

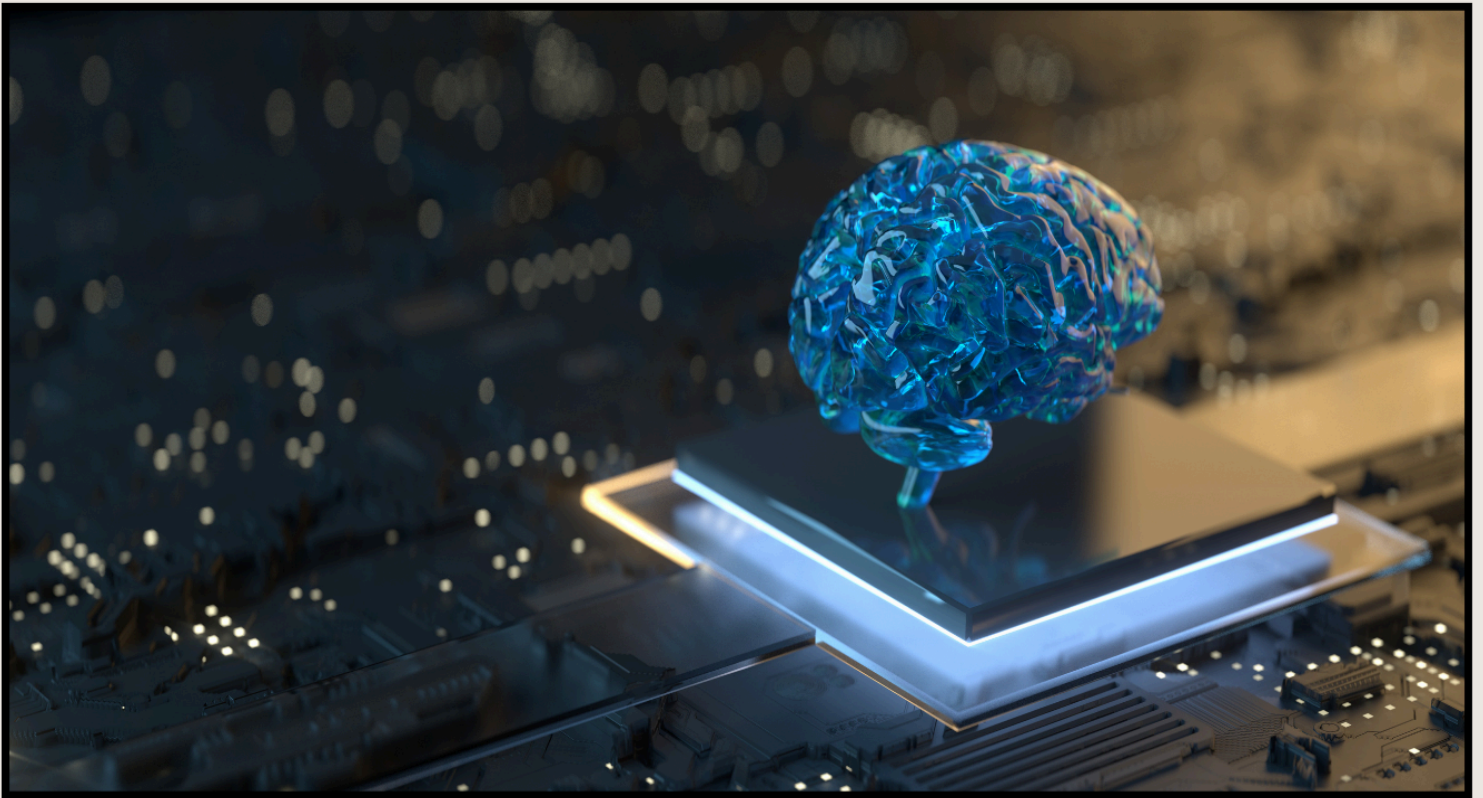


CoCoSys

CENTER FOR THE
CO-DESIGN OF COGNITIVE SYSTEMS

Research Scholar Catalog

Annual Review, March 12-13, 2024



Semiconductor
Research
Corporation



Table of Contents

<u>COCOSYS PI DIRECTORY</u>	<u>5</u>
<u>COCOSYS THEMES & TASKS</u>	<u>7</u>
<u>THEME 1: NEURAL, SYMBOLIC, AND PROBABILISTIC ALGORITHMS</u>	<u>11</u>
<u>I.1 Timur Ibrayev</u>	<u>11</u>
<u>I.2 Christopher Kymn</u>	<u>12</u>
<u>I.3 Yinghan Long</u>	<u>12</u>
<u>I.4 Quanling Zhao</u>	<u>13</u>
<u>I.5 Yudi Xie</u>	<u>13</u>
<u>I.6 Onat Gungor</u>	<u>14</u>
<u>I.7 Amogh Joshi</u>	<u>14</u>
<u>I.8 Vignesh Sundaresha</u>	<u>15</u>
<u>I.9 Shiting (Ginny) Xiao</u>	<u>15</u>
<u>I.10 Flavio Ponzina</u>	<u>16</u>
<u>I.11 Xiaofan Yu</u>	<u>16</u>
<u>I.12 Arijit Dasgupta</u>	<u>17</u>
<u>I.13 Mathieu Huot</u>	<u>17</u>
<u>I.14 Foroozan Karimzadeh</u>	<u>18</u>
<u>I.15 Connor Bybee</u>	<u>18</u>
<u>I.16 Keming Fan</u>	<u>19</u>
<u>I.17 Jeffrey Yu</u>	<u>19</u>
<u>I.18 Jesse Michel</u>	<u>20</u>
<u>I.19 Tiffany Luong & Yoni Friedman</u>	<u>20</u>
<u>I.20 Tian Jin</u>	<u>21</u>
<u>I.21 Guangyu Jiang</u>	<u>21</u>
<u>THEME 2: HARDWARE ALGORITHM CO-DESIGN</u>	<u>24</u>
<u>2.1 Zishen Wan</u>	<u>24</u>
<u>Ritik Raj</u>	<u>24</u>
<u>Hanchen Yang</u>	<u>24</u>



2.2 Zhenyu Wang	25
2.3 Amrit Nagarajan	25
2.4 Surya Selvam	26
2.5 Yuan Liao	26
2.6 Soonha Hwang	27
2.7 Kaining Zhou	27
2.8 Jian Meng	28
2.9 Sixu Li	28
2.10 Yonggan Fu	29
2.11 Haoran You	29
2.12 Yongan Zhang	30
2.13 Neelson Li	30
2.14 Abhimanyu Bambhaniya	31
2.15 Akshat Ramachandran	31
2.16 Jamin Seo	32
2.17 Flavio Ponzina	32
2.18 Sai Saketika Chekuri	33
2.19 Reena Elangovan	33
2.20 Hao Kang	34
2.21 Jingbo Sun	34
2.22 Amey Agrawal	35
THEME 3: TECHNOLOGY-DRIVEN HARDWARE MOTIFS	38
3.1 Abhiroop Bhattacharjee	38
3.2 Piyush Kumar	39
3.3 Md. Nahid Haque Shazon	39
3.4 Siri Narla	40
3.5 Deepika Sharma	40
3.6 Kartik Prabhu	41
3.7 Revanth Koduru	41
3.8 Saion Roy	42
3.9 Akul Malhotra	42
3.10 Connor Talley	43
3.11 Hyung Joon Byun	43
3.12 Jeffry Louis Victor	44



3.13 Che-Kai Liu	44
3.14 Gopikrishnan Raveendran Nair	45
3.15 Aviral Pandey	45
3.16 Mohamed Ibrahim	46
3.17 Tianyi Zhang	46
THEME 4: COLLABORATIVE INTELLIGENCE	48
4.1 Youngeun Kim	48
4.2 Zishen Wan	49
4.3 Yeshwanth Venkatesha	49
4.4 Christopher Richardson	50
4.5 Tyler Lizzo	50
4.6 John Taylor	51
4.7 Benjamin Reichman	51
4.8 Chaojian Li	52
4.9 Anirudh Sundar	52
4.10 Arghadip Das	53
4.11 Mariah Schrum	53
4.12 Xiaofan Yu	54
4.13 Sai Aparna Aketi	54
4.14 Edward Sadler	55
4.15 Cambridge Yang	55



COCOSYS PI DIRECTORY

PRINCIPAL INVESTIGATORS	TASKS	EMAIL
Anand Raghunathan Purdue University	3131.005, 3131.006, 3131.007, 3131.008	raghunathan@purdue.edu
Anca Dragan University of California, Berkeley	3131.013	anca@berkeley.edu
Arijit Raychowdhury Georgia Institute of Technology	3131.011, 3131.012, 3131.015, 3131.003, 3131.007, 3131.009	arijit.raychowdhury@ ece.gatech.edu
Azad Naemi Georgia Institute of Technology	3131.010	an42@gatech.edu
Bruno Olshausen University of California, Berkeley	3131.002, 3131.003, 3131.004	baolshausen@berkeley.edu
Jae-sun Seo Cornell Tech	3131.009, 3131.010, 3131.011, 3131.012	js3528@cornell.edu
James DiCarlo MIT	3131.001	dicarlo@mit.edu
Jan Rabaey University of California, Berkeley	3131.002, 3131.007, 3131.008, 3131.005, 3131.006	jan@eecs.berkeley.edu
Josh Tenenbaum MIT	3131.001, 3131.004, 3131.013	jbt@mit.edu
Kaushik Roy Purdue University	3131.001, 3131.003, 3131.009, 3131.010	kaushik@purdue.edu
Larry Heck Georgia Institute of Technology	3131.013, 3131.014	larryheck@gatech.edu
Michael Carbin MIT	3131.006	mcarbin@csail.mit.edu
Naresh Shanbhag University of Illinois at Urbana-Champaign	3131.001, 3131.002, 3131.003, 3131.004, 3131.015	shanbhag@illinois.edu



PRINCIPAL INVESTIGATORS	TASKS	EMAIL
Priyadarshini Panda Yale University	3131.013, 3131.014, 3131.015	priya.panda@yale.edu
Priyanka Raina Stanford University	3131.006, 3131.009, 3131.012,	praina@stanford.edu
Sumeet K. Gupta Purdue University	3131.01	guptask@purdue.edu
Tajana S. Rosing University of California, San Diego	3131.002, 3131.005, 3131.015	tajana@ucsd.edu
Tushar Krishna Georgia Institute of Technology	3131.005, 3131.006, 3131.007, 3131.008	tushar@ece.gatech.edu
Vijay Raghunathan Purdue University	3131.014	vr@purdue.edu
Yingyan (Celine) Lin Georgia Institute of Technology	3131.006, 3131.009, 3131.012	ylin715@gatech.edu
Yu (Kevin) Cao University of Minnesota	3131.007, 3131.008, 3131.011	yucao@umn.edu

COCOSYS SUPPORT STAFF

NAME	ROLE	EMAIL
Emily Watson	Program & Operations Manager	emily.watson@ece.gatech.edu
Janna Young	Faculty Support Coordinator	jyoung381@gatech.edu
Melissa Donahoe	Financial Analyst	melissa.donahoe@ ece.gatech.edu



COCOSYS THEMES & TASKS

COCOSYS aims to enable the next generation of collaborative human-AI systems through synergistic advances in algorithms, hardware motifs, algorithm-hardware co-design, and collective and collaborative intelligence. To pursue its overarching vision and goals, the center will adopt a vertically integrated approach consisting of synergistic efforts in neural, symbolic, and probabilistic algorithms; algorithm-hardware co-design; technology-driven hardware motifs; and collective and collaborative intelligence.

THEME 1: NEURAL, SYMBOLIC, AND PROBABILISTIC ALGORITHMS

Theme 1 will create the next generation of explainable algorithms, expand the scope of neuro-inspired algorithms from perception to reasoning and decision-making, and uncover the fundamental accuracy-robustness-efficiency tradeoffs in cognitive systems.

Task Number	Task Name
3131.001	Unifying Neural, Symbolic and Probabilistic Models
3131.002	Hyper-dimensional (HD) Information Representations & Processing
3131.003	Computing with Emergent and Dynamical Systems
3131.004	Theoretical Underpinnings of Robustness-accuracy-efficiency Tradeoffs

THEME 2: HARDWARE ALGORITHM CO-DESIGN

Theme 2 will distill the key computational characteristics of future cognitive workloads developed by Theme 1 and use them to drive the design of the next generation of programmable hardware architectures for cognitive computing. This theme will play a key role in ensuring that the developed algorithms are well-matched to the proposed hardware fabrics and vice-versa.

Task Number	Task Name
3131.005	Architectures for Neuro-symbolic-probabilistic Workloads
3131.006	Full-stack Optimization and Software Frameworks for Cognitive Systems
3131.007	Technology and Integration-driven Cognitive Architectures
3131.008	System Evaluation and Benchmarking



THEME 3: TECHNOLOGY-DRIVEN HARDWARE MOTIFS

Theme 3 will design the building blocks of future cognitive hardware platforms by matching the unique capabilities of various CMOS and beyond-CMOS devices and integration technologies to the needs of the workloads, seeking quantum improvements in energy efficiency and performance.

Task Number	Task Name
3131.009	Digital, Mixed-signal and Mixed-mode Cognitive Circuits
3131.010	Technology (Logic, Memory, Interconnect) Evaluation
3131.011	Heterogeneous Integration Driven Cognitive HW Design
3131.012	Hardware Prototyping and Benchmarking

THEME 4: COLLABORATIVE INTELLIGENCE

Theme 4 will specifically focus on the challenges involved in collections of AI agents and how AI agents interact with humans.

Task Number	Task Name
3131.013	Human-AI Collaboration Through Visual and Natural Language Understanding
3131.014	AI-AI Collaboration and Multi-agent Systems
3131.015	Robust, Secure and Privacy-preserving Intelligence



THEME I SESSION OVERVIEW

Tuesday, March 12 from 11:30 AM-12:30 PM

POSTER NO.	PRESENTER	TITLE
I.1	Timur Ibrayev	Machine Perception Based on Foveation and Saccades
I.2	Christopher Kymn	Efficient Visual Disentanglement with Convolutional Sparse Coding and Resonator Networks
I.3	Yinghan Long	Segmented Recurrent Transformer: An Efficient Sequence-to-Sequence Model
I.4	Quanling Zhao	Bridging the Gap between Hyperdimensional Computing and Kernel Methods via the Nyström Method
I.5	Yudi Xie	What's Behind? Unveiling Probabilistic Visual Inference in Humans and Primates
I.6	Onat Gungor	Effective Defenses for Machine Learning-based Intrusion Detection Against Adversarial Attacks
I.7	Amogh Joshi	Unraveling Hybrid Architectures for high-speed Event-based Object Tracking and Autonomous Navigation: A Neuro-symbolic Approach
I.8	Vignesh Sundaresha	Growing Efficient Accurate and Robust Neural Networks on the Edge
I.9	Shiting (Ginny) Xiao	ReSpike: Efficient Residual Frame Video Perception with Spiking Neural Networks
I.10	Flavio Ponzina	Dynamic Ensembling for Dependable and Energy-Efficient Edge AI
I.11	Xiaofan Yu	Lifelong Intelligence Beyond the Edge using Hyperdimensional Computing
I.12	Arijit Dasgupta	Intuitive Physical Reasoning with Probabilistic Programs
I.13	Mathieu Huot	Variational Inference By Automatic Differentiation of Expected Values
I.14	Foroozan Karimzadeh	CIM-Aware Quantization for Energy Efficient Generative AI Models




I.15	Connor Bybee	Efficient Optimization with Higher-Order Ising Machines
I.16	Keming Fan	HDnn: Accelerating few-shot learning using Hyperdimensional Computing with RRAM
I.17	Jeffrey Yu	8-bit Transformer Inference and Fine-tuning for Edge Accelerators
I.18	Jesse Michel	Distributions for Compositionally Differentiating Parametric Discontinuities
I.19	Tiffany Luong and Yoni Friedman	Towards Robust Computer Vision – A Benchmark of Human Perception in Challenging Visual Conditions
I.20	Tian Jin	The Cost of Down-Scaling Language Models: Fact Recall Deteriorates Before In-Context Learning
I.21	Guangyu Jiang	Visual Sensor Fusion for Neuromorphic SLAM

THEME I: NEURAL, SYMBOLIC, AND PROBABILISTIC ALGORITHMS

PRESENTATION DETAILS

Tuesday, March 12 from 11:30 AM-12:30 PM

SCHOLAR	POSTER DETAILS
<p>I.I Timur Ibrayev Purdue University</p>  <p>Email: tibrayev@purdue.edu PI: Kaushik Roy Level: PhD student Expected Graduation Date: November 2024</p>	<p>Title: Machine Perception Based on Foveation and Saccades</p> <p>Abstract: Deep neural networks have become the de facto choice as feature extraction engines, ubiquitously used for computer vision tasks. The current approach is to process every input with uniform resolution in a one-shot manner and make all of the predictions at once. However, human vision is an “active” process that not only actively switches from one focus point to another within the visual field, but also applies spatially varying attention centered at such focus points. Hence, we propose incorporating the bio-plausible mechanisms of foveation and saccades to transform the methods of machine perception. In particular, we present two object localization frameworks that incorporate foveation and saccades with a variable degree of supervision to provide improved performance and enable resiliency under automated annotation system scenarios. Next, we present a collection of works that stem out of such bio-plausible frameworks, allowing (a) simultaneous learning of visual semantics and visual syntax of an image, (b) robustness to adversarial perturbations, and (c) efficient object localization in videos.</p>



I.2 Christopher Kymn
UC Berkeley



Email: cjkymn@berkeley.edu
PI: Bruno A. Olshausen
Level: PhD student
Expected Graduation Date:
May 2025

Title: Efficient Visual Disentanglement with Convolutional Sparse Coding and Resonator Networks

Abstract: Disentanglement is a central problem for understanding visual scenes: examples include separating the effects of form from motion, and of lighting from surface reflectance. The resonator network is an HD Computing/VSA-based algorithm for performing disentanglement with distributed representations, with strong performance relative to standard methods. However, an unsolved problem is how to learn useful representations of visual scenes as input for the resonator. We propose a method for integrating resonator networks with the latent representations produced by sparse coding, a well-known unsupervised learning framework for signal representation. We show how this integration helps with the capacity limits of distributed representations and reduces collisions in the combinatorial search space. Conversely, we also show that the resonator network outperforms alternative methods used by similar generative models. Finally, we discuss how our proposal relates and contributes to neuroscientific theories of hierarchical inference in mammalian visual cortex.

I.3 Yinghan Long
Purdue University



Email: long273@purdue.edu
PI: Kaushik Roy
Level: PhD student
Expected Graduation Date:
May 2024

Title: Segmented Recurrent Transformer: An Efficient Sequence-to-Sequence Model

Abstract: Transformers have shown dominant performance across a range of domains including language and vision. However, their computational cost grows quadratically with the sequence length, making their usage prohibitive for resource-constrained applications. To counter this, our approach is to divide the whole sequence into segments and apply attention to the individual segments. We propose a segmented recurrent transformer (SRformer) that combines segmented (local) attention with recurrent attention. The loss caused by reducing the attention window length is compensated by aggregating information across segments with recurrent attention. SRformer leverages Recurrent Accumulate-and-Fire (RAF) neurons' inherent memory to update the cumulative product of keys and values. The segmented attention and lightweight RAF neurons ensure the efficiency of the proposed transformer. Such an approach leads to models with sequential processing capability at a lower computation/memory cost. We apply the proposed method to T5 and BART transformers. The modified models are tested on summarization datasets including CNN-dailymail, XSUM, ArXiv, and Media-SUM. Notably, using segmented inputs of varied sizes, the proposed model achieves 6–22% higher ROUGE1 scores than a segmented transformer and outperforms other recurrent transformer approaches. Furthermore, compared to full attention, the proposed model reduces the computational complexity of cross-attention by around 40%.



I.4 Quanling Zhao UC San Diego



Email: quzhao@ucsd.edu

PI: Tajana Rosing

Level: Bachelor's student

Expected Graduation Date:
2023

Title: Bridging the Gap between Hyperdimensional Computing and Kernel Methods via the Nyström Method

Abstract: Hyperdimensional computing (HDC) is an approach for solving cognitive information processing tasks using data represented as high-dimensional vectors. The technique has a rigorous mathematical backing and is easy to implement in energy-efficient and highly parallel hardware like FPGAs and "in-memory" architectures. The success of HDC-based machine learning approaches is heavily dependent on the mapping of raw data to high-dimensional space. In this work, we propose a new method for constructing this mapping that is based on the Nyström method from the literature on kernel approximation. Our approach provides a simple recipe to turn any user-defined positive-semidefinite similarity function into an equivalent mapping in HDC. There is a vast literature on the design of such functions for learning problems. Our approach provides a mechanism to import them into the HDC setting, expanding the types of problems that can be tackled using HDC (e.g. graph with attributes). An empirical comparison of our approach against existing HDC encoding methods on a variety of classification tasks shows that we can achieve 15%-40% and 3%-18% better classification accuracy on graph and string datasets respectively.

I.5 Yudi Xie MIT



Email: yu_xie@mit.edu

PI: James DiCarlo

Level: PhD student

Expected Graduation Date:
May 2026

Title: What's Behind? Unveiling Probabilistic Visual Inference in Humans and Primates

Abstract: Both humans and primates exhibit remarkable capabilities in performing complex visual inferences with astonishing speed and demonstrate a profound ability to generalize systematically. Previous studies in neuroscience have primarily focused on simple stimuli like gradings or oriented bars, leaving the understanding of the neural and cognitive mechanism of mid-to-high-level probabilistic visual computations in the brain unexplored. Here, we introduce "What's Behind," a novel task designed to prob probabilistic visual reasoning in humans and non-human primates. This task challenges experimental subjects to identify objects that are partially occluded, creating a scenario where the object's identity is highly ambiguous due to the occlusion. To explore the underlying cognitive and neural mechanisms, we propose several computational models implemented using convolutional neural networks and structured probabilistic programs. These models serve as hypotheses for the observed phenomena. Our future research will focus on how experimental subjects provide probability judgments and demonstrate systematic generalization in this task, employing these responses to distinguish between the various computational models. This approach aims to deepen our understanding of the complex processes involved in visual cognition and its neural basis, bridging a significant gap in current neuroscience research.



I.6 Onat Gungor
UC San Diego



Email: ogungor@ucsd.edu

PI: Tajana Rosing

Level: Post-doc

Title: Effective Defenses for Machine Learning-based Intrusion Detection Against Adversarial Attacks

Abstract: Due to increased inter-connectivity, the Internet of Things (IoT) has various security vulnerabilities. Machine learning-based intrusion detection system (ML-IDS) is an IoT security measure relying on ML models to detect malicious activity. However, these methods are susceptible to adversarial attacks. We propose two defenses for ML-IDS against adversarial attacks: robust layered defense, and adaptive adversarial training defense for hyperdimensional computing (HD). The former approach detects if a sample comes from an adversarial attack and eliminates the adversarial component. This solution improves model prediction performance by 114% and 50% with respect to no defense and the state-of-the-art adversarial training defense. The latter approach discovers the most effective adversarial attack from a set of adversarial attacks and includes it into our adaptive adversarial training. We could improve HD prediction performance by up to 145% and 35% with respect to no defense and the state-of-the-art adversarial training defense.

I.7 Amogh Joshi
Purdue University



Email: joshi157@purdue.edu

PI: Kaushik Roy

Level: PhD student

Expected Graduation Date:

December 2026

Title: Unraveling Hybrid Architectures for high-speed Event-based Object Tracking and Autonomous Navigation: A Neuro-symbolic Approach

Abstract: Hybrid architectures based on the fusion of traditional Analog Neural Networks (ANNs) and bio-inspired Spiking Neural Networks (SNNs) show promising potential by leveraging the complementary advantages of each. They offer highly efficient, lightweight yet capable systems for carrying out real-world sequential applications in resource-constrained edge systems. This poster delves into the effectiveness of hybrid SNN-ANN architectures within the context of a real-world autonomous navigation application. The scenario involves a cognitive system performing object detection/tracking and a planning algorithm to obtain the motion trajectory. Event-based sensors are used to obtain sparse and asynchronous streams of events at a high temporal resolution, crucial for tracking high-speed objects. A lightweight hybrid SNN-ANN architecture processes the event-stream dynamically, discerning objects based on their speed and direction of motion. Subsequently, a physics-guided neuro-symbolic planner handles the detected objects along with depth information, generating collision-free trajectories for the realization of autonomous navigation.



I.8 Vignesh Sundaresha

University of Illinois
Urbana-Champaign



Email: vs49@illinois.edu

PI: Naresh Shanbhag

Level: PhD student

Expected Graduation Date:

December 2027

Title: Growing Efficient Accurate and Robust Neural Networks on the Edge

Abstract: The ubiquitous deployment of deep networks on resource-constrained Edge devices is hindered by their high computational complexity and their fragility to out-of-distribution (OOD) data, especially to naturally occurring corruptions. Current solutions rely on the Cloud to train and compress these models before deploying them to the Edge. This incurs high energy and latency costs while also raising privacy concerns. We propose GEARnn for Growing Efficient, Accurate, and Robust neural networks in-situ (i.e., completely on the Edge device). Starting with a low-complexity initial backbone network, GEARnn employs One-Shot Growth (OSG) to grow a network satisfying the memory constraints of the Edge device, and then robustly trains the network using Efficient Robust Augmentation (ERA). We also employ Test-Time Adaptation (TTA) for continual fine-tuning of model parameters to handle variations in locally available data. We demonstrate results on NVIDIA Jetson Xavier-NX and analyze the trade-offs between accuracy, robustness, training complexity, and model size.

I.9 Shiting (Ginny) Xiao

Yale University



Email: ginny.xiao@yale.edu

PI: Priya Panda

Level: PhD student

Expected Graduation Date:

May 2028

Title: ReSpike: Efficient Residual Frame Video Perception with Spiking Neural Networks

Abstract: Modeling temporal dynamics in videos is essential for accurate motion capture. While current methods utilizing optical flows or 3D convolutions have been proven effective, they suffer from computational inefficiencies. To address this problem, we introduce ReSpike, an innovative framework combining Spiking Neural Networks (SNNs) and Artificial Neural Networks (ANNs) for video action recognition, where a spiking ResNet is used to efficiently capture the temporal dynamics. Utilizing a novel partitioning of video frames into Key and Residual frames, the framework leverages the strengths of ANNs in processing dense RGB information and that of SNNs in handling sparse, time-sensitive information. In addition, we propose a multi-modal fusion Transformer to integrate spatial features from ANN branch and temporal features from SNN branch using cross-domain attention. Extensive experiments on HMDB-51 and UCF-101 datasets validate that the proposed ReSpike network outperforms existing state-of-the-art methods in video action recognition, and achieves better accuracy-efficiency tradeoffs.



I.10 Flavio Ponzina
UC San Diego



Email: fponzina@ucsd.edu
PI: Tajana Rosing
Level: Post-doc

Title: Dynamic Ensembling for Dependable and Energy-Efficient Edge AI

Abstract: Ensemble learning is a meta-learning approach that combines the predictions of multiple learners, demonstrating improved accuracy and robustness performance. Nevertheless, ensembling models like Convolutional Neural Networks (CNNs) result in high memory and computing overheads that usually prevent the use of ensembles in embedded systems. When deployed in nature, these devices are equipped with small batteries that provide a power supply and usually include a limited number of solar cells to harvest energy from the environment. In this work, we first propose an improved ensembling method that outperforms previous works in terms of accuracy performance. Then, we leverage the multi-CNN structure of the designed ensemble to implement a novel model selection policy for edge AI computation in energy-harvesting AI systems. We show that our solution outperforms state-of-the-art (i) by improving system dependability for limited and dynamic energy budgets and (ii) by enabling concurrent on-device training and inference stages at the edge.

I.11 Xiaofan Yu
UC San Diego



Email: xlyu@ucsd.edu
PI: Tajana Rosing
Level: PhD student
Expected Graduation Date:
December 2024

Title: Lifelong Intelligence Beyond the Edge using Hyperdimensional Computing

Abstract: On-device learning has emerged as a prevailing trend that avoids the slow response time and costly communication in cloud-based learning. The ability to learn continuously and indefinitely in a changing environment, and with resource constraints, is critical for real sensor deployments. However, existing designs are inadequate for practical scenarios with (i) streaming data input, (ii) lack of supervision, and (iii) limited on-board resources. In this paper, we design and deploy the first on-device lifelong learning system called LifeHD for general IoT applications with limited supervision. LifeHD is designed based on a novel brain-inspired and lightweight learning paradigm called Hyperdimensional Computing (HDC). We utilize a two-tier associative memory organization to intelligently store and manage high-dimensional, low-precision vectors, which represent the historical patterns as cluster centroids. We additionally propose two variants of LifeHD to cope with scarce labeled inputs and power constraints. We implement LifeHD on off-the-shelf edge platforms and perform extensive evaluations across three scenarios. Our measurements show that LifeHD improves the quality of unsupervised clustering by up to 3.25x compared to the state-of-the-art NN-based unsupervised lifelong learning baselines with as much as 43.2x better energy efficiency.



I.12 Arijit Dasgupta
MIT



Email: arijitdg@mit.edu

PI: Vikash Mansinghka and
Joshua Tenenbaum

Level: PhD student

Expected Graduation Date:
December 2028

Title: Intuitive Physical Reasoning with Probabilistic Programs

Abstract: A key requisite to the deployment of vision-based commonsense machines is the building of general-purpose, interpretable models of 3D objects and their interactions. We introduce a hierarchical probabilistic program that encapsulates uncertainties over 3D object attributes and their causal interactions in dynamic scenes. This program synergistically combines renderer-in-the-loop inverse graphics for perception and approximate dynamics for intuitive physics, from which one can query facets of intuitive physical reasoning. We focus on the Violation-of-Expectation (VoE) paradigm for 3D scene plausibility. The program runs in tandem with a GPU-accelerated physics-informed sequential monte-carlo algorithm for zero-shot learning of object attributes and pose-tracking. We show that our approach works robustly on an established VoE benchmark. We also illustrate that our approach transcends existing computational VoE models by inferring posterior estimates of extrinsic (pose, shape) and intrinsic (mass, friction) object attributes. This paves a new standard for model interpretability and explainability in this paradigm while retaining robust performance.

I.13 Mathieu Huot
MIT



Emails: mhuot@mit.edu

PI: Josh Tenenbaum and Vikash
Mansinghka

Level: Post-doc

Title: Variational Inference By Automatic Differentiation of
Expected Values

Abstract: A common task in many fields of science is to optimize a function defined as an expected value. In Lew & Huot et al, the authors introduced ADEV, an extension to forward mode automatic differentiation which derives unbiased gradient estimators for loss functions defined as expected values, for an expressive class of probabilistic computations. We introduce a new reverse-mode algorithm for computing unbiased gradient estimates in the style of ADEV, and language automation for constructing variational loss objectives for an expressive trace-based PPL. Our new algorithm and language automation support the concise expression of several advanced variational algorithms, including importance-weighted autoencoders (IWAE), hierarchical variational inference (HVI), and reweighted wake-sleep (RWS) — including novel low variance gradient estimators which cannot be implemented in systems like Pyro. We provide an implementation in GenJAX, a GPU-accelerated probabilistic programming system, and obtain order of magnitude speed-ups in convergence over equivalent automation in Pyro.



I.14 Foroozan Karimzadeh
Georgia Institute of Technology



Email: fkarimzadeh6@gatech.edu
PI: Arijit Rauchowdhury
Level: Post-doc

Title: CIM-Aware Quantization for Energy Efficient Generative AI Models

Abstract: Large language models (LLMs) with generative capabilities have showcased outstanding performance across diverse applications. However, deploying these models on resource-constrained edge and mobile devices during inference mode is very challenging due to their massive computations. In this paper, we proposed a compute-in-memory (CIM) aware post-training quantization and sparsity technique to achieve superior energy efficiency compared to the traditional CMOS architecture. In CIM, multiplying to "0" bit requires less energy than "1." Therefore, we quantized the LLMs to the desired numbers where they have fewer ones in their binary representations than zero. We used a 2bit/cell resistive CIM test chip to evaluate our method. Since in the 2bit/cell RCIM E00<E01<E10<E11, we demonstrated that our method achieved 2.6x less energy than the baseline method, which eventually enabled running LLMs on edge devices.

I.15 Connor Bybee
UC Berkeley



Email: bybee@berkeley.edu
PI: Bruno Olshausen
Level: Post-doc

Title: Efficient Optimization with Higher-Order Ising Machines

Abstract: A prominent approach to solving combinatorial optimization problems on parallel hardware is Ising machines, i.e., hardware implementations of networks of interacting binary spin variables. Most Ising machines leverage second-order interactions although important classes of optimization problems, such as satisfiability problems, map more seamlessly to Ising networks with higher-order interactions. Here, we demonstrate that higher-order Ising machines can solve satisfiability problems more resource-efficiently in terms of the number of spin variables and their connections when compared to traditional second-order Ising machines. Further, our results show on a benchmark dataset of Boolean k-satisfiability problems that higher-order Ising machines implemented with coupled oscillators rapidly find solutions that are better than second-order Ising machines, thus, improving the current state-of-the-art for Ising machines.



I.16 Keming Fan
UC San Diego



Email: k4fan@ucsd.edu

PI: Tajana Rosing

Level: PhD student

Expected Graduation Date:
June 2027

Title: HDnn: Accelerating few-shot learning using Hyperdimensional Computing with RRAM

Abstract: Few-shot learning (FSL) is a machine learning algorithm that rapidly trains classification models with just a few samples (1-10 per class). However, existing FSL classifiers are either computationally expensive or lack accuracy. In this work, we introduce the first fabricated FSL-HD accelerator, a fabricated FSL chip based on hyperdimensional computing (HDC). HDC is a lightweight, brain-inspired algorithm that is hardware-friendly, by using highly parallel simple operations. It achieves good accuracy in FSL without the need to retrain feature extractor models. This chip supports feature extraction, HD encoding, training, and inference, achieving superior accuracy and broader features compared to the state-of-the-art FSL accelerator, SAPIENS [T-ED'21]; 1000x energy efficiency improvement than CHIMERA [JSSC'22]. Furthermore, we extend FSL-HD by enabling in-memory inference using analog computing in emerging non-volatile memory cell RRAM). The custom analog circuit supports high-density storage of encoded data, and can efficiently perform in-memory matrix-vector multiplication. This approach ensures high memory density for optimal on-chip storage of encoded hypervectors and is 3x more energy-efficient and 4x latency improvement than near-memory digital implementations.

I.17 Jeffrey Yu
Stanford University



Email: jeffreyy@stanford.edu

PI: Priyanka Raina

Level: PhD student

Expected Graduation Date:
May 2027

Title: 8-bit Transformer Inference and Fine-tuning for Edge Accelerators

Abstract: Transformer models excel in natural language processing and vision tasks but require substantial computational and memory resources, posing challenges for edge accelerators. Quantization, particularly to lower precision types like 8-bit integer (int8) and 8-bit floating-point (FP8), can mitigate these demands. However, int8's limited precision hinders its use in training, whereas prior FP8 applications only partially quantized Transformer operations. This study explores Transformer inference and fine-tuning at the edge using two 8-bit types: FP8 and Posit8, an 8-bit posit format with variable exponent and fraction lengths, offering higher precision around the value 1 and suitability for storing Transformer weights and activations. Our comprehensive approach quantizes all operations in both forward and backward passes, not just matrix multiplications. Key contributions include the following: (1) Transformer inference in FP8 and Posit8 with less than 1% accuracy loss compared to bfloat16, using operation fusion and no scaling factors. (2) Transformer fine-tuning in 8 bits via adapted low-rank adaptation (LoRA) for Posit8 and FP8, enhancing GEMM operations efficiency and reducing memory use. (3) An efficient posit softmax design, significantly smaller and less power-intensive than bfloat16 equivalents, yet maintaining accuracy. Overall, Posit8 and FP8 achieve comparable accuracy to bfloat16 for inference and fine-tuning, with substantial reductions in area and energy use, and a threefold decrease in fine-tuning memory requirements.



I.18 Jesse Michel

MIT



Email: jmmichel@csail.mit.edu

PI: Michael Carbin

Level: PhD student

Expected Graduation Date:

May 2025

Title: Distributions for Compositionally Differentiating Parametric Discontinuities

Abstract: Computations in physical simulation, computer graphics, and probabilistic inference often require the differentiation of discontinuous processes due to contact, occlusion, and the switching of a controller on/off. Popular differentiable programming languages, such as PyTorch and JAX, do not support the differentiation of these processes. We introduce a differentiable programming language, Potto, that is the first, first-order language to support differentiation of parametric discontinuities (conditionals containing one or more real-valued variables of integration and parameters in the condition). We present denotational semantics for programs and for program derivatives and show the two accord. From this, we describe the implementation of Potto, which enables separate compilation of programs. Our implementation of Potto overcomes previous compile-time bottlenecks. We showcase the features of Potto by implementing a prototype differentiable renderer with separately compiled shaders..

I.19 Tiffany Luong &

Yoni Friedman

MIT



Email: tluong@mit.edu and
yf@mit.edu

PI: Vikash Mansinghka

Level: Research Scientist and
PhD student, respectively

Title: Towards Robust Computer Vision –
A Benchmark of Human Perception in Challenging Visual Conditions

Abstract: Human perception can recover 4D information from a wide range of environments, often in spite of poor visual conditions (rain, snow, poor lighting, etc...). Similarly, we can make rapid inferences about object categories, agents, and actions by observing just a handful of key points. This robust, efficient, and generalizable perceptual mechanism presents a basic challenge to modern computer vision systems, which tend to be data-inefficient, and brittle. Here, we present a series of stimuli and psychophysics experiments designed to probe those perceptual abilities in humans, and illustrate shortcomings in standard approaches to computer vision. Using a class of Just-Noticeable-Difference tasks on a range of abstract, challenging, and noisy visual conditions, our work illustrates the range of human perceptual abilities and presents a strong benchmark for more efficient, reliable models of computer vision.



I.20 Tian Jin

MIT



Email: tianjin@csail.mit.edu

PI: Michael Carbin

Level: PhD student

Expected Graduation Date:

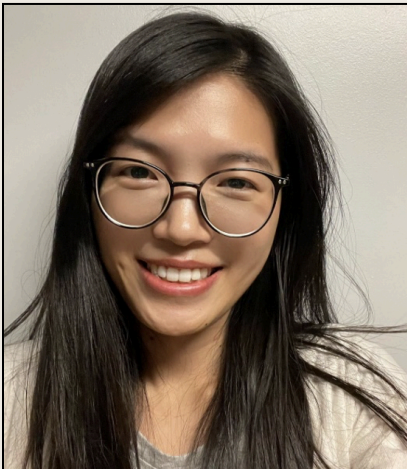
May 2026

Title: The Cost of Down-Scaling Language Models: Fact Recall Deteriorates Before In-Context Learning

Abstract: How does scaling the number of parameters in large language models (LLMs) affect their core capabilities? We study two natural scaling techniques — weight pruning and simply training a smaller or larger model, the latter of which we refer to as dense scaling — and their effects on two core capabilities of LLMs: (a) recalling facts presented during pre-training and (b) processing information presented in-context during inference. By curating a suite of tasks that help disentangle these two capabilities, we find a striking difference in how these two capabilities evolve due to scaling. Reducing the model size by more than 30% (via either scaling approach) significantly decreases the ability to recall facts seen in pre-training. Yet, a 60--70% reduction largely preserves the various ways the model can process in-context information, ranging from retrieving answers from a long context to learning parameterized functions from in-context exemplars.

I.21 Guangyu Jiang

Kennesaw State University



Email:

gjiang@students.kennesaw.edu

PI: Yan Fang

Level: PhD student

Expected Graduation Date:

August 2025

Title: Visual Sensor Fusion for Neuromorphic SLAM

Abstract: For autonomous robotics, Simultaneous Localization and Mapping (SLAM) is a challenge due to the limited computing resources of edge computers. Recent works have made progress through the integration of event-based cameras. In this work, we aim to explore the synergy of dynamic vision sensors (DVS) and regular cameras that facilitate a neuromorphic SLAM framework, inspired by the RatSLAM framework. Our motivation is to address the issue of RatSLAM encountered in scenarios with similar landmarks or visual data with low quality and to enhance the visual odometry performance. The proposed framework incorporates two cameras in a sensor fusion style and leverages both deep neural networks and spiking neural networks for visual processing. We expect the proposed framework can improve the accuracy and energy efficiency of neuromorphic SLAM tasks.



THEME 2 SESSION OVERVIEW

Tuesday, March 12 from 3:00-4:00 PM

POSTER NO.	PRESENTER	TITLE
2.1	Zishen Wan, Ritik Raj, and Hanchen Yang	CogSys: Efficient and Scalable Neuro-Vector-Symbolic Cognition System via Algorithm-Hardware Co-Design
2.2	Zhenyu Wang	Analytical Modeling and Electro-Thermal Benchmarking of Heterogenous Integration with 2.5D/3D for Cognitive Computing
2.3	Amrit Nagarajan	TokenDrop + BucketSampler: Towards Efficient Padding-free Fine-tuning of Language Models
2.4	Surya Selvam	Efficient Batched Inference in Conditional Neural Networks
2.5	Yuan Liao	A 28nm Scalable and Flexible Accelerator for Universal Sparse Transformer Models
2.6	Soonha Hwang	Benchmarking of In-Memory Computing (IMC) Architectures
2.7	Kaining Zhou	IMCsim: An In-Memory Computing Simulation Framework for Deep Learning Workloads
2.8	Jian Meng	Torch2Chip: An End-to-end Customizable Deep Neural Network Compression and Deployment Toolkit for Prototype Hardware Accelerator Design
2.9	Sixu Li	Easy-3D: A 1.47W, 591 million points per second 3D reconstruction accelerator with End-to-End acceleration, Optimized Sampling and Reconfigurable Renderer
2.10	Yonggan Fu	Auto-CARD: Efficient and Robust Codec Avatar Driving for Real-time Mobile Telepresence
2.11	Haoran You	ShiftAddViT: Mixture of Multiplication Primitives Towards Efficient Vision Transformer
2.12	Yongan Zhang	GPT4AIGChip: Towards Next-Generation AI Accelerator Design Automation via Large Language Models





2.13	Nealson Li	E-Track: Eye Tracking with Event Camera for Extended Reality (XR) Applications
2.14	Abhimanyu Bambhaniya	Algorithmic optimization and system design of LLMs
2.15	Akshat Ramachandran	Algorithm-Hardware Co-Design of Distribution-Aware Logarithmic-Posit Encodings for Efficient DNN Inference
2.16	Jamin Seo	Machine Learning System for XR Workloads - Benchmark Suite and Accelerator
2.17	Flavio Ponzina	Dynamic Ensembling for Dependable and Energy-Efficient Edge AI
2.18	Sai Saketika Chekuri	Co-Designing AR/VR Workloads and Neural Network Accelerators
2.19	Reena Elangovan	VCQ: Vector Clustered Quantization for 4-bit LLM inference
2.20	Hao Kang	Reducing Memory Foot Print in Large Language Models through Low Rank Approximation
2.21	Jingbo Sun	Adaptive Graph Learning for Efficient Thermal Analysis of the Chiplet System under Interface Variations
2.22	Amey Agrawal	Vidur: A Large-Scale Simulation Framework for LLM Inference



THEME 2: HARDWARE ALGORITHM CO-DESIGN

PRESENTATION DETAILS

Tuesday, March 12 from 3:00-4:00 PM

SCHOLAR	POSTER DETAILS
<p>2.1 Zishen Wan Georgia Institute of Technology</p>  <p>Ritik Raj Georgia Institute of Technology</p>  <p>Hanchen Yang Georgia Institute of Technology</p>  <p>Emails: zishenwan@gatech.edu, ritik.raj@gatech.edu, and hanchen@gatech.edu</p> <p>PIs: Arijit Raychowdhury and Tushar Krishna</p> <p>Level: PhD students</p> <p>Expected Graduation Date: August 2025, May 2028, and December 2027, respectively</p>	<p>Title: CogSys: Efficient and Scalable Neuro-Vector-Symbolic Cognition System via Algorithm-Hardware Co-Design</p> <p>Abstract: Raven's progressive matrices (RPM) is widely used as a standardized test for logical thinking and abstract reasoning evaluation. State-of-the-art research in neurosymbolic AI, especially with neuro-vector-symbolic architecture (NVSA), has demonstrated encouraging results on solving RPM tests. Despite their promise, executing NVSA on current computing systems comes with significant latency and memory overheads due to their holographic vector computation, preventing NVSA from being efficient and scalable. This work proposes CogSys, an algorithm-hardware co-design framework dedicated to neurosymbolic AI acceleration, aiming to win both reasoning efficiency and cognition scalability. On the algorithm side, CogSys implements an efficient factorization technique with a multi-level parallelism strategy to alleviate compute and memory overhead in vector-symbolic operations. On the hardware side, CogSys proposes an architecture dedicated to further accelerating NVSA to maximize bandwidth efficiency and data reuse opportunities by proposing reconfigurable and scalable neuro/symbolic compute arrays with an adaptive folding strategy. CogSys enables real-time abduction reasoning towards human fluid intelligence, requiring only 0.3 seconds per spatial-temporal reasoning task with 4mm² area and 1.18 W power consumption.</p>



2.2 Zhenyu Wang

Arizona State University



Email: zwang586@asu.edu

PI: Yu (Kevin) Cao

Level: PhD student

Expected Graduation Date:
July 2024

Title: Analytical Modeling and Electro-Thermal Benchmarking of Heterogenous Integration with 2.5D/3D for Cognitive Computing

Abstract: Monolithic designs face significant fabrication costs and data movement challenges, especially when dealing with complex and diverse models for cognitive computing. Advanced 2.5D/3D packaging promises high bandwidth and density to overcome these challenges but also introduces new electro-thermal constraints. This work presents a suite of analytical performance models to enable efficient benchmarking of a 2.5D/3D system for cognitive computing. These models encompass various performance metrics related to computing units, network-on-chip, and network-on-package. The results are summarized into a new tool, HISIM, which is 104 – 106 faster than state-of-the-art AI benchmark tools. The simulator, HISIM, enables us to evaluate the potential of 2.5D/3D heterogeneous integration on representative AI algorithms under thermal constraints.

2.3 Amrit Nagarajan

Purdue University



Email: nagaraj9@purdue.edu

PI: Anand Raghunathan

Level: PhD student

Expected Graduation Date:
December 2023

Title: TokenDrop + BucketSampler: Towards Efficient Padding-free Fine-tuning of Language Models

Abstract: The great success of Language Models (LMs) for various Natural Language Processing (NLP) tasks is accompanied by computational challenges during both pre-training and fine-tuning. During fine-tuning, the presence of variable-length input sequences necessitates the use of padding tokens when batching sequences. These padding tokens lead to ineffectual computations, adversely impacting the efficiency of fine-tuning. We also observe that LMs memorize the limited task-specific training data despite the use of known regularization methods. Based on these insights, we present TokenDrop + BucketSampler, a framework that simultaneously improves the efficiency and accuracy of LM fine-tuning. BucketSampler generates batches of samples with lower variance in sequence lengths to reduce the number of padding tokens. TokenDrop is a new regularizer that prunes a random subset of insignificant tokens from each input sequence in every epoch to prevent overfitting. TokenDrop drops more tokens from the longer sequences in each batch to further reduce variance in input lengths and the need for padding. TokenDrop + BucketSampler accelerates fine-tuning on diverse downstream tasks by up to 10.61X, while also producing models that are up to 1.17% more accurate compared to conventional fine-tuning.



2.4 Surya Selvam

Purdue University



Email: selvams@purdue.edu

PI: Anand Raghunathan

Level: PhD student

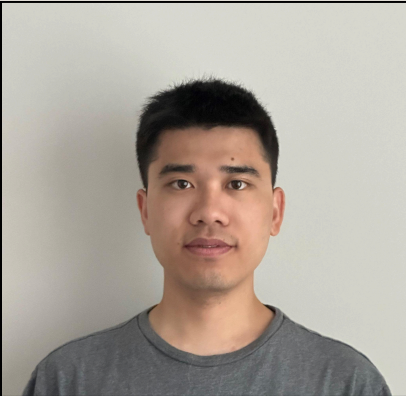
Expected Graduation Date:
May 2025

Title: Efficient Batched Inference in Conditional Neural Networks

Abstract: Conditional Neural Networks (NNs), which modulate computational effort on an input-by-input basis, have gained popularity in recent years. However, batching, a technique widely used to improve hardware utilization and throughput during NN inference, becomes challenging in conditional NNs due to the varying computational footprints across inputs in the batch. We propose BatchCond, an optimized batching framework for conditional NNs that consists of two key steps: (i) Computational Similarity Driven Batching, which batches inputs that are likely to share similar computational patterns, and (ii) Hardware-aware Batch Reorganization, which addresses residual computational irregularity by dynamically splitting batches into computationally-similar sub-batches in order to improve performance. BatchCond improves the throughput of batched inference by up to 6.6x (mean of 2.4x) on conditional NNs including early exit, slimmable and auto-regressive transformer NNs.

2.5 Yuan Liao

Cornell Tech



Email: yl3662@cornell.edu

PI: Jae-sun Seo

Level: PhD student

Expected Graduation Date:
May 2026

Title: A 28nm Scalable and Flexible Accelerator for Universal Sparse Transformer Models

Abstract: Recently, transformer-based models have prospered in natural language processing and computer vision. However, the current hardware designs for the Transformer-based models have yet to explore the potential of software-hardware co-design fully. Four difficulties need to be addressed in the software-hardware co-design of the accelerator for Transformer-based models: scalability, flexibility, sparsity handling, and non-linear function calculations. In this work, we propose the row-wise matrix multiplication processing elements (RMMPE) and the post-PE processors (PPE). RMMPE computes matrix multiplication in row-wise products, which increases data reuse opportunity and improves scalability and flexibility. RMMPE can also efficiently handle various dimensional, unstructured sparse matrix multiplication. PPE computes complex functions in linear approximation. The experiment results indicate 1.27 TOPS throughput and 1.45 TOPS/W energy efficiency, among the most efficient Transformer accelerators.



2.6 Soonha Hwang

University of Illinois
Urbana-Champaign



Email: soonhah2@illinois.edu

PI: Naresh Shanbhag

Level: PhD student

Expected Graduation Date:
May 2028

Title: Benchmarking of In-Memory Computing (IMC) Architectures

Abstract: We present a 2023 update on our on-going IMC benchmarking activity. The purpose of this activity is to enable the IMC design community within COCOSYS and elsewhere to evaluate trends in this important area. This benchmarking activity involves extracting more than 30 metrics from IC prototypes of IMC and digital accelerators published in major circuit conferences such as ISSCC, VLSI, CICC, and ESSCIRC since 2019. It further involves analysis of these metrics to extract trends. A GitHub database: <https://github.com/UIUC-IMC/UIUC-IMC-Benchmarking/> has been made available for users to download both the database and accompanying Python scripts for generating plots. An analysis of the 2023 update reveals: 1) SRAM-based IMCs continue to maintain superiority in terms of energy efficiency and compute density at the bank level, but the gap with digital accelerators has significantly narrowed from 10x to 2.5x; 2) eNVM-based IMCs still lag behind their SRAM-based counterparts in both energy efficiency and compute density; 3) digital IMC (DIMCs) designs have begun to proliferate due to its scalability and accuracy advantages over analog IMCs; and 4) IMC publications continue to underreport critical bank-level metrics such as accuracy and the trade-off between energy and accuracy.

2.7 Kaining Zhou

University of Illinois
Urbana-Champaign



Email: kainingz@illinois.edu

PI: Naresh Shanbhag

Level: PhD student

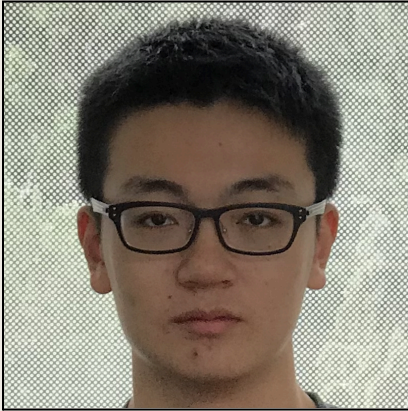
Expected Graduation Date:
May 2027

Title: IMCsim: An In-Memory Computing Simulation Framework for Deep Learning Workloads

Abstract: In-memory computing (IMC) has been proven to be an effective approach to overcoming well-known memory bottlenecks. To facilitate IMC development, prior studies have built several specific IMC simulators for small neural network workloads. However, as we scale deep learning (DL) towards larger models, new scalability challenges with IMC-based designs need to be overcome. Hereby, we present a generic full-system IMC simulation framework named IMCsim by directly integrating the software run-time libraries for DNN models into the IMC device simulator. IMCsim introduces a new set of ISA extensions to express the common tensor operators in the IMC simulator. By offering the flexibility of mapping tensor operators to different IMC architectures, IMCsim enables users to explore the trade-offs between performance, energy, and computation accuracy for different IMC design choices. To exhibit the functionality, efficiency, and value of IMCsim, we conduct case studies with LLaMA and report our study results in the evaluation.



2.8 Jian Meng
Cornell Tech



Email: jm2787@cornell.edu

PI: Jae-sun Seo

Level: PhD student

Expected Graduation Date:
June 2025

Title: Torch2Chip: An End-to-end Customizable Deep Neural Network Compression and Deployment Toolkit for Prototype Hardware Accelerator Design

Abstract: The development of model compression is motivated by the evolution of various neural network accelerator designs. On the algorithm side, the ultimate goal of quantization or pruning is accelerating the expensive DNN computations on hardware. Although the state-of-the-art quantization algorithm can achieve ultra-low precision with negligible accuracy degradation, the deep learning framework can only support non-customizable data format and extraction workflow. Secondly, the current SoTA algorithm treats the quantized integer as an intermediate result, while the final output of the quantizer is the “discretized” floating-point values. Finally, the industry-standard toolkits are constrained to their in-house product. The limited degree of freedom in the current toolkit hinders the prototype accelerator design. In this work, we propose Torch2Chip, an open-sourced, customizable, and high-performance toolkit that supports user-designed compression followed by automatic model fusion and parameter extraction. The user-customized compression algorithm will be directly packed into the deployment-ready format for prototype verification.

2.9 Sixu Li
Georgia Institute of Technology



Email: sli941@gatech.edu

PI: Yingyan (Celine) Lin

Level: PhD student

Expected Graduation Date:
May 2028

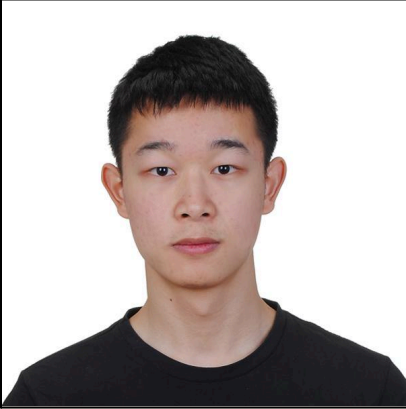
Title: Easy-3D: A 1.47W, 591 million points per second 3D reconstruction accelerator with End-to-End acceleration, Optimized Sampling and Reconfigurable Renderer

Abstract: Recent breakthroughs in Neural Radiance Field (NeRF) based 3D reconstruction and rendering have spurred the possibility of immersive experiences in augmented and virtual reality (AR/VR). However, current NeRF acceleration techniques are still inadequate for real-world AR/VR applications due to: the lack of end-to-end pipeline acceleration support, causing impractical DRAM bandwidth demands for edge devices. To tackle this limitation, we have developed Easy-3D, an end-to-end, 3D reconstruction accelerator capable of instant scene reconstruction and real-time rendering. A 28nm CMOS-based chip has been taped out to validate the proposed accelerator architecture and a simulator is built upon the chip. Results show that the proposed accelerator can achieve 2.5× and 6× improvement in training and inference, respectively, compared to the latest accelerators.



2.10 Yonggan Fu

Georgia Institute of Technology



Email: yfu314@gatech.edu

PI: Yingyan (Celine) Lin

Level: PhD student

Expected Graduation Date:
May 2025

Title: Auto-CARD: Efficient and Robust Codec Avatar Driving for Real-time Mobile Telepresence

Abstract: Real-time and robust photorealistic avatars for telepresence in AR/VR have been highly desired for enabling immersive photorealistic telepresence. However, there still exists one key bottleneck: the considerable computational expense needed to accurately infer facial expressions captured from headset-mounted cameras with a quality level that can match the realism of the avatar's human appearance. To this end, we propose a framework called Auto-CARD, which for the first time enables real-time and robust driving of Codec Avatars when exclusively using merely on-device computing resources. This is achieved by minimizing two sources of redundancy. First, we develop a dedicated neural architecture search technique called AVE-NAS for avatar encoding in AR/VR, which explicitly boosts both the searched architectures' robustness in the presence of extreme facial expressions and hardware friendliness on fast-evolving AR/VR headsets. Second, we leverage the temporal redundancy in consecutively captured images during continuous rendering and develop a mechanism dubbed LATEX to skip the computation of redundant frames. Specifically, we first identify an opportunity from the linearity of the latent space derived by the avatar decoder and then propose to perform adaptive latent extrapolation for redundant frames. For evaluation, we demonstrate the efficacy of our Auto-CARD framework in real-time Codec Avatar driving settings, where we achieve a 5.05x speed-up on Meta Quest 2 while maintaining a comparable or even better animation quality than state-of-the-art avatar encoder designs.

2.11 Haoran You

Georgia Institute of Technology



Email: haoran.you@gatech.edu

PI: Yingyan (Celine) Lin

Level: PhD student

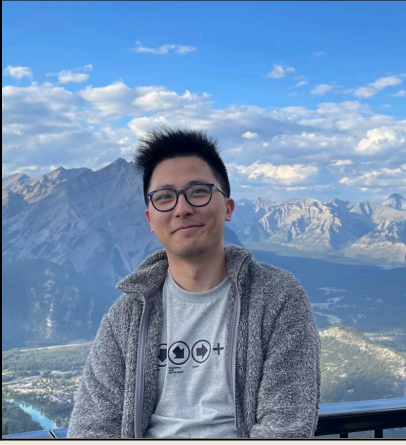
Expected Graduation Date:
May 2025

Title: ShiftAddViT: Mixture of Multiplication Primitives Towards Efficient Vision Transformer

Abstract: Vision Transformers (ViTs) excel in various vision tasks but suffer from inefficiencies in attention mechanisms and multi-layer perceptrons (MLPs), due to dense multiplications. Our solution, ShiftAddViT, reparameterizes pre-trained ViTs using simpler operations like bitwise shifts and additions. This approach creates a multiplication-reduced model, enhancing GPU inference speed without retraining from scratch. We reparameterize MatMuls in attention using additive kernels after converting queries and keys into binary codes in Hamming space. The remaining MLPs are reparameterized with shift kernels, optimized for GPUs using TVM. This maintains attention accuracy but slightly reduces MLP accuracy. To counterbalance, we introduce a mixture of experts (MoE) framework for MLP reparameterization, balancing multiplication, and its primitives. A new latency-aware loss in this framework helps assign input tokens dynamically, optimizing for latency. ShiftAddViT demonstrates up to $5.18\times$ faster GPU latency and 42.9% energy savings, with comparable accuracy to standard efficient ViTs in diverse 2D/3D Transformer-based vision tasks.



2.12 Yongan Zhang
Georgia Institute of Technology



Email: yzhang919@gatech.edu
PI: Yingyan (Celine) Lin
Level: PhD student
Expected Graduation Date:
May 2025

Title: GPT4AIGChip: Towards Next-Generation AI Accelerator Design Automation via Large Language Models

Abstract: We develop GPT4AIGChip, a framework intended to democratize AI accelerator design by leveraging human natural languages instead of domain-specific languages. Specifically, we first perform an in-depth investigation into LLMs' limitations and capabilities for AI accelerator design, thus aiding our understanding of our current position and garnering insights into LLM-powered automated AI accelerator design. Furthermore, drawing inspiration from the above insights, we develop a framework called GPT4AIGChip, which features an automated demo-augmented prompt-generation pipeline utilizing in-context learning to guide LLMs toward creating high-quality AI accelerator design. To our knowledge, this work is the first to demonstrate an effective pipeline for LLM-powered automated AI accelerator generation. Accordingly, we anticipate that our insights and framework can serve as a catalyst for innovations in next-generation LLM-powered design automation tools.

2.13 Nealson Li
Georgia Institute of Technology



Email: nealson@gatech.edu
PI: Arijit Raychowdhury
Level: PhD student
Expected Graduation Date:
May 2025

Title: E-Track: Eye Tracking with Event Camera for Extended Reality (XR) Applications

Abstract: Eye tracking is essential to enable extended reality (XR) applications on headsets with tight latency and power constraints. Unlike RGB cameras, the event camera generates asynchronous sparse events with high temporal resolution, which are suitable characteristics for eye tracking in XR systems. Nonetheless, processing an event-based data stream is a challenging task. In this demonstration, we present an event-based eye-tracking system that extracts pupil features, which is the first system that operates only with an event camera without additional sensing hardware. First, the event-to-frame conversion method encodes the events triggered by eye motion into a 3-channel frame. Secondly, a convolutional neural network (CNN) is trained on 24 subjects to classify the events representing the pupil. Finally, the region of interest mechanism reduces 96% of CNN inference. Our system locates the pupil with an error of 3.68 pixels in real time and preserves the sparse and asynchronous nature of event-based data stream.



2.14 Abhimanyu Bambhaniya

Georgia Institute of Technology



Email:

abambhaniya3@gatech.edu

PI: Tushar Krishna

Level: PhD student

Expected Graduation Date:

May 2026

Title: Algorithmic optimization and system design of LLMs

Abstract: LLMs have become an integral part of the Modern eco-system. The exponential growth scale poses many exciting challenges. My research delves into advancing Large Language Models (LLMs) inference through algorithmic optimization and system design. A key focus lies in integrating structured sparsity into attention-based models and developing tools to elucidate hardware trade-offs at different sparsity levels. Using roofline-based analytical tools, design-space exploration helps us understand and develop efficient hardware for LLM inference. This multifaceted approach aims to enhance the efficiency and performance of LLMs by optimizing algorithms, incorporating structured sparsity, and systematically exploring design spaces informed by hardware considerations.

2.15 Akshat Ramachandran

Georgia Institute of Technology



Email: akshat.r@gatech.edu

PI: Tushar Krishna

Level: Master's student

Expected Graduation Date:

August 2028

Title: Algorithm-Hardware Co-Design of Distribution-Aware Logarithmic-Posit Encodings for Efficient DNN Inference

Abstract: Traditional Deep Neural Network (DNN) quantization methods using integer, fixed-point, or floating-point data types struggle to capture diverse DNN parameter distributions at low precision, and often require large silicon overhead and intensive quantization-aware training. We introduce Logarithmic Posits (LP), an adaptive, hardware-friendly data type inspired by posits that dynamically adapts to DNN weight/activation distributions by parameterizing LP bit fields. We develop a novel algorithm-hardware co-design framework involving an automated framework to identify optimal layer-wise LP parameters and a unified mixed-precision accelerator (LPA) incorporating LP in the computational datapath. The co-design framework demonstrates a minimal drop in top-1 accuracy across various models, coupled with significant improvements in performance per unit area and enhanced energy efficiency compared to state-of-the-art quantization accelerators using different data types.



2.16 Jamin Seo

Georgia Institute of Technology



Email: jseo89@gatech.edu

PI: Tushar Krishna

Level: PhD student

Expected Graduation Date:

May 2026

Title: Machine Learning System for XR Workloads - Benchmark Suite and Accelerator

Abstract: Extended reality(XR) applications deploy heterogeneous machine learning(ML) models for diverse tasks and modalities. Designing efficient hardware systems for such multi-task multi-model(MTMM) workloads is challenging due to unique characteristics: (1)concurrent or cascading execution of multiple models, grouped per different XR usage scenarios; (2)dynamic task graphs from user-context interaction; (3)processing tied to input sources with specific frame per second(fps). To address the lack of prior exploration, we contribute to two threads. First, we developed “XR Bench”, an open-source benchmark suite for ML-based XR applications. It features industry-inspired usage scenarios with dynamic dependencies and fps requirements and proposes a comprehensive scoring metric for system evaluation. Second, drawing insights from the benchmark, we explore hardware acceleration methodology, focusing on mapping flexibility. We investigate the effectiveness of providing partial flexibility in mapping support, to reduce hardware overhead while maintaining real-time performance comparable to a fully flexible accelerator.

2.17 Flavio Ponzina

UC San Diego



Email: fponzina@ucsd.edu

PI: Tajana Rosing

Level: Post-doc

Title: MicroHD: An Accuracy-Driven Optimization of Hyperdimensional Computing Algorithms for TinyML systems

Abstract: Hyperdimensional computing (HDC) is emerging as a promising AI approach that can effectively target TinyML applications thanks to its lightweight computing and memory requirements. Previous works on HDC showed that limiting the standard 10k dimensions of the hyperdimensional space to much lower values is possible, reducing even more HDC resource requirements. Similarly, other studies demonstrated that binary values can be used as elements of the generated hypervectors, leading to significant efficiency gains at the cost of some degree of accuracy degradation. Nevertheless, current optimization attempts do not concurrently co-optimize HDC hyper-parameters, and accuracy degradation is not directly controlled, resulting in sub-optimal HDC models providing several applications with unacceptable output qualities.

In this work, we propose MicroHD, a novel accuracy-driven HDC optimization approach that iteratively tunes HDC hyper-parameters, reducing memory and computing requirements while ensuring user-defined accuracy levels. The proposed method can be applied to HDC implementations using different encoding functions, demonstrates good scalability for larger HDC workloads, and achieves compression and efficiency gains up to 200x when compared to baseline implementations for accuracy degradations lower than 1%.



2.18 Sai Saketika Chekuri
Stanford University



Email: saketika@stanford.edu
PI: Priyanka Raina
Level: Master's student
Expected Graduation Date:
June 2025

Title: Co-Designing AR/VR Workloads and Neural Network Accelerators

Abstract: Emerging AR-VR applications employ complex deep-learning models, and dedicated hardware accelerators tailored for these applications are increasingly becoming popular. Traditional approaches either tailor neural architectures for fixed hardware or customize hardware for specific neural models. In contrast, Hardware/Software Co-Design involves iteratively designing the hardware system as well as optimizing software algorithms to jointly optimize the accuracy of the models, hardware performance, and utilization. This poster presents the results of running various AR-VR tasks on an SoC. Specifically, the evaluation focuses on the performance of both the hardware and deep learning models, examining the impact of adjusting parameters in the original deep learning algorithms to better align with the hardware and vice versa. This work summarizes the impact of this optimization via metrics such as utilization, RTL runtime, and energy.

2.19 Reena Elangovan
Purdue University



Email: elangovr@purdue.edu
PI: Anand Raghunathan
Level: PhD student
Expected Graduation Date:
February 2024

Title: VCQ: Vector Clustered Quantization for 4-bit LLM inference

Abstract: Fine-grained quantization, where the weights and activations in a model are quantized at a per-vector granularity, has emerged as an effective technique for sub-8-bit quantization of LLMs. Recent per-vector quantization proposals that associate unique scale factors to each vector, and further quantize the scale factors themselves, represent the current state-of-the-art for post-training quantization (PTQ) of LLMs. To advance the state-of-the-art, we propose a new quantization method called vector clustered quantization (VCQ) that clusters the operand vectors and creates a dedicated quantizer for each vector cluster, going beyond simply associating a scale factor to each vector. As a specific embodiment of VCQ, we propose `\ovcq`, a PTQ algorithm that iteratively clusters operand vectors and quantizes each vector cluster with a Lloyd-Max based optimal quantizer. We show `\ovcq` achieves a local minima of quantization MSE for a given bitwidth. We demonstrate that `\ovcq` achieves superior bitwidth-vs-perplexity compared to previous state-of-the-art methods. During 4-bit post-training quantization (PTQ) of both weights and activations, `\ovcq` limits perplexity degradation to within 0.8 across GPT3 models on the Wikitext-103 dataset and degradation in F1 score to within 0.6 across BERT models on the SQuADv1.1 dataset compared to an unquantized baseline.



2.20 Hao Kang

Georgia Institute of Technology



Email: hkang342@gatech.edu

PI: Tushar Krishna

Level: PhD student

Expected Graduation Date:
August 2028

Title: Reducing Memory Foot Print in Large Language Models through Low-Rank Approximation

Abstract: Large Language Models (LLMs) have become increasingly powerful in natural language processing tasks. However, the extensive memory requirements of these models pose challenges for their deployment in practical applications, particularly in resource-constrained environments. In this paper, we propose an efficient approach to reduce the memory footprint of LLMs through low-rank approximation. Our research focuses on developing a low-rank approximation technique specifically tailored for LLMs. By leveraging the inherent structure and redundancy in the model, we aim to significantly reduce its memory consumption without compromising its inference performance. Through extensive experimentation and evaluation, we demonstrate the effectiveness of our approach in efficiently reducing the memory footprint of LLMs. Furthermore, we present a thorough analysis of the trade-offs between memory reduction and inference accuracy, highlighting the potential benefits of our proposed method. Our findings indicate that by applying low-rank approximation, substantial memory savings can be achieved without significantly sacrificing the overall performance of LLMs. This paves the way for the deployment of LLMs in resource-constrained systems, opening up new avenues for leveraging their capabilities in real-world applications.

2.21 Jingbo Sun

Arizona State University



Email: jsun127@asu.edu

PI: Yu (Kevin) Cao

Level: PhD student

Expected Graduation Date:
May 2024

Title: Adaptive Graph Learning for Efficient Thermal Analysis of the Chiplet System under Interface Variations

Abstract: In thermal analysis of a chiplet system, conventional numerical methods or machine learning-based surrogate models face tremendous challenges in computation cost and accuracy, especially in the presence of process and material variations. We propose Graph Neural Networks (GNNs) as a mathematical framework for efficient and robust thermal analysis with composite materials. By modeling each region and their thermal interactions as a graph, we continually adapt the GNN model under thermal interface variations. We validate our approach with numerical solutions and real thermal images from a crossbar unit and demonstrate its speedup and accuracy in a 2.5D chiplet system.



2.22 Amey Agrawal

Georgia Institute of Technology



Email:

agrawalamey12@gmail.com

PI: Alexey Tumanov

Level: PhD student

Expected Graduation Date:

May 2026

Title: Vidur: A Large-Scale Simulation Framework for LLM Inference

Abstract: Large language models (LLMs) are rapidly adopted for their ability to perform tasks that require human-like skills. However, LLM inference is expensive and is a function of both the model and the workload, which doesn't scale. Furthermore, optimizing LLM inference is challenging. Its performance depends on many configuration options, parallelization strategies, batching, scheduling, etc. Optimal configuration depends on the model, request trace, and application requirements. Thus, identifying the optimal configuration by experimentally running hundreds of configuration combinations for each deployment scenario is infeasible. To tackle this challenge, we present Vidur, the first large-scale, high-fidelity simulation framework for LLM inference. We validate our simulator on several LLMs and show that it estimates inference latency and throughput with less than 5% error rate. We also propose a configuration search tool, Vidur-Search, which leverages these high fidelity simulations to identify the most cost effective deployment configuration subject to provided SLOs. Vidur is able to find the best deployment configuration for LLaMA2-70B across a pool of A100 / A40 GPUs under 6 hours on a CPU machine, in contrast to a deployment-based exploration (310K GPU hours, costing \approx 479K). We will open-source Vidur so that researchers and practitioners can collaboratively explore model and systems optimizations for efficient deployment of LLMs.



THEME 3 SESSION OVERVIEW

Wednesday, March 13 from 10:30-11:30 AM

POSTER NO.	PRESENTER	TITLE
3.1	Abhiroop Bhattacharjee	Key-Value Clipping of Transformers on Memristive Crossbars for Write Noise Mitigation
3.2	Piyush Kumar	Vertically integrated end-to-end technology evaluation platform: 7nm MRAM design and 3nm PDK development
3.3	Md. Nahid Haque Shazon	Memory Compiler Development for Emerging Memories: Evaluating Spin Orbit-torque MRAMs for Future Cognitive Systems
3.4	Siri Narla	Content Addressable Memory for Hardware Accelerated Search at Advanced Nodes
3.5	Deepika Sharma	An Input Sparsity-aware Reconfigurable Digital CIM SNN Accelerator Core for Event-based Vision Tasks
3.6	Kartik Prabhu	MINOTAUR: Enabling Transformer Models at the Edge with Posits and Resistive RAM
3.7	Revanth Koduru	Small Signal Capacitance in Ferroelectrics: Mechanisms and Physical Insights from Phase-field simulations
3.8	Saion Roy	Compute SNDR-boosted 22nm MRAM IMC using Statistical Error Compensation
3.9	Akul Malhotra	BNN-Flip: Enhancing the Fault Tolerance and Security of Compute-in-Memory Enabled Binary Neural Network Accelerators
3.10	Connor Talley	A 40nm VLIW Edge Accelerator with 5MB of 0.256pJ/b RRAM and a Localization Solver for Bristle Robot Surveillance
3.11	Hyung Joon Byun	Evaluation and Optimization of 3D IC Architecture with Digital Compute-in-Memory Design
3.12	Jeffry Louis Victor	Rearranging Crossbar Weights for Enhanced DNN Accuracy in Deeply Scaled Technologies




3.13	Che-Kai Liu	Heterogeneous 3D Integrated CIM for Factorization with Holographic Perceptual Representations
3.14	Gopikrishnan Raveendran Nair	Addressing the Diversity in AI Computing: A 16nm RISC-V based SOC Chiplet for Graph and DNNs
3.15	Aviral Pandey	A Seizure Prediction SoC with a 17.2nJ/cIs Unsupervised Online-Learning Classifier and ZOOM Analog Frontends
3.16	Mohamed Ibrahim and Youbin Kim	Efficient Co-Design of a Programmable Hyperdimensional Processing Unit for Multi-Layer Cognition
3.17	Tianyi Zhang	3D In-Sensor Transformer-based Selection and Refinement for Early Object Detection

THEME 3: TECHNOLOGY-DRIVEN HARDWARE MOTIFS

THEME 3 PRESENTATION DETAILS

Wednesday, March 13 from 10:30-11:30 AM

SCHOLAR	POSTER DETAILS
<p>3.1 Abhiroop Bhattacharjee Yale University</p>  <p>Email: abhiroop.bhattacharjee@yale.edu</p> <p>PI: Priyadarshini Panda</p> <p>Level: PhD student</p> <p>Expected Graduation Date: 2025</p>	<p>Title: ClipFormer: Key-Value Clipping of Transformers on Memristive Crossbars for Write Noise Mitigation</p> <p>Abstract: Transformers have brought about a significant revolution in a variety of real-world applications, ranging from natural language processing to computer vision. Nevertheless, the conventional von-Neumann computing paradigm encounters challenges related to memory and bandwidth limitations when it comes to accelerating transformers, primarily due to their massive model sizes. To this end, In-memory Computing (IMC) crossbars based on Non-volatile Memories (NVMs) have emerged as a promising alternative for accelerating the inference of transformers. IMC crossbars exhibit the capability to execute highly parallelized Matrix-Vector-Multiplications (MVMs) with notable energy-efficiency and compactness, thereby overcoming the hindrances posed by traditional computing approaches. Despite the advantages, the use of analog MVM operations in crossbars introduces non-idealities such as stochastic read and write noise from the NVM devices. These non-idealities significantly impact the inference accuracy of deployed transformers. Specifically, our study revolves around the vulnerability of pre-trained Vision Transformers (ViTs) on crossbars due to the influence of write noise on the dynamically-generated Key (K) & Value (V) matrices in the attention layers. This effect has not been accounted for in prior works. Thus, for robust inference of ViTs on crossbars, we propose a transformation on the K & V matrices during inference termed as ‘ClipFormer’. ClipFormer aims to enhance the non-ideal inference accuracies of ViT models by mitigating write noise, without requiring additional hardware or incurring any training overhead. Importantly, ClipFormer is amenable to transformers deployed across any analog crossbar platform. Experimental results on the Imagenet-1k dataset using pre-trained DeiT-S transformers, subjected to both standard and variation-aware training, reveal a substantial increase of >10-40% in non-ideal accuracies at the high write noise regime when applying ClipFormer, while also achieving a 7 – 8% reduction in the total attention area & energy.</p>



3.2 Piyush Kumar

Georgia Institute of Technology



Email: pkumar315@gatech.edu

PI: Azad Naeemi

Level: PhD student

Expected Graduation Date:
December 2024

Title: Vertically integrated end-to-end technology evaluation platform:
7nm MRAM design and 3nm PDK development

Abstract: The goal of this task is to develop a vertically integrated end-to-end technology evaluation platform for various hardware motifs for cognitive systems based on CMOS and beyond CMOS devices. Toward that goal, we have been working on a rigorous technology modeling framework for CAM and MRAM arrays for various cognitive systems such as hyperdimensional computing and neural networks. At the 7nm node, we are working towards the characterization of complete sub-system for SOT/STT-based MRAM/CAM arrays using ASAP7 PDK. We have designed bitcell layouts for MRAM and CAM cells along with various peripherals which are used in the sub-array level read/write characterization. To understand the impact of design choices at advanced nodes, we are developing a 3nm GAAFET-based PDK with the goal of performing PnR using the PDK. Based on rigorous TCAD simulations, we are modeling interconnect with wire width down to 12nm, where we are considering various novel BEOL options.

3.3 Md. Nahid Haque Shazon

Georgia Institute of Technology



Email: mshazon3@gatech.edu

PI: Azad Naeemi

Level: PhD student

Expected Graduation Date:
August 2027

Title: Memory Compiler Development for Emerging Memories:
Evaluating Spin Orbit-torque MRAMs for Future Cognitive Systems

Abstract: In alignment with the goal to bridge the hardware designs of cognitive systems based on beyond CMOS devices, our work focuses on evaluating cell-level and array-level performances of Spin Orbit-torque MRAMs (SOT-MRAMs), which have emerged as promising non-volatile candidates for last-level cache due to their high endurance, sufficiently low read/write latency, and scalability. Since conventional drift-diffusion models are unable to capture the non-uniformity of spin current in nanoscale, our work focuses on calculating the spin current distribution based on finite element simulations and drift-diffusion equations in micromagnetic simulation to investigate the effect of spin current non-uniformity on magnetization switching dynamics. Furthermore, in order to evaluate the array-level performances of SOT-MRAMs while being used as cache memories, our work outlines a framework for a generic memory compiler, which will be able to generate a complete memory array along with necessary peripheral circuitry, irrespective of memory types and technology generations.



3.4 Siri Narla

Georgia Institute of Technology



Email: snarla6@gatech.edu

PI: Azad Naeemi

Level: PhD student

Expected Graduation Date:
December 2024

Title: Content Addressable Memory for Hardware Accelerated Search at Advanced Nodes

Abstract: The goal of this task is to thoroughly evaluate the parallel search enabled by CAMs at advanced nodes and highlight the advantages and challenges of using them for similarity search in big data applications. Towards that goal, we have designed CAM cells based on CMOS, spin-orbit torque (SOT), and ferroelectric field effect transistor (FeFET) devices and from their ASAP7-based physical layouts extracted cell parasitics using state-of-the-art EDA tools. These parasitics were used to develop SPICE netlists to model the search operations. We used a CAM-based dataset search and a sequential recommendation system with CAM-based similarity search to highlight the application-level performance degradation due to interconnect parasitics at the 7nm node. We proposed and evaluated two solutions to mitigate these interconnect effects. We are also working on further benchmarking CAM-based similarity search against graph and non-graph-based approximate nearest neighbor search techniques using the SIFT 1 billion dataset.

3.5 Deepika Sharma

Purdue University



Email: sharm444@purdue.edu

PI: Kaushik Roy

Level: PhD student

Expected Graduation Date:
May 2025

Title: An Input Sparsity-aware Reconfigurable Digital CIM SNN Accelerator Core for Event-based Vision Tasks

Abstract: This paper introduces a reconfigurable digital computing-in-memory (CIM) core for processing Spiking Neural Networks (SNNs). It has three key features: 1) reduces data movement at both macro and core levels, 2) supports three weight/Vmem bit precisions seamlessly, and 3) leverages input spike sparsity to enhance energy efficiency and reduce latency with minimal overhead. The proposed SNN core, fabricated in 65nm, achieves an energy efficiency of 2.5 - 5 TOPS/W at 50 MHz and 0.9V.



3.6 Kartik Prabhu
Stanford University



Email: kprabhu7@stanford.edu
PI: Priyanka Raina
Level: PhD student
Expected Graduation Date:
September 2024

Title: MINOTAUR: Enabling Transformer Models
at the Edge with Posits and Resistive RAM

Abstract: Transformer models achieve state-of-the-art accuracy but are challenging to run in resource-constrained edge environments, as they are large and difficult to quantize to 8b integers. MINOTAUR overcomes these challenges and enables both inference and training of machine learning models at the edge through (1) an alternative 8b floating-point data type, (2) a deep neural network accelerator optimized for operator fusion, and (3) temporal power-gating of on-chip non-volatile resistive RAM (RRAM). MINOTAUR uses 8b posits, an alternative to the IEEE-754 floating point standard that achieves a higher dynamic range for the same number of bits. Compared to bfloat16, 8b posit reduces both the memory capacity required and the memory access energy by 2x, with a 1.36x smaller area and 1.1x lower power multiply-accumulate (MAC) operator. Posit operations require type conversions that can introduce quantization errors, leading to accuracy degradation. MINOTAUR addresses this challenge through fusion of transformer operations, enabled by a configurable vector datapath; this improves inference accuracy and reduces the number of memory accesses. Finally, MINOTAUR fits large models (e.g., 12 MB MobileBERT-tiny) entirely on-chip without any external memory, and further reduces memory power with workload-optimized power gating. MINOTAUR is fabricated in 40nm, and demonstrates inference on ResNet-18 and MobileBERT-tiny, as well as finetuning of MobileBERT-tiny.

3.7 Revanth Koduru
Purdue University



Email: kodurur@purdue.edu
PI: Sumeet K. Gupta
Level: PhD student
Expected Graduation Date:
December 2025

Title: Small Signal Capacitance in Ferroelectrics: Mechanisms and Physical Insights
from Phase-field simulations

Abstract: In this work, we explore the underlying physical mechanisms governing the small signal capacitance in ferroelectrics, with a focus on Hafnium-oxide based ferroelectrics. Our study employs a time-dependent Ginzburg Landau (TDGL) equation-based phase-field framework, which emulates the polycrystalline nature of Hafnium Oxide by utilizing clustered distributions of the TDGL parameters. The simulation methodology utilizes a Metal-Ferroelectric-insulator-Metal (MFIM) capacitor structure and closely mirrors the experimental procedure for measuring small signal capacitance in ferroelectrics. The simulation outcomes successfully reproduce the distinctive butterfly C-V characteristics observed in ferroelectrics and unveils two dominant mechanisms – one operating near domain walls and the other within the bulk of the domain – responsible for the C-V characteristics. Our work intricates explores their interplay and impact on the ferroelectric capacitance under different domain configurations and densities. This work enhances the understanding of small signal capacitance in ferroelectrics and aids in driving the optimization strategies for ferroelectric capacitive applications.



3.8 Saion Roy

University of Illinois
Urbana-Champaign



Email: saionkr2@illinois.edu

PI: Naresh Shanbhag

Level: PhD student

Expected Graduation Date:

August 2024

Title: Compute SNDR-boosted 22nm MRAM IMC using Statistical Error Compensation

Abstract: Recent benchmarking data based on published prototypes clearly highlight that resistive in-memory computing (IMC) architectures currently lag behind SRAM IMCs and digital accelerators in both energy efficiency and compute density due to their low compute accuracy. In this work, we propose two methods for boosting the compute signal-to-noise-plus-distortion ratio (SNDR) of a 22nm MRAM IMC macro. They include a circuit approach of offset-compensating current sensing to reduce the static mismatch across ADC columns, and the algorithmic technique of statistical error compensation (SEC) to reduce the impact of non-linearity due to wire parasitics. We demonstrate that an SEC-enabled SNDR boost of 2.7-to-6 dB can be traded off to achieve a substantial 5x reduction in energy per 1b operation. Lastly, we use the chip data to validate the noise models for analog non-idealities which are then employed to analyze the fundamental limits on the compute SNDR of eNVM IMCs. Our analysis reveals that the maximum SNDR is primarily limited by pre-ADC analog non-idealities.

3.9 Akul Malhotra

Purdue University



Email: malhot23@purdue.edu

PI: Sumeet Gupta

Level: PhD student

Expected Graduation Date:

May 2025

Title: BNN-Flip: Enhancing the Fault Tolerance and Security of Compute-in-Memory Enabled Binary Neural Network Accelerators

Abstract: Compute-in-memory based binary neural networks or CiM- BNNs offer high energy/area efficiency for the design of edge deep neural network (DNN) accelerators, with only a mild accuracy reduction. However, for successful deployment, the design of CiM-BNNs must consider challenges such as memory faults and data security that plague existing DNN accelerators. In this work, we aim to mitigate both these problems simultaneously by proposing BNN-Flip, a training-free weight transformation algorithm that not only enhances the fault tolerance of CiM-BNNs but also provides protection from weight theft attacks. BNN- Flip inverts the rows and columns of the BNN weight matrix in a way that reduces the impact of memory faults on the CiM-BNN's inference accuracy while preserving the correctness of the CiM operation. Concurrently, our technique encodes the CiM-BNN weights, imparting security against weight theft. Our experiments on various CiM-BNNs show that BNN-Flip achieves an inference accuracy increase of up to 10.55% over the baseline (i.e. CiM-BNNs not employing BNN-Flip) in the presence of memory faults. Additionally, we show that the encoded weights generated by BNN-Flip furnish extremely low (near 'random guess') inference accuracy for the adversary attempting weight theft. The benefits of BNN-Flip come with an energy, latency, and area overhead of < 3%, < 2%, and <7%, respectively.



3.10 Connor Talley
Georgia Institute of Technology



Email: connor.talley@gatech.edu
PI: Arijit Raychowdhury
Level: PhD student
Expected Graduation Date:
May 2027

Title: A 40nm VLIW Edge Accelerator with 5MB of 0.256pJ/b RRAM and a
Localization Solver for Bristle Robot Surveillance

Abstract: We present a resistive random-access memory (RRAM)-based accelerator in 40nm CMOS with embedded RRAM for bristle robot navigation in stealth-surveillance tasks. The design supports accurate perception by maintaining high nonvolatile memory (NVM) density and using digital multiply-accumulates to improve accuracy. 10 very long instruction word (VLIW)-controlled NVM matrix units (NMUs) are integrated at the top level. The VLIW controller enables parallel coordination of MACs, memories, and program flow. This is combined with a 10T SRAM-based state-update accelerator utilizing voltage-controlled adjacent-cell shifting that enables 16× latency reduction compared to a 6T SRAM-based method of data-readout followed by addition and write-back. The design includes 5MB of RRAM, post-ECC, and improves chip-level NVM density 3.6× to 2.07Mb/mm², access energy 3.79× to 0.256pJ/b, and area-normalized bandwidth 3.64× to 0.63GB/s/mm². The design supports a retentive sleep mode through core voltage reduction and setting RRAM modules to power down (PD), reducing chip-level measured leakage to 110μW.

3.11 Hyung Joon Byun
Cornell Tech



Email: hb479@cornell.edu
PI: Jae-sun Seo
Level: PhD student
Expected Graduation Date:
May 2028

Title: Evaluation and Optimization of 3D IC Architecture with
Digital Compute-in-Memory Design

Abstract: Due to the compute-intensive DNN algorithms, several two-dimensional architectures have been suggested including systolic arrays or compute-in-memory (CIM) arrays for energy-efficient DNN inference or training. To increase the energy efficiency within a constrained area, three-dimensional technologies have been actively investigated with the potential to decrease the data path length or increase the activation buffer size. Several works have reported the three-dimensional architectures using non-CIM designs, but investigations on three-dimensional architectures with CIM macros have not been conducted extensively where our work focused on. In this talk, we were to leverage digital CIM (DCIM) macros and various three-dimensional architectures to find the opportunity of increased energy efficiency compared to two-dimensional structures. We have built in-house simulators calculating energy and area given high-level hardware descriptions and DNN workloads. We have investigated different types of 3D DCIM architectures and dataflows which have shown 1.74x energy savings on average while reducing footprint area up to 68.56% compared to the 2D systolic arrays with the same throughput.



3.12 Jeffrey Louis Victor

Purdue University



Email: louis8@purdue.edu

PI: Sumeet Gupta

Level: PhD student

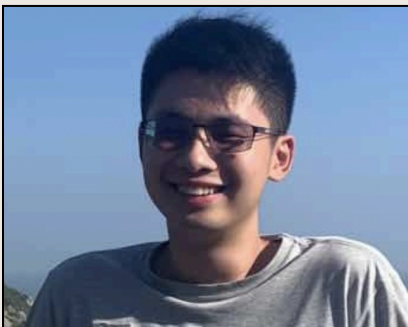
Expected Graduation Date:
May 2025

Title: Rearranging Crossbar Weights for Enhanced DNN Accuracy in Deeply Scaled Technologies

Abstract: Crossbars are central to in-memory computing (IMC) based acceleration of Deep Neural Networks (DNNs). But the associated non-idealities such as interconnect resistance limit their robustness, leading to a significant reduction in DNN inference accuracy. To address this, we propose a technique, which rearranges weights in crossbar arrays to mitigate the detrimental effects of aggravated wire resistance observed in deeply scaled nodes. We analyze the efficacy of our technique on 8T-SRAM and FeFET arrays of varying sizes (from 64x64 to 256x256) at the 7 nm node and observe an increase in inference accuracy from 47.78% to 83.5% for ResNet-20/CIFAR-10. We also study the synergistic use of our technique with Partial-Wordline-Activation which increases inference accuracy from 47.78% to 89.86%. Further, we establish that the proposed technique incurs minimal energy and latency overheads.

3.13 Che-Kai Liu

Georgia Institute of Technology



Email: che-kai@gatech.edu

PI: Arijit Raychowdhury

Level: PhD student

Expected Graduation Date:
2029

Title: Heterogeneous 3D Integrated CIM for Factorization with Holographic Perceptual Representations

Abstract: Disentangling attributes of various sensory signals is central to human-like perception and reasoning and a critical task for higher-order cognitive and neuro-symbolic AI systems. An elegant approach to represent this intricate factorization is via high-dimensional holographic vectors drawing on brain-inspired vector symbolic architectures. However, holographic factorization involves iterative computation with high-dimensional matrix-vector multiplications and suffers from non-convergence problems. In this paper, we present H3DFact, a heterogeneous 3D integrated in-memory compute engine capable of efficiently factorizing high-dimensional holographic representations. H3DFact exploits the computation-in-superposition capability of holographic vectors and the intrinsic stochasticity associated with memristive-based 3D compute-in-memory. Evaluated on large-scale factorization and perceptual problems, H3DFact demonstrates superior capability in factorization accuracy and operational capacity by up to five orders of magnitude, with 5.5x compute density, 1.2x energy efficiency improvements, and 5.9x less silicon footprint compared to iso-capacity 2D designs.



3.14 Gopikrishnan

Raveendran Nair

University of Minnesota



Email: ravee013@umn.edu

PI: Yu (Kevin) Cao

Level: PhD student

Expected Graduation Date:

May 2025

Title: Addressing the Diversity in AI Computing:
A 16nm RISC-V based SOC Chiplet for Graph and DNNs

Abstract: AI algorithms are increasingly diverse, from dense to sparse, and from regular to irregular. Hardware platforms, such as in-memory computing (IMC) are optimal for dense matrix/vector computation but significantly underutilized for sparse and random data workloads, such as GCNs. To efficiently manage such diversity in hardware, we propose a programmable SOC chiplet that dynamically balances the computation requirements across different design levels. The chiplet comprises a RISC-V CPU core and a heterogenous co-processor accelerator core. We perform a comprehensive analysis of representative DNNs and GCNs, identify the type of computations required, and propose a common set of design macros to build the accelerator. We integrate two types of processing elements:(1) Latch-based digital IMCs to dense computation, and (2) The digital SIMD array with fine-grained control to handle irregular and sparse workloads. These PEs are integrated into a programmable architecture, enabling support for various memory access and computation patterns. Based on Intel's 16nm design data, the new SOC accelerator achieves an 11x improvement in latency compared to state-of-the-art homogeneous accelerators.

3.15 Aviral Pandey

UC Berkeley



Email: aviral0607@berkeley.edu

PI: Jan Rabaey

Level: PhD student

Expected Graduation Date:

May 2025

Title: A Seizure Prediction SoC with a 17.2nJ/cIs Unsupervised
Online-Learning Classifier and ZOOM Analog Frontends

Abstract: This work presents SPIRIT, a SoC integrating an unsupervised online-learning seizure prediction classifier with eight 13-bit ENOB ZOOM Analog Frontends. SPIRIT achieves, on average, 97.5%/96.2% sensitivity/specificity predicting seizures 8.4 minutes before they occur. SPIRIT consumes 132.4μW/17.2μW and 3mm²/0.14mm² (with and without the frontends), the lowest reported for a prediction SoC by >21,000x and 10x respectively, and for an online tuning classifier by 134x and 28x respectively.



3.16 Mohamed Ibrahim

Georgia Institute of Technology



Email: mibrahim81@gatech.edu

and youbin_kim@berkeley.edu

PI: Arijit Raychowdhury and
Jan Rabaey

Level: Post-doc and
PhD Student, respectively

Title: Efficient Co-Design of a Programmable Hyperdimensional Processing Unit for Multi-Layer Cognition

Abstract: The methodology used to design and co-optimize the very first general-purpose hyperdimensional (HD) processing unit capable of executing a broad spectrum of HD workloads (called "HPU") is presented. HD computing is a brain-inspired computational paradigm that uses the principles of high-dimensional mathematics to perform cognitive tasks. While considerable efforts have been spent toward realizing efficient HD processors, all of these targeted specific application domains, most often pattern classification. In contrast, the HPU design addresses the multiple layers of a cognitive process. A structured methodology identifies the kernel HD computations recurring at each of these layers and maps them onto a unified and parameterized architectural model. The effectiveness in terms of runtime and energy consumption of the approach is evaluated. The results show that the resulting HPU efficiently processes the full range of HD algorithms, and far outperforms a baseline implementation.

3.17 Tianyi Zhang

University of Minnesota



Email: zhan9167@umn.edu

PI: Yu (Kevin) Cao

Level: PhD student

Expected Graduation Date:
May 2025

Title: 3D In-Sensor Transformer-based Selection and Refinement for Early Object Detection

Abstract: Early object detection (OD) is pivotal for ensuring the safety of dynamic systems. Current OD algorithms have limited success for small objects at a long distance. While employing super-resolution can magnify objects for improved performance, the process of upscaling entire images introduces substantial computational overhead from the background. Inspired by the selective attention in human vision, we propose a saliency-based processing step in the sensor. Leveraging 3D integration to stack the AI computing layer with the sensing layer, such in-sensor computing (ISC) will select pixels corresponding to salient objects. Subsequently, we elaborate on the details of selected small objects. Our approach is built upon a transformer-based network and integrates the diffusion model to improve the detection accuracy. As demonstrated on BDD100K, our algorithms enhance the mAP for small objects from 1.03 to 8.93 and reduce the data volume in computation by more than 77%.



THEME 4 SESSION OVERVIEW


Wednesday, March 13 from 3:00-3:45 PM

POSTER NO.	PRESENTER	TITLE
4.1	Youngeun Kim	One-stage Prompt-based Continual Learning: Towards computation-efficient and privacy-preserving continual learning framework
4.2	Zishen Wan	MuIBERRY: Enabling Bit-Error Robustness for Energy-Efficient Multi-Agent Autonomous UAV Systems
4.3	Yeshwanth Venkatesha	Accelerating Resource-Constrained Federated Learning for Vision and Language Models
4.4	Christopher Richardson	Alignment with Language Models through Language Feedback and Retrieval Augmentation
4.5	Tyler Lizzo	Rank Reduction using Weight Matrix Decomposition
4.6	John Taylor	The A.I. Virtual Assistant Pipeline: AVA Digital Human
4.7	Benjamin Reichman	Dense Passage Retrieval... Is it retrieving?
4.8	Chaojian Li	On-Device AR/VR 3D Reconstruction and Rendering
4.9	Anirudh Sundar	gTBLS: Generating Tables from Text by Conditional Question Answering
4.10	Arghadip Das	Towards Energy-Efficient Collaborative Inference
4.11	Mariah Schrum	Towards Online Adaptive Deep Brain Stimulation
4.12	Xiaofan Yu	SensorQA: Sensor-based Question Answering for Real-Life Applications
4.13	Sai Aparna Aketi	Global Update Tracking: A Decentralized Learning Algorithm for Heterogeneous Data
4.14	Edward Sadler	Collaborative Robots in Automated and Energy-Efficient Fruit Monitoring and Harvesting
4.15	Cambridge Yang	On the Learnability of Reinforcement Learning Objectives

THEME 4: COLLABORATIVE INTELLIGENCE

PRESENTATION DETAILS

Wednesday, March 13 from 3:00-3:45 PM

SCHOLAR	POSTER DETAILS
<p>4.1 Youngeun Kim Yale University</p>  <p>Email: youngeun.kim@yale.edu PI: Priyadarshini Panda Level: PhD student Expected Graduation Date: May 2024</p>	<p>Title: One-Stage Prompt-Based Continual Learning: Towards Computation-Efficient And Privacy-Preserving Continual Learning Framework</p> <p>Abstract: Training models effectively and efficiently on a continuous stream of data presents a significant practical hurdle. A straightforward approach would entail accumulating both prior and new data and then updating the model using this comprehensive dataset. However, as data volume grows, fully retraining a model on such extensive data becomes increasingly impractical. Additionally, storing past data can raise privacy issues, such as those highlighted by the EU General Data Protection Regulation (GDPR). An alternative solution is to adapt the model based solely on currently incoming data, eschewing any access to past data. This paradigm is termed as rehearsal-free continual learning, and the primary goal is to diminish the effects of catastrophic forgetting on previously acquired data. Among the rehearsal-free continual learning methods, Prompt-based Continual Learning (PCL) stands out as it has demonstrated state-of-the-art performance in image classification tasks, even surpassing rehearsal-based methods. PCL utilizes a pre-trained Vision Transformer (ViT) and refines the model by training learnable tokens on the given data. PCL adopts a prompt pool-based training scheme where different prompts are selected and trained for each continual learning stage. Although PCL methods show state-of-the-art performance, huge computational costs from the two ViT feed-forward stages make the model difficult to deploy into resource-constrained devices. Specifically, the PCL method requires two-stage ViT feedforward steps. We refer to this approach as a two-stage PCL method. To address this, we introduce a one-stage PCL framework by directly using the intermediate layer's token embedding as a prompt query. This design removes the need for an additional feed-forward stage, resulting in ~50% computational cost reduction for both training and inference with marginal accuracy drop (< 1%) on public class-incremental continual learning benchmarks including CIFAR-100 and ImageNet-R.</p>



4.2 Zishen Wan

Georgia Institute of Technology



Email: zishenwan@gatech.edu

PI: Arijit Raychowdhury &
Tushar Krishna

Level: PhD student

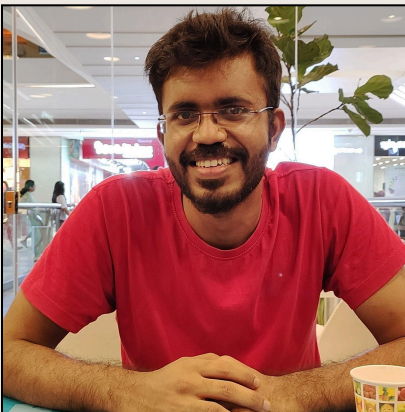
Expected Graduation Date:
August 2025

Title: MulBERRY: Enabling Bit-Error Robustness for Energy-Efficient Multi-Agent Autonomous UAV Systems

Abstract: The adoption of autonomous swarms, consisting of a multitude of unmanned aerial vehicles (UAVs), has become prevalent in mainstream applications. These swarms are expected to collaboratively carry out navigation tasks and employ complex reinforcement learning models within stringent onboard size-weight-and-power constraints. While techniques such as reducing onboard voltage can improve energy efficiency, they can lead to on-chip bit failures that are detrimental to mission safety and performance. To this end, we propose MulBERRY, a multi-agent robust learning framework to enhance bit error robustness and energy efficiency for autonomous UAV swarms. MulBERRY supports multi-agent robust learning, both offline and on-device, with adaptive and collaborative agent-server optimizations. For the first time, MulBERRY demonstrates the practicality of robust low-voltage operation on UAV swarms leading to energy savings in both compute and mission quality-of-flight. We conduct extensive system-level multi-UAV experiments with algorithm-level robust learning and hardware-level characterizations, demonstrating that MulBERRY achieves robustness-performance-efficiency co-optimizations, with up to 18.97% flight energy reduction and 22.07% more number of successful missions.

4.3 Yeshwanth Venkatesha

Yale University



Email:

yeshwanth.venkatesha@yale.edu

PI: Priyadarshini Panda

Level: PhD student

Expected Graduation Date:
May 2025

Title: Accelerating Resource-Constrained Federated Learning for Vision and Language Models

Abstract: Federated Learning is gaining attention for its applications in privacy-preserving AI solutions at the edge. However, edge devices often face limitations in compute power, memory, and communication bandwidth. Traditional federated approaches often overlook these constraints. In this context, we examine two studies to reduce the training effort at the clients and improve federated aggregation for faster convergence. Firstly, we address the challenge of client-specific architecture using a novel Neural Architecture Search (NAS) methodology integrated into the federated framework. We enhance the search process by employing an innovative sampling strategy that efficiently divides the search space. Our method achieves near iso-accuracy as compared to baseline models with 50% fewer resources on the CIFAR10 dataset. Secondly, we explore the training efficiency of Parameter-Efficient Fine Tuning (PEFT) for language models in a federated environment. This involves reducing the training complexity with head pruning followed by weighted aggregation and strategic client selection to accelerate convergence. We reduce the communication complexity by 1.8x while maintaining an accuracy drop of less than 2% on MutiNLI dataset.



4.4 Christopher Richardson

Georgia Institute of Technology



Email: crichardson8@gatech.edu

PI: Larry Heck

Level: PhD student

Expected Graduation Date:
May 2025

Title: Alignment with Language Models through Language Feedback and Retrieval Augmentation

Abstract: Large language models (LLMs) have become a central focus of AI research across a wide variety of domains. While state-of-the-art LLMs have shown significant parsing and understanding abilities, there remain major challenges with aligning model outputs to users' needs. This problem is called the alignment problem and has become a central thread in the AI research community. We seek to explore the usage of language feedback to address this problem. Utilizing direct user feedback is one of the most promising ways to improve alignment, with language feedback enabling a rich and dynamic way for users to express the alignment gap to the model. How to fully and optimally utilize language feedback remains an open problem. We propose a method to improve feedback utilization with retrieval-augmentation. Our method involves using the language feedback itself to query a database of examples in order to retrieve past samples where similar feedback was given. Our goal is to show performance improvement using our method relative to traditional retrieval-augmentation approaches.

4.5 Tyler Lizzo

Georgia Institute of Technology



Email: lizzo@gatech.edu

PI: Larry Heck

Level: PhD student

Expected Graduation Date:
May 2028

Title: Rank Reduction using Weight Matrix Decomposition

Abstract: An important challenge for conversational AI is the balance between the size and accuracy of large language models (LLMs). Research in LLMs has focused on greatly scaling size, with little consideration given to computational restrictions. For resource-constrained environments, techniques must be developed to bridge this computation gap while maintaining the system's accuracy. Prior works have looked at matrix decomposition on the weight matrices in the layers of a transformer, relying on the rank deficiency of these matrices to maintain accuracy while greatly reducing the number of parameters in the system. This work builds upon this idea for task-specific situations by integrating matrix decompositions with selective fine-tuning. This work will begin to bridge the gap between the algorithmic and hardware groups in COCOSYS by designing hardware-aware matrix decompositions.



4.6 John Taylor

Georgia Institute of Technology



Email: john.taylor@gatech.edu

PI: Larry Heck

Level: Research Engineer

Title: The A.I. Virtual Assistant Pipeline: AVA Digital Human

Abstract: The A.I. Virtual Assistant, also known as Ava, is a full-bodied digital avatar that can have conversations with a user by listening and speaking. The avatar features a realistic human appearance, idle body animations, and facial animations with real-time lip synch. The pipeline orchestrating Ava's various subsystems includes state-of-the-art language models, automatic speech recognition, and text-to-speech engines. Each part of the pipeline works individually, allowing the components to be swapped out or altered easily. The pipeline's modular nature enables rapid evolution, simplifying the implementation of emerging technological advances and integration with other CoCoSys projects. We envision a plethora of expansions for Ava, ranging from continual learning for more personalized and adaptive collaboration with users, to multimodal systems for more flexible methods of interaction, handling communication-based on vision and body language.

4.7 Benjamin Reichman

Georgia Institute of Technology



Email: bZR@gatech.edu

PI: Larry Heck

Level: PhD student

Expected Graduation Date:

May 2026

Title: Dense Passage Retrieval... Is it retrieving?

Abstract: Dense passage retrieval (DPR) is the first step in the retrieval augmented generation (RAG) paradigm for improving large language models (LLM) performance. DPR fine-tunes pre-trained networks to enhance the alignment of the embeddings between queries and relevant textual data. A deeper understanding of DPR fine-tuning will be required to fundamentally unlock the full potential of this approach. In this work, we explore DPR-trained models mechanistically by using a combination of probing, layer activation analysis, and model editing. Our experiments show that DPR training decentralizes how knowledge is stored in the network, creating multiple access pathways to the same information. We also uncover a limitation in this training style: the internal knowledge of the pre-trained model bounds what the retrieval model can retrieve.



4.8 Chaojian Li

Georgia Institute of Technology



Email: cli851@gatech.edu

PI: Yingyan (Celine) Lin

Level: PhD student

Expected Graduation Date:

May 2025

Title: On-Device AR/VR 3D Reconstruction and Rendering

Abstract: Neural Radiance Field (NeRF) based 3D reconstruction is highly desirable for immersive Augmented and Virtual Reality (AR/VR) applications, but achieving instant (i.e., < 5 seconds) on-device NeRF training and real-time (i.e., > 30 FPS) high-quality NeRF rendering remains a challenge. In this work, we first identify the inefficiency bottleneck during NeRF training: the need to interpolate NeRF embeddings up to 200,000 times from a 3D embedding grid during each training iteration. To alleviate this, we propose an algorithm-hardware co-design acceleration framework that achieves instant on-device NeRF training. Our algorithm decomposes the embedding grid representation in terms of color and density. Our hardware accelerator further reduces the dominant memory accesses for embedding grid interpolation. Moreover, we developed a web-based real-time NeRF rendering viewer to allow the reconstructed scenes can be viewed on cross-platform devices even without the proposed hardware accelerator.

4.9 Anirudh Sundar

Georgia Institute of Technology



Email: asundar34@gatech.edu

PI: Larry Heck

Level: PhD student

Expected Graduation Date:

May 2025

Title: gTBLS: Generating Tables from Text by Conditional Question Answering

Abstract: Distilling large, unstructured text into a structured, condensed form such as tables is an open research problem in Natural Language Processing. Prior approaches address this task through additional parameters in the Transformer's attention mechanism. This poster presents Generative Tables gTBLS, a two-stage, parameter-efficient solution to automatically construct structured tables from text. The first stage infers table structure (row and column headers) from the text, and the second stage formulates questions and fine-tunes a causal language model to answer them. gTBLS improves prior approaches by up to 21% in BERTScore on the table content generation task of the E2E, WikiTableText, WikiBio, and RotoWire datasets with 66% fewer parameters.



4.10 Arghadip Das

Purdue University



Email: das169@purdue.edu

PI: Vijay Raghunathan

Level: PhD student

Expected Graduation Date:

December 2026

Title: Towards Energy-Efficient Collaborative Inference

Abstract: Collaborative inference applications based on distributed deep neural networks (DDNNs) are becoming increasingly popular. In these applications, DDNNs are used to classify 3D objects from a set of 2D images or views, known as multiview convolutional neural networks (MVCNN). However, due to the intensive computational demands, substantial communication overhead, high inference delay, and energy limits, it is difficult to deploy MVCNN on resource-constrained edge devices. We propose, for the first time, the concept of jointly optimizing distributed collaborative inference systems by employing a set of optimizations that explores the performance, accuracy, and energy trade-off space. Our proposed techniques prune the large design space using the non-uniform contribution of various perspectives/views in a multiview CNN to achieve an optimized quality-energy trade-off. In addition, we also propose a novel energy-aware heuristic, which dynamically configures edge inference systems based on application quality bounds and increases the system lifetime. Experimental results on an Intel Stratix IV FPGA development board-based prototype that executes 12-view 3D object classification show significant energy savings ($2.6\times$ – $8\times$) for minimal ($<1\%$) application-level quality loss.

4.11 Mariah Schrum

UC Berkeley



Email:

mariahschrum@berkeley.edu

PI: Anca Dragan

Level: Post-doc

Title: Towards Online Adaptive Deep Brain Stimulation

Abstract: Deep Brain Stimulation (DBS) can treat neurological conditions such as Parkinson's disease and post-stroke motor deficits by influencing abnormal neural activity. Because of patient heterogeneity, each patient requires a unique DBS control policy to achieve optimal neural responses. Model-free reinforcement learning (MFRL) holds promise in learning effective policies for a variety of control tasks similar to DBS. However, MFRL's limitation lies in its need for numerous environmental interactions, making it impractical for DBS in which interactions with the patient (i.e., brain stimulations) are costly. In this work, we introduce a novel, model-based reinforcement learning (MBRL) approach for learning neural coprocessor policies for DBS. Our key insight is to break down co-processor policy learning into two steps: 1) learn the value of world actions via a biomechanically realistic simulation, and 2) learn the mapping from coprocessor stimulation to world actions via online interaction. We show that our approach surpasses the limitations of traditional MFRL methods in terms of sample efficiency and task success and outperforms baseline MBRL approaches in a neurologically realistic model of an injured brain. This work establishes a foundation for improving the understanding and efficacy of RL solutions for DBS.



4.12 Xiaofan Yu

UC San Diego



Email: xlyu@ucsd.edu

PI: Tajana Rosing

Level: PhD student

Expected Graduation Date:
December 2024

Title: SensorQA: Sensor-based Question Answering for Real-Life Applications

Abstract: In recent years, billions of sensors have been deployed for providing continuous monitoring in real-life applications. Machine learning techniques are essential to extract hidden patterns in the sensor readings and perform comprehensive analyses. Although it accounts for many real-world scenarios, providing only classification or regression results limits the interpretation of sensor data to passive information query. Users, instead, may actively raise questions about the specific type of information that she would like to learn. For example, with IMU and heart rate data, a user may want to question whether her daily activities are sufficient to keep fit. PI Rosing's and PI Heck's teams collaborate together to design and implement a practical QA system for sensor-based applications. The project includes two phases: (1) building a QA dataset from sensor data, (2) training a multimodal model to jointly process the sensor data and input question, targeting at generating appropriate answers to arbitrary questions about arbitrary sensors.

4.13 Sai Aparna Aketi

Purdue University



Email: saketi@purdue.edu

PI: Kaushik Roy

Level: PhD student

Expected Graduation Date:
April 2024

Title: Global Update Tracking: A Decentralized Learning Algorithm for Heterogeneous Data

Abstract: Decentralized learning enables the training of deep learning models over large distributed datasets generated at different locations, without the need for a central server. However, in practical scenarios, the data distribution across these devices can be significantly different, leading to a degradation in model performance. In this paper, we focus on designing a decentralized learning algorithm that is less susceptible to variations in data distribution across devices. We propose Global Update Tracking (GUT), a novel tracking-based method that aims to mitigate the impact of heterogeneous data in decentralized learning without introducing any communication overhead. We demonstrate the effectiveness of the proposed technique through an exhaustive set of experiments on various Computer Vision datasets (CIFAR-10, CIFAR-100, Fashion MNIST, and ImageNet), model architectures, and network topologies. Our experiments show that the proposed method achieves state-of-the-art performance for decentralized learning on heterogeneous data via a 1–6% improvement in test accuracy compared to other existing techniques.



4.14 Edward Sadler
Kennesaw State University



Email:

esadler4@students.kennesaw.edu

PI: Yan Fang

Level: Undergraduate

Expected Graduation Date:

May 2025

Title: Collaborative Robots in Automated and
Energy-Efficient Fruit Monitoring and Harvesting

Abstract: This project aims to develop a collaborative robots system capable of automating monitoring and harvesting of fruits in agriculture. Using visual sensors, the drones will detect, recognize, and localize target fruits, and then cut their pedicel to make them fall and be caught by a collector. A ground vehicle robot serves as a mobile landing platform to carry the drone. The project will focus on designing a collaborative UAV-UGV intelligence, optimizing energy efficiency using tiny machine-learning neural networks, and implementing end-to-end control. We expect to demonstrate a customized mini drone that can fly around a plant and accurately approach target fruits. This project aims to address the challenges of labor-intensive fruit harvesting and revolutionize agriculture with automated solutions.

4.15 Cambridge Yang
MIT



Email: camyang@csail.mit.edu

PI: Michael Carbin

Level: PhD student

Expected Graduation Date:

August 2024

Title: On the Learnability of Reinforcement Learning Objectives

Abstract: In reinforcement learning, the classic objectives of maximizing discounted and finite-horizon cumulative rewards are PAC-learnable: Some algorithms learn a near-optimal policy with high probability using a finite amount of samples and computation. Researchers have recently introduced objectives and corresponding reinforcement-learning algorithms beyond the classic cumulative rewards, such as objectives specified as linear temporal logic formulas. However, questions about the PAC-learnability of these new objectives have remained open. This talk discusses the PAC-learnability of general reinforcement-learning objectives. First, we discuss a sufficient and necessary condition of PAC-learnable linear temporal logic objective. Second, we discuss a sufficient condition of PAC-learnable general objective. We conclude with future directions for the specifications of learnable reinforcement learning objectives.



CoCoSys

CENTER FOR THE
CO-DESIGN OF COGNITIVE SYSTEMS

