characterization of the implemented prototype, and publication of impactful papers in flagship conferences and journals in the field of solid-state circuits and microwave engineering. Other duties for this position include mentoring and co-advising graduate students and research projects within the research team, developing and contributing to proposals to secure funding for future projects, and organizing project review meetings. The researcher will contribute to ongoing projects in the group by leading, working with and co-advising graduate students and helping them grow their professional skills, mentoring them toward successful dissertation projects, and assisting them with technical aspects of their projects and publications of their research. For more information, please refer to https://facultyvacancies.com/postdoc-position-in-electrical-engineering,i38967.html.

Notice to authors