A3D3 all-hands meeting

Friday 7 October 2022 - Friday 7 October 2022

Southern Methodist University

Book of Abstracts

Contents

1

Welcome and Introduction

Research Area / 2

Hardware-Algorithm Co-development

Research Area / 3

HEP

Research Area / 4

MMA

Research Area / 5

Neuroscience

Research Area / 6

Heterogeneous system

Research Area / 7

Heterogenous & Targeted system

Activity Group / 8

Education

Activity Group / 9

Community Engagement

Activity Group / 10

HDR Ecosystem

Activity Group / 11

Equity and Diversity

12

Panel review with external committee

Co-authors: Frank Wuerthwein [1]; Mitra Taheri [2]; Nhan Tran [3]; Patrick Brady [4]; Saskia de Vries [5]; Stephen Neuendorffer [6]

[1] *Univ. of California San Diego (US)*
[2] *Johns Hopkins University*
[3] *Fermi National Accelerator Lab. (US)*
[4] *University of Wisconsin-Milwaukee*
[5] *Allen Institute*
[6] *AMD Xilinx*

Moderator: Philip Coleman Harris

13

Closing

Lightening talks / 14

Neural encoding of proprioception of the limbs in the mouse primary somatosensory and motor cortices

Authors: Maria Dadarlat[None]; Megan Hope Lipton[None]

Rodents rely on proprioceptive information from the periphery to guide and coordinate precise forelimb and hindlimb movements, a process called sensorimotor integration. The mouse primary somatosensory (S1) and primary motor (M1) cortices are known to be necessary for adapting motor commands to new sensory environments, and recent work suggests neurons in the forelimb area of S1 encode proprioceptive information about contralateral forelimb movement. However, we do not know how proprioception of all four limbs is represented across multiple brain regions. To address this question and to isolate pure somatosensory responses (proprioception and touch) from

motor commands that would be present in awake animals, we recorded neural responses to passive movement of ipsilateral and contralateral limbs in eight mice under anesthesia. Using stereotaxic coordinates to locate S1 and M1 forelimb and hindlimb areas, we performed unilateral two-photon imaging over these two regions simultaneously in mice expressing GCaMP6s, a highly sensitive fluorescent indicator of neuronal activity. A brushing motion was used to provide cutaneous and proprioceptive stimulation to each limb (blocks of five trials per limb were repeated across three cycles). Altogether, we recorded the activity of 12,895 neurons, of which 2,053 neurons (16%) were significantly modulated by passive movement of at least one limb ($p < 0.02$, Wilcoxon rank-sum test on single trial responses vs. baseline). Of significantly modulated neurons, 48% responded to movement of the contralateral hindlimb, 15% to the ipsilateral hindlimb, 30% to the contralateral forelimb, and 7% to the ipsilateral forelimb. A subset of neurons (9%) was significantly modulated by more than one type of limb movement, most often ipsilateral and contralateral hindlimb movement. In terms of response amplitude, neurons that were significantly modulated by contralateral movements had larger responses than those modulated by ipsilateral movements (hindlimb: $dF = 0.90 \pm 0.01$ SEM contralateral vs. $dF = 0.79 \pm 0.01$ SEM ipsilateral, $p = 5.1 \times 10\text{-}39$; forelimb: $dF = 0.78 \pm 0.01$ SEM contralateral vs. $dF = 0.74 \pm 0.01$ SEM ipsilateral, $p = 0.012$). In summary, we found evidence of proprioceptive signals related to both ipsilateral and contralateral forelimbs and hindlimbs across primary somatosensory and motor cortices of the mouse. The distributed nature of these responses, across cortical regions and limbs, could be an indication of how proprioception guides the formation of motor commands within the mouse cortex.

Lightening talks / 15

Deep Learning for the High Granularity Calorimeter L1Trigger

Author: Rohan Shenoy[1]

[1] *Univ. of California San Diego (US)*

The High Granularity Calorimeter (HGCAL) is part of the High Luminosity upgrade of the CMS detector at the Large Hadron Collider (HL-LHC). For the trigger primitive generation of the 6 million channels in this detector, data compression at the front end may be accomplished by using deep-learning techniques using an on-ASICs network. The Endcap Trigger Concentrator (ECON-T) ASIC foresees an encoder based on a convolutional neural network (CNN). The performance is evaluated using the earth mover's distance (EMD). Ideally, we would like to quantify the loss between the input and the decoded image at every step of the training using the EMD. However, the EMD is not differentiable and can therefore not be used directly as a loss function for gradient descent. The task of this project is to approximate the EMD using a separate set of CNNs and then implement the EMD NN as a custom loss for the ASIC encoder training, with the goal of achieving better physics performance.

Lightening talks / 16

Interaction Network Autoencoder in the Level-1 Trigger

Author: Sukanya Krishna[1]

[1] *Univ. of California San Diego (US)*

At the LHC, the FPGA-based real-time data filter system that rapidly decides which collision events to record, known as the level-1 trigger, requires small models because of the low latency budget and other computing resource constraints. To enhance the sensitivity to unknown new physics, we want to put generic anomaly detection algorithms into the trigger. Past research suggests that graph neural network (GNN) based autoencoders can be effective mechanisms for reconstructing particle jets and isolating anomalous signals from background data. Rather than treating particle jets as ordered sequences or images, interaction networks embed particle jet showers as a graph and exploit particle-particle relationships to efficiently encode and reconstruct particle-level information within jets. This project investigates graph-based standard and variational autoencoders. The two objectives in this project are to evaluate the anomaly detection performance against other kinds of autoencoder structures (e.g. convolutional or fully-connected) and implement the model on an FPGA to meet L1 trigger requirements.

Lightening talks / 17

Semi-supervised Graph Neural Networks for Pileup Noise Removal

Authors: Nhan Tran[1]; Shikun Liu[None]; Tianchun Li[None]

Co-authors: Garyfallia Paspalaki [2]; Miaoyuan Liu [2]; Pan Li ; Yongbin Feng [1]

[1] *Fermi National Accelerator Lab. (US)*

[2] *Purdue University (US)*

The high instantaneous luminosity of the CERN Large Hadron Collider leads to multiple proton-proton interactions in the same or nearby bunch crossings (pileup). Advanced pileup mitigation algorithms are designed to remove this noise from pileup particles and improve the performance of crucial physics observables. This study implements a semi-supervised graph neural network for particle-level pileup noise removal, by identifying individual particles produced from pileup. The graph neural network is firstly trained on charged particles with known labels, which can be obtained from detector measurements on data or simulation, and then inferred on neutral particles for which such labels are missing. This semi-supervised approach does not depend on the neutral particle pileup label information from simulation, and thus allows us to perform training directly on experimental data. The performance of this approach is found to be consistently better than widely-used domain algorithms and comparable to the fully-supervised training using simulation truth information. The study serves as the first attempt at applying semi-supervised learning techniques to pileup mitigation and opens up a new direction for fully data-driven machine learning pileup mitigation studies.
In the semi-supervised pileup mitigation study, model transferability from charged particles to neutral particles depends on the assumption that the features of training charged particles and testing neutral particles are from the same distribution. This motivates us to think of a broader problem that the simulation data and experimental data have different distributions and how the model may generalize. We would like to present some of our recent findings on how to make graph neural networks more generalizable when such a distribution gap exists.

Lightening talks / 18

Contrastive learning for correction of data/monte-carlo disagreements

Authors: Dylan Sheldon Rankin[1]; Philip Coleman Harris[1]; Simon Rothman[1]

[1] *Massachusetts Inst. of Technology (US)*

Many high-energy physics analyses rely on various machine learning models for both event reconstruction and signal/background discrimination. One of the major sources of systematic uncertainty in these analyses is due to residual mismodelling in the detailed simulation samples used to train these algorithms. In this work, we will discuss a novel approach to correcting for these systematic effects using contrastive learning and demonstrate some preliminary results.

Lightening talks / 19

PyLog: An Algorithm-Centric Python-Based FPGA Programming and Synthesis Flow

Author: Jialiang Zhang[None]

The fast-growing complexity of new applications and new use scenarios poses serious challenges for computing systems. Heterogeneous systems consist of different types of processors and accelerators, and provide unique combined benefits of hardware acceleration from each individual component. CPU-FPGA heterogeneous systems provide both programmable logic and general-purpose processors, and they have demonstrated great flexibility, performance, and efficiency. Heterogeneous systems have been created and deployed in many different applications and scenarios. However, as system complexity and application complexity grow rapidly, programming and optimizing heterogeneous systems require great manual efforts and consume a lot of time. In this work, we propose a Python-based high-level programming framework to simplify programming and optimization of CPU-FPGA heterogeneous systems. The proposed high-level operations isolate underlying hardware details from programmers and provide more optimization opportunities for the compiler.

Lightening talks / 20

ScaleFlow: Scalable High-Level Synthesis for Large Dataflow Applications

Authors: Hanchen Ye[None]; HyeGang Jun[None]; Deming Chen[None]

Efficient coarse-grained dataflow and well-organized data movement are essential for large-scale HLS (High-level Synthesis) designs to achieve promising performance and energy efficiency. Traditional HLS tools typically rely on designers to empirically specify the desired dataflow and data movement scheme and iteratively tune the design choices until reaching a satisfactory result. Although recent HLS optimization tools have achieved the automatic generation of dataflow implementation, these tools don't support a systematic representation of dataflow structures thus can only handle HLS designs without complicated hierarchies. Meanwhile, insufficient dataflow optimization and DSE (Design Space Exploration) often lead to sub-optimal and unscalable design solutions. To address the challenges, this poster proposes a scalable HLS framework called *ScaleFlow* that can explore the design space of large-scale dataflow and generate highly efficient HLS designs. ScaleFlow is built on top of a state-of-the-art compilation infrastructure called MLIR (Multi-Level Intermediate Representation) and proposes a new Dataflow MLIR dialect to model the multi-level dataflow hierarchy through a structural and abstracted representation. Meanwhile, in order to fully leverage the coarse-grained parallelism, the Dataflow dialect enables the functional level synchronization of off-chip memory accesses by explicitly modeling the streaming communications between dataflow nodes. On top of the new representation, ScaleFlow proposes a new DSE engine that decomposes the DSE problem into multiple levels according to the intrinsic dataflow hierarchy, and at each level, the DSE problem is further partitioned into local intra-node explorations and a global inter-node exploration. The hierarchical decomposition enables ScaleFlow to conduct a comprehensive and scalable search of the solution space given the design constraints.

Lightening talks / 22

Interaction Network Autoencoders vs Deep Neural Network Autoencoders for Anomaly Detection at the CMS Level One Trigger

Author: Andrew Skivington[1]

[1] *University of California-San Diego, Duarte Lab*

The Large Hadron Collider has recently started Run 3, which means protons are being collided at an astonishing energy of 13.6 TeV. Within the LHC is the CMS detector. Moreover, the Level one trigger is an integral part of the CMS detector and in many ways is considered the first responder of the High Level Trigger system. Its job is to make important initial cuts to the massive amounts of incoming data; therefore, it stores the data it was designed to flag as interesting for offline analysis and what is believed to be "uninteresting" data is thrown out forever and is never analyzed. However, what if within this "uninteresting" data lies some new physics unbeknownst to us, but the Level one trigger was able to pick out by deploying unsupervised machine learning algorithms onto the trigger? This is where autoencoders (AE) and variational autoencoders (VAE) enter the picture. The idea is that if the autoencoder is trained on standard collision events, then when the AE encounters events vastly different than the training data, the AE will produce a high reconstruction loss, which signifies an anomalous event that can be saved for offline analysis.

Currently, we are looking at two different AE and VAE architectures. One of the architectures is a deep neural network (DNN) VAE, meaning it is a fully connected network with dense Keras layers that make up the encoder and decoder of the AE and VAE, respectively. The other AE and VAE is an interaction network model, which is a kind of graphical neural network (GNN) AE that takes advantage of the natural graphical representation of particle collision data. The initial goal of the research is to determine which model, the DNN or GNN, is better for anomaly detection at Level one. Depending on the conclusion, a method to decrease the the computational resources used must be established in order to deploy the model on the Level one trigger. If the GNN proves to be the viable model, then the plan is to look to the method of knowledge distillation to train a student network, from a larger teacher network, which will meet the resource requirements without loss of generalization. Furthermore, one of these models could have the potential to be deployed on the Level one trigger to perform anomaly detection in the future.

Lightening talks / 23

Deep Learning Development and Deployment for Low-Latency Gravitational-Wave Astronomy

Author: Will Benoit[None]

The successful electromagnetic observation of the neutron star merger GW170817 led to explosive growth in the field of multi-messenger astronomy. With that growth has come new challenges and opportunities. The computational needs of gravitational-wave astronomy have risen alongside the sensitivity of the global network of gravitational-wave detectors, and will continue to rise as more detectors with even greater sensitivity come online in the next decade. As the scale of data ramps up, new techniques are desired that will allow for low-latency detection of gravitational-waves and enable multi-messenger followup. We present two deep learning networks that are being developed to address this demand: DeepClean and BBHnet. In combination, these networks form an end-to-end pipeline capable of denoising gravitational-wave strain data and detecting binary black hole mergers. We also present steps that have been taken in the development of these algorithms that will encourage their widespread adoption and use. Taking lessons from industry and the field of machine learning operations, tools and procedures have been created that simplify the process of consistently training, testing, and implementing machine learning networks. This lowers the barrier to entry for end users, and ensures that effective analysis tools are actually applied to important science questions.

24

Keynote speaker

Research Area / 25

Low-latency EM-Bright source property inference from GW data

Author: Deep Chatterjee[None]

The detection of the binary neutron star (BNS) merger, GW170817, was the first success story of multi-messenger observations of compact binary mergers. However, while the number of GW events have increased, there have been no joint electromagnetic counterparts detected since. A rapid assessment of properties that could lead to a counterpart is essential to aid time-sensitive follow-up operations, especially robotic telescopes. At minimum, this needs the possibility of a neutron star (NS). Also, the tidal disruption physics is important to determine the remnant matter post merger, the dynamics of which could result in the counterparts. The main challenge, however, is that the binary system parameters such as masses and spins estimated from the real-time, GW template-based searches are often dominated by statistical and systematic errors. Here, I'll present an application of supervised machine learning that was used in the third observing run to correct for and report EM-bright source properties in real-time GW discovery alerts.

Lightening talks / 26

Electromagnetic Counterpart Identification of Gravitational-wave candidates using deep-learning

Author: Deep Chatterjee[1]

[1] *Massachusetts Institute of Technology*

As gravitational-wave (GW) detectors become more sensitive and probe ever more distant reaches, the number of detected binary neutron star mergers will increase. However, detecting more events farther away with GWs does not guarantee corresponding increase in the number of electromagnetic counterparts of these events. Current and upcoming wide-field surveys that participate in GW follow-up operations will have to contend with distinguishing the kilonova from the ever increasing number of transients they detect, many of which will be consistent with the GW sky-localization. We have developed a novel tool based on a temporal convolutional neural network architecture, trained on sparse early-time photometry and contextual information for Electromagnetic Counterpart Identification (El-CID). The overarching goal for El-CID is to slice through list of new transient candidates that are consistent with the GW sky localization, and determine which sources are consistent with kilonovae, allowing limited target-of-opportunity resources to be used judiciously. In addition to verifying the performance of our algorithm on an extensive testing sample, we validate it on AT2017gfo - the only EM counterpart of a binary neutron star merger discovered to date - and AT2019npv - a supernova that was initially suspected as a counterpart of the gravitational-wave event, GW190814, but was later ruled out after further analysis.

Lightening talks / 27

Sleep Spindle as a Driver of Low Latency, Low Power ML in HLS4ML

Author: Xiaohan Liu[None]

A specific type of Electroencephalography (EEG) signal, sleep spindle, is believed to contribute to neuronal plasticity and memory consolidation. In this project, we proposed a system that is based on ultra-low latency and power FPGA to detect and interact with the sleep spindles to help neuroscientists to understand the mechanism behind the theory. The proposed system will have a programmed

FPGA and a headstage. The headstage will record the subject's brain signals and the FPGA will be connected with the headstage and process those signals to detect and interact with the sleep spindles.

This system is currently under development. We are working on implementing the HLS4ML to support the baseline deep learning model for this system. The baseline model is named Latent Factor Analysis via Dynamical Systems (LFADs). LFADs is an RNN variational autoencoder for analyzing spiking neural data. LFADs follows the encoder-decoder structure, and the main component for the LFADs'encoder is a bidirectional GRU, and the decoder is a unidirectional GRU. A dense layer is followed by the decoder to reduce the data to a set of low-dimensional temporal factors, which is LFADs latent. Another dense layer is followed by the latent information to create the log firing rate. The LFADs latent and the log firing rate are two important outputs from LFADs.

We faced several challenges during implementation. First, LFADs is a custom model, and this type of model is not supported by HLS4ML. We recreated the LFADs based on the Keras functional API and copied the weights and bias to test and ensure the recreated model has the same model architecture as the original one. Besides, the gaussian sampling and the bidirectional GRU layer are also not supported in HLS4ML. We removed the gaussian sampling from LFADs and found the performance does not decrease much based on the same dataset. Thus, we decided to have HLS4ML support this non-variational LFADs first and add the gaussian sampling back later. For the current stage, we are implementing the bidirectional GRU layer for HLS4ML and hope to finish it soon.

Lightening talks / 28

Transformer in HLS4ML

Authors: Zhixing "Ethan" Jiang[1]; Elham E Khoda[1]

Co-authors: Scott Hauck ; Shih-Chieh Hsu [2]

[1] *University of Washington (US)*

[2] *University of Washington Seattle (US)*

The transformer become widely use for the Natural language processing (NLP) task. Besides the NLP tasks, the transformer could be used to analyze any other time series signal processing data, such as images prediction, sensor detection, or glitch detection.
However, most of studies on transformer were implemented on GPU. The implementation of transformer on the FPGA was quite undiscovered. In this poster, we present an implementation of transformer on FPGA within the hls4ml framework.
We demonstrate how the transformer block, especially the multi-head attention layer, would be implemented on FPGA. We show a small transformer model as a benchmark to indicate the latency and resource usage on FPGA. Eventually, we discuss our next steps and expectation of the transformer on FPGA.