A3D3 All-hands meeting

Report of Contributions

https://indico.cern.ch/e/1405903

Contribution ID: **1** Type: **not specified**

Welcome

Monday 14 October 2024 09:00 (10 minutes)

Presenters: PARK, Seungbin; ZHANG, Yulei (University of Washington (US))

Session Classification: Morning Plenary

Contribution ID: **2** Type: **not specified**

Plenary Speaker 1

Presenter: Dr HAN, Song (MIT)

Session Classification: Morning Plenary

Contribution ID: **3** Type: **not specified**

Introduction to HDR ML Challenge

Session Classification: Morning Plenary

Contribution ID: **4** Type: **not specified**

Trainee-led: Overview of MMA

Presenter: BENOIT, Will

Session Classification: Morning Plenary

Contribution ID: **5** Type: **not specified**

Trainee-led: Overview of Neuro

Monday 14 October 2024 10:15 (15 minutes)

Presenter: PARK, Seungbin

Session Classification: Morning Plenary

Contribution ID: **6** Type: **not specified**

Trainee-led: Overview of HEP

Monday 14 October 2024 10:30 (15 minutes)

Presenter: CHOU, Yuan-Tang (University of Washington (US))

Session Classification: Morning Plenary

Contribution ID: **7** Type: **not specified**

Trainee-led: Overview of HAC

Presenter: MIAO, Siqi

Session Classification: Morning Plenary

Contribution ID: **8** Type: **poster**

Decoding multi-limb trajectories of naturalistic running from calcium imaging using deep learning

Monday 14 October 2024 17:15 (5 minutes)

Decoding neural activity into behaviorally-relevant variables such as speech or movement is an essential step in the development of brain-machine interfaces (BMIs)and can be used to clarify the role of distinct brain areas in relation to behavior. Two-photon (2p) calcium imaging provides access to thousands of neurons withsingle-cell resolution in genetically-defined populations and therefore is a promising tool for next-generation optical BMIs. However, decoding 2p calcium imagingrecordings into behavioral variables for use in real-time applications has traditionally been challenging due to the low sampling rate of the signal as well as the indirectand non-linear relationship between the underlying neural activity and the slow fluorescent signal. Here, we show an approach using deep learning to decode thenaturalistic multi-limb trajectories of running mice from neural recordings made with 2p calcium imaging over the sensorimotor cortex in a single hemisphere. Thework demonstrates the feasibility of using deep learning methods to identify and characterize populations of neurons that encode behaviorally-relevant variables. Thisapproach will be critical in the future implementation of neural decoding for next-generation optical BMIs that will improve the lives of patients suffering fromneurological injury and disease.

Author: PARK, Seungbin

Co-authors: DADARLAT, Maria; LIPTON, Megan Hope

Presenter: PARK, Seungbin

Session Classification: Poster Session / Reception

Contribution ID: **9** Type: **talk**

Implementation of ‘traccc‘ as-a-service

Monday 14 October 2024 14:15 (13 minutes)

Particle tracking at Large Hadron Collider (LHC) experiments is a crucial component of particle reconstruction, yet it remains one of the most computationally challenging tasks in this process. As we approach the High-Luminosity LHC era, the complexity of tracking is expected to increase significantly. Leveraging coprocessors such as GPUs presents a promising solution to the rising computational demands. The traccc project is a tracking demonstrator under the ACTS software designed to harness GPU resources for tracking. Despite promising initial results, the deployment of GPU algorithms such as traccc in production chains remains a significant challenge. In this talk, we present an as-a-service (aaS) approach to address these deployment challenges. A dedicated backend written to efficiently manage multiple concurrent requests from the server and load multiple model instances onto a dedicated GPU server is presented. Our results demonstrate increased resource utilization and a significant improvement in throughput compared to standalone traccc implementations.

Author: COCHRAN-BRANSON, Miles (University of Washington (US))

Co-authors: ZHAO, Haoran (University of Washington (US)); JU, Xiangyang (Lawrence Berkeley National Lab. (US)); CHOU, Yuan-Tang (University of Washington (US))

Presenter: COCHRAN-BRANSON, Miles (University of Washington (US))

Session Classification: Afternoon Plenary

Contribution ID: **11** Type: **talk**

Closed-loop Brain Stimulation in NHPs using hardware-accelerated Machine Learning

Monday 14 October 2024 16:09 (13 minutes)

Non-Human Primates (NHPs) are central to neuroscience research due to their complex behavioral interactions and physiological similarities to the human brain. A principal motivation behind the NHP research in the aoLab at the University of Washington is to understand and model neural circuits, which can be translated for practical applications for humans. However, the nonlinear interconnections within the brain challenge the isolation and study of specific neuronal structures during experiments, necessitating real-time stimulation of those regions.

Modern CPUs and GPUs cannot meet the millisecond-order latency constraint for real-time brain stimulation, in which the incoming data must be processed, passed through a mathematical model of the region, and encoded into stimulation patterns. To address this bottleneck, we integrate Field Programmable Gate Arrays (FPGAs) within an existing experimental apparatus using Neuropixels probes to collect spiking data from 300+ electrode channels. Additionally, we use the Latent Factor Analysis via Dynamical Systems (LFADS) architecture based on a sequential Variational Autoencoder (VAE) to model the dynamics of the brain region and predict future firing activity during experimentation.

Typically, passing a single batch of data (Channels x TimeSteps) for inference through LFADS requires 30.42ms on an Intel i7 CPU and 28.8ms on an NVIDIA A4000 GPU; exceeding the millisecond-order timeline of a single spike. The same process requires only 0.65ms on a Xilinx U55C FPGA, enabling the system to respond to inferred spike activity by stimulating the region. This gap in performance widens significantly with larger batches due to the multi-staged processing pipeline in the FPGA. However, the LFADS model weights were quantized due to limited hardware resources on the FPGA, reducing model accuracy from 91.3% for the full floating-point variant to 80.2% on test data. Even so, the system provides a platform to deploy and test hypotheses of functional and behavioral attributes of neuronal circuits by leveraging hardware acceleration.

Author: BOTADRA, Rajeev

Co-authors: ORSBORN, Amy; CHEN, ChiJui; SHLIZERMAN, Eli; SCHOLL, Leo; HAUCK, Scott; HSU, Shih-Chieh (University of Washington Seattle (US)); LE, Trung

Presenter: BOTADRA, Rajeev

Session Classification: Afternoon Plenary

Contribution ID: **12** Type: **poster**

Multi-messenger Astronomy Observations Via Alert Stream Filtering

Monday 14 October 2024 17:40 (5 minutes)

In this work we show advancements in follow-up methods for detection of electromagnetic counterparts to gravitational wave signals. These multi-messenger observations are important targets for their ability to unlock science including measurement of the Hubble constant, which is a current major effort in cosmology. In this work we include a data-driven heuristic to select anomalous flares which are candidate electromagnetic counterparts to the gravitational wave signals of merging black holes. This work also includes a systematic analysis of filtering done on telescope alert streams in order to identify astronomical transients, including these binary black hole merger flares. This analysis takes advantage of the six year history of observation and filtering on data from the Zwicky Transient Facility (ZTF). This work targets improvement of filter efficiency and interpretability, and also searches for interesting objects in the gaps between the range of filters operating on the ZTF alert stream. All of this work is done with the goal of minimizing followup latency for quickly evolving and fading signals. It is motivated by the upcoming Rubin Observatory, which will produce an alert stream that is magnitudes larger than that of ZTF.

Author: NOLAN, Kira (California Institute of Technology)

Co-author: Prof. GRAHAM, Matthew (California Institute of Technology)

Presenter: NOLAN, Kira (California Institute of Technology)

Session Classification: Poster Session / Reception

Contribution ID: **13** Type: **poster**

Estimating the Hubble Constant by combining posteriors from Multi-Messenger Kilonovae Observations

Monday 14 October 2024 17:30 (5 minutes)

Multi-Messenger observations of kilonovae can be used to measure the Hubble Constant by combining distance posteriors from the gravitational wave observations with a redshift measurement of the source's host galaxy. There is a significant discrepancy between two existing, prominent estimates of the Hubble constant: Planck, utilizing cosmic microwave background radiation and lambda cold dark matter (ΛCDM) cosmology, and SH0ES which uses supernovae as standard candles. Observations of gravitational waves and the associated kilonovae emitted from compact binary mergers also show promise for estimating the Hubble Constant. Here, we devise a framework to estimate the Hubble constant using simulated observations of neutron star mergers. We obtain a probability distribution of Hubble constant values for each event using Kernel Density Estimation (KDE) and Bayes'Theorem. We then tested various methods of KDE combination to arrive at our final model. We hope to put this framework into use with future observations in the current observing run and beyond, to arrive at an accurate measurement of the Hubble Constant.

Author: AVERILL, Megan

Co-author: TOIVONEN, Andrew (University of Minnesota)

Presenter: AVERILL, Megan

Session Classification: Poster Session / Reception

Contribution ID: **14** Type: **not specified**

Introducing A3D3 trainee-union

Session Classification: Morning Plenary

Contribution ID: **15** Type: **not specified**

Plenary Speaker 2

Session Classification: Morning Plenary

Contribution ID: **16** Type: **not specified**

Plenary Speaker 2

Presenter: LI, Pan

Session Classification: Morning Plenary

Contribution ID: **17** Type: **talk**

Combining Neural Networks for IceCube DeepCore Reconstruction and Classification Tasks

Monday 14 October 2024 14:41 (13 minutes)

IceCube DeepCore is an infill of the IceCube Neutrino Observatory designed to study neutrinos with energies as low as 5 GeV. Reconstruction and classification tasks near the lower energy threshold of IceCube DeepCore are especially difficult due to the low number of detected photons per neutrino event. Many neural networks have been developed for these tasks, and there are many ways we could consider training these neural networks to facilitate specific behaviors in the neural networks. In order to take advantage of this variety in architectures and training techniques we could use model stacking techniques with boosted decision trees (BDTs) to combine the outputs of multiple neural networks and improve overall performance. In this talk we show how this technique can improve inelasticity reconstruction with convolutional neural networks (CNNs) trained with different methods and improve neutrino flavor classification using two CNNs with different model architectures.

Author: PETERSON, Josh

Presenter: PETERSON, Josh

Session Classification: Afternoon Plenary

Contribution ID: **18** Type: **poster**

Noise or Astrophysical: Developing Machine Learning Classifiers for Characterizing Gravitational Wave Events

Monday 14 October 2024 17:20 (5 minutes)

The detection of gravitational waves with the Laser Interferometer Gravitational Wave Observatory (LIGO) has provided the tools to probe the furthest reaches of the universe. A rapid follow up to compact binary coalescence (CBC) events and their electromagnetic counterparts is crucial to find short lived transients. After a gravitational wave (GW) detection, another particular challenge is determining a fast and efficient way of characterizing events as astrophysical or terrestrial in origin. The mergers themselves provide many data products from low-latency CBC search pipelines which can aid in discerning whether or not a GW signal is astrophysical. We present an efficient low-latency method of alert classification by applying data products available in low-latency into three machine learning classification algorithms: Random Forest (RF), K-Nearest Neighbors (KNN), and Neural network (NN) using simulated event data from the Mock Data Challenge (MDC). We report the accuracy of the RF, KNN, and NN classifiers on the MDC events are 0.82, 0.84, and 0.89 respectively.

Author: TSUKAMOTO, Seiya

Presenter: TSUKAMOTO, Seiya

Session Classification: Poster Session / Reception

Contribution ID: **19** Type: **talk**

Neural Architecture Codesign for Fast Physics Applications

Monday 14 October 2024 15:30 (13 minutes)

We develop an automated pipeline to streamline neural architecture codesign for physics applications, to reduce the need for ML expertise when designing models for a novel task. Our method employs a two-stage neural architecture search (NAS) design to enhance these models, including hardware costs, leading to the discovery of more hardware-efficient neural architectures. The global search stage explores a wide range of architectures within a flexible and modular search space to identify promising candidate architectures. The local search stage further fine-tunes hyperparameters and applies compression techniques such as quantization aware training (QAT) and network pruning. We synthesize the optimal models to high level synthesis code for FPGA deployment with the hls4ml library. Additionally, our hierarchical search space provides greater flexibility in optimization, which can easily extend to other tasks and domains. We demonstrate this with two case studies: Bragg peak finding in materials science and jet classification in high energy physics.

Authors: DEMLER, Dmitri (UCSD); WEITZ, Jason (UCSD); DUARTE, Javier Mauricio (Univ. of California San Diego (US)); MCDERMOTT, Luke (UCSD); TRAN, Nhan (Fermi National Accelerator Lab. (US))

Presenter: WEITZ, Jason (UCSD)

Session Classification: Afternoon Plenary

Contribution ID: **20** Type: **poster**

Smart Pixels: A Machine Learning Approach Towards Data Reduction in Next-Generation Particle Detectors

Monday 14 October 2024 18:15 (13 minutes)

Pixel detectors are highly valuable for their precise measurement of charged particle trajectories. However, next-generation detectors will demand even smaller pixel sizes, resulting in extremely high data rates surpassing those at the HL-LHC. This necessitates a "smart"approach for processing incoming data, significantly reducing the data volume for a detector's trigger system to select interesting events. As charged particles pass through an array of pixel sensors, they leave behind clusters of deposited charge. The shape of these charge clusters can be useful, especially when fed into customized neural networks, which can extract the physical properties of the charged particle. These weights and biases of these neural networks can then be later implemented on-chip onto ASICs for installation at future pixel detectors.

We propose a "feature regression network", which uses TensorFlow and QKeras and takes as an input the 2-D shape of the charge clusters at different slices of time. These inputs are passed through a convolutional network and dense network to regress 14 quantities. As a result, we can predict the position (x, y), incidence angle (cot alpha, cot beta), and their covariance matrix. This customized model has been trained and evaluated on 7 different sets of pixel pitches (varying in x, y, and thickness) placed in a 13x21 pixel array, where their performance is analyzed through residual and pull study.

Author: JIANG, David (Univ. Illinois at Urbana Champaign (US))

Co-authors: GANDRAKOTA, Abhijith (Fermi National Accelerator Lab. (US)); BEAN, Alice (The University of Kansas (US)); BADEA, Anthony (University of Chicago (US)); DAS, Arghya Ranjan; PARPILLON, Benjamin (Fermi National Accelerator Lab. (US)); SYAL, Chinar (Fermi National Accelerator Lab. (US)); MILLS, Corrinne (University of Illinois at Chicago (US)); WEN, Dahai (Nanjing Normal University (CN)); SHEKAR, Danush (University of Illinois at Chicago (US)); BERRY, Douglas Ryan (Fermi National Accelerator Lab. (US)); HOWARD, Eliza Claire (University of Chicago (US)); FAHIM, Farah (Fermi National Accelerator Lab. (US)); Ms PRADHAN, Gauri; DI GUGLIELMO, Giuseppe (Fermilab); DICKINSON, Jennet Elizabeth (Cornell University (US)); Ms YOO, Jieun (UIC); HIRSCHAUER, Jim (Fermi National Accelerator Lab. (US)); DI PETRILLO, Karri Folan (University of Chicago); GRAY, Lindsey (Fermi National Accelerator Lab. (US)); VALENTIN, Manuel; NEUBAUER, Mark (Univ. Illinois at Urbana Champaign (US)); LIU, Miaoyuan (Purdue University (US)); WADUD, Mohammad Abrar (University of Illinois at Chicago (US)); SWARTZ, Morris (Johns Hopkins University (US)); TRAN, Nhan (Fermi National Accelerator Lab. (US)); MAKSIMOVIC, Petar (Johns Hopkins University (US)); KOVACH-FUENTES, Rachel Elizabeth (University of Chicago (US)); LIPTON, Ronald (Fermi National Accelerator Lab. (US)); KULKARNI, Shruti R (Oak Ridge National Laboratory)

Presenter: JIANG, David (Univ. Illinois at Urbana Champaign (US))

Session Classification: Poster Session / Reception

Contribution ID: **21** Type: **poster**

Accelerating Subglacial Bed Topography Prediction in Greenland: A Performance Evaluation of Spark-Optimized Machine Learning Models

Monday 14 October 2024 17:55 (5 minutes)

Accurate estimation of subglacial bed topography is crucial for understanding ice sheet dynamics and their responses to climate change. In this study, we employ machine learning models, enhanced with Spark parallelization, to predict subglacial bed elevation using surface attributes such as ice thickness, flow velocity, and surface elevation. Radar track data serves as ground truth for model validation. Our primary objective is to leverage Spark's distributed computing framework to accelerate model training and evaluation, enabling scalable analysis of large datasets. We tested several machine learning algorithms compatible with Spark, including XGBoost, Gradient Boosting (GBoost), Random Forest, and Kernel Regression. XGBoost emerged as the most efficient model, achieving substantial speed-ups as the number of computing nodes increased.

Our findings underscore the importance of distributed computing in enhancing the scalability and efficiency of machine learning models for large-scale climate data. The transition to the High-Performance Computing Facilities (HPCF), along with Spark parallelization, significantly reduced training time, demonstrating the effectiveness of distributed computing for complex datasets. This approach not only improves computational performance but also accelerates experimentation and analysis, contributing to a deeper understanding of ice sheet behavior and its implications for climate change. Future work will focus on applying these methods to even larger datasets, further leveraging Spark's capabilities to advance predictive modeling and support climate change mitigation and adaptation efforts. Moreover, we plan to explore deep neural networks and their performance on this application, leveraging multi-GPU architectures to further accelerate model training.

Author: CHAM, Mostafa (iHARP)

Co-authors: Mr SHAKERI, Ehsan (UMBC); Ms TABASSUM, Tartela (UMBC); Dr WANG, Jianwu (iHARP)

Presenter: CHAM, Mostafa (iHARP)

Session Classification: Poster Session / Reception

Contribution ID: **22** Type: **poster**

Foundation Model for Real-Time Model Selection and Fitting

Monday 14 October 2024 18:28 (13 minutes)

Fitting data to a variety of models is a fundamental challenge in the monitoring and control of dynamical systems across science and manufacturing domains. In this work, we present a compact foundation model designed for adaptive function selection and regression. The proposed architecture utilizes 1D convolutional neural networks (CNNs), augmented by physical constraints, to facilitate the selection and generation of functional forms that fit the underlying data. The model is capable of processing input data from spectroscopic measurements or time series and is grounded in a collection of 9 common mathematical functions derived from physical principles.

The data generation process for training involves randomly sampling parameters within predefined ranges for each function, evaluating these functions over a fixed interval of the independent variable (x). Models are trained on a large dataset with a sample size of 10,000 per function. The model is trained to accommodate up to five variable equations in the form f(x), allowing it to handle diverse datasets where different subsets may be optimally described by distinct functional forms. The architecture is inherently extensible to accommodate additional functions as needed.

Once trained, the model can estimate the parameters for multiple functions in a single forward pass, generating predictions by evaluating each function with its corresponding parameter set. A built-in model selection mechanism ensures that the most appropriate function is chosen to represent the data. This approach offers an alternative to traditional methods like symbolic regression and sparse optimization of nonlinear dynamics, both of which are notoriously difficult to optimize due to high computational complexity and non-convexity of the solution space.

We further evaluate the performance of this model under varying noise levels, showcasing its robustness through stochastic averaging during batch processing. Despite its compact size, with only ˜5,000 parameters, the model exhibits impressive scalability and performance improvements over time. Notably, the architecture is simple to convert and deploy on hardware platforms like FPGAs using the HLS4ML framework, making it highly suitable for real-time applications.

By leveraging this physics-constrained fitting framework, the model provides an efficient and lightweight solution for signal processing and real-time monitoring and control, paving the way for advanced, low-latency adaptive systems in a wide range of scientific and industrial applications.

Authors: TSAI, Cymberly; AGAR, Joshua

Presenter: TSAI, Cymberly

Session Classification: Poster Session / Reception

Contribution ID: **23** Type: **talk**

Trainee-led: Overview of MMA

Monday 14 October 2024 10:00 (15 minutes)

Multi-messenger astronomy is one of the pillars of A3D3. It spans optical, neutrino, and gravitational wave astronomy, each of which is a field with exciting physics and the potential to apply advanced machine learning techniques. In this presentation, I will give an overview of the work that is currently being done across all of the groups in this area.

Authors: QUEEN, Joshua; BENOIT, Will

Presenter: BENOIT, Will

Session Classification: Morning Plenary

Contribution ID: **24** Type: **talk**

ConNAS4ML: Gradient-Based NAS Framework with Hardware Constraints for hls4ml

Monday 14 October 2024 15:43 (13 minutes)

In software-hardware co-design, balancing performance with hardware constraints is critical, especially when using FPGAs for real-time applications in scientific fields with hls4ml. Limited resources and stringent latency requirements exacerbate this challenge. Existing frameworks such as AutoQKeras use Bayesian optimization to balance model size/energy, and accuracy, but they are time-consuming, rely on early-stage training, and often lead to inaccurate configuration evaluations, requiring significant trial and error. Additionally, these metrics often fail to reflect actual hardware usage.

In this work, we present **ConNAS4ML**, a **gradient-based**, **constraint-aware Neural Architecture Search (NAS) framework** for hardware-aware optimization within the hls4ml workflow. Our approach incorporates practical hardware resource metrics into the search process and dynamically adapts to different HLS designs, tool versions, and FPGA devices. Unlike AutoQKeras, ConNAS4ML performs simultaneous training and searching, requiring only minimal fine-tuning afterward. Users can either explore trade-offs between model performance and hardware usage or apply user-defined hardware constraints to ensure selected architectures stay within resource limits while maximizing performance. Key contributions include: (1) a user-friendly interface for customizing search space, hardware metrics, and constraints; (2) deep integration with hls4ml, allowing users to define and experiment with their own HLS synthesis configurations for FPGA; and (3) efficient hardware-aware optimization, exploring architectures under hardware constraints in a single shot manner, avoiding the time-consuming trial and error.

Preliminary results show our approach's effectiveness in two tasks. The first task optimized filter numbers of a 1.8M parameter CNN for energy reconstruction in calorimeters, achieving a 48.01% parameter reduction, and reductions in LUT usage (29.73%), FF (31.62%), BRAM (16.06%), and DSP (23.92%), with only a 0.84% increase in MAE after fine-tuning. The second task focuses on Jet Tagging classification using a precision search under various constraints. Even without fine-tuning, models stayed within constraints, with accuracy differences less than 0.37% from the baseline. Both tasks were efficient, with the architecture search taking 2 GPU hours and the precision search taking 0.26 GPU hours on one GPU. This framework can greatly accelerate FPGA deployment in resource-constrained environments, benefiting various fields beyond HEP such as edge computing and autonomous systems.

Author: CHEN, ChiJui

Co-author: LAI, Bo-Cheng

Presenter: CHEN, ChiJui

Session Classification: Afternoon Plenary

Contribution ID: **25** Type: **poster**

Finding Binary Black Hole Mergers

Monday 14 October 2024 18:05 (5 minutes)

Binary black hole mergers can be located by collecting and analyzing the unique gravitational wave signals they produce. Deep learning computational models, specifically Aframe, are used to identify and filter gravitational wave signals more accurately and in less time than traditional matched filtering analyses. The current machine learning model that we use, Aframe, was originally developed with only the LIGO detectors in mind. However, VIRGO detectors provide additional data from gravitational wave signals that could potentially enhance the detection of binary black hole mergers. We developed a model with Aframe incorporating both LIGO and VIRGO data. Our results showed that the addition of VIRGO data increased the detection of higher mass mergers, however, there was a decrease in the detection of lower mass mergers. This drop in performance for lower masses could be due to VIRGO running on a vastly different sensitivity than LIGO detectors which is something we are continuing to investigate.

Author: HENDERSON, Steven (UMN - Twin Cities)

Presenter: HENDERSON, Steven (UMN - Twin Cities)

Session Classification: Poster Session / Reception

Contribution ID: **26** Type: **poster**

Improving Sensitivity to Neutron Star Gravitational Wave Events using the Qp Transform

Monday 14 October 2024 17:45 (5 minutes)

Aframe is a gravitational wave search pipeline being constructed at the University of Minnesota Twin Cities. It is a machine learning pipeline designed to look for gravitational wave signals. It performs well for events with a high chirp mass, but could be better for ones with a lower chirp mass. My work is on improving Aframe's performance for these lower mass events using a new type of Q transform, the Qp transform.

Author: DE BRUIN, Emma

Presenter: DE BRUIN, Emma

Session Classification: Poster Session / Reception

Contribution ID: **27** Type: **poster**

Real-time compression of CMS detector data with machine learning

Monday 14 October 2024 17:35 (5 minutes)

The upcoming high-luminosity upgrade to the LHC will involve a dramatic increase in the number of simultaneous collisions delivered to the Compact Muon Solenoid (CMS) experiment. To deal with the increased number of simultaneous interactions per bunch crossing as well as the radiation damage to the current crystal ECAL endcaps, a radiation-hard high-granularity calorimeter (HGCAL) will be installed in the CMS detector. With its six million readout channels, the HGCAL will produce information on the energy and position of detected particles at a rate of 5 Pb/s. These data rates must be reduced by several orders of magnitude in a few microseconds in order to trigger on interesting physics events. We explore the application of machine learning for data compression performed by the HGCAL front-end electronics. We have implemented a conditional autoencoder which compresses data on the ECON-T ASIC before transmission off-detector to the rest of the trigger system.

Authors: PECZAK, Mariel; CREMONESI, Matteo (Carnegie-Mellon University (US)); WOODWARD, Nate; HARRIS, Philip Coleman (Massachusetts Inst. of Technology (US)); ROTHMAN, Simon (Massachusetts Inst. of Technology (US)); BALDWIN, Zachary Allen (CERN)

Presenter: PECZAK, Mariel

Session Classification: Poster Session / Reception

Contribution ID: **28** Type: **poster**

Early-time Classification of Astronomical Transients

Monday 14 October 2024 18:00 (5 minutes)

In time-domain astronomy, rapid classification of astronomical transients is critical for determining candidates for follow-up observations. With the advent of the Vera Rubin Observatory's Legacy Survey of Space and Time, the backlog of astronomical data will increase by terabytes a night. Machine learning models capable of processing and analyzing large quantities of data can advance the discovery process. Building upon astronet, a python package for classifying transients, this research moves towards a model capable of early-time classification of transients using a time-series transformer.

Author: BROWN, Alexandra Junell

Presenter: BROWN, Alexandra Junell

Session Classification: Poster Session / Reception

Contribution ID: **29** Type: **talk**

Channel tagging: identifying events with supernova pointing information

Monday 14 October 2024 14:28 (13 minutes)

Core collapse supernova explosions offer a rich potential of physics to explore. The emitted neutrinos are the first signals to reach the earth. Detecting these neutrinos and their direction can provide valuable information to optical detection systems in a multi messenger astronomy approach.
In liquid argon time projection chambers such as DUNE the charge interactions are the most abundant for supernova neutrinos, while only elastic neutrino electron scatters with an about ten times smaller cross section carry directly accessible pointing information. Thus it is crucial to be able to reliably distinguish between the two types of interactions. In my talk I will explore the idea of using a gradient boosted decision tree for this purpose.

Author: HAKENMUELLER, Janina Dorin (Duke University)

Presenter: HAKENMUELLER, Janina Dorin (Duke University)

Session Classification: Afternoon Plenary

Contribution ID: **30** Type: **talk**

NSF HDR ML Anomaly Detection Challenge

Monday 14 October 2024 11:00 (15 minutes)

Harnessing the Data Revolution (HDR), is an effort by the National Science Foundation (NSF) to promote the exploration of fundamental scientific questions using data-driven techniques. To raise interest in these approaches, and the HDR community, we have developed a Machine Learning (ML) challenge for anomaly detection, taking advantage of widespread data from several HDR institutes. This challenge seeks to connect these endeavors across several scientific disciplines, using a range of datasets spanning climate science, phylogenetics, materials science, and gravitational wave data from LIGO. The aim of participants is to design a novel ML model for their anomaly detection task. Using a single metric, their algorithm should detect anomalies across various datasets. We utilize the open-source benchmark ecosystem Codabench to host the challenge and ensure a Findable, Accessible, Interoperable, and Reusable (FAIR) dataset and workflow for participants from any community to contribute. In involving participating members from various communities, we promote collaboration and advance the broader goal of data-driven discovery.

Authors: GOVORKOVA, Katya (Massachusetts Inst. of Technology (US)); HARRIS, Philip Coleman (Massachusetts Inst. of Technology (US)); CHOU, Yuan-Tang (University of Washington (US))

Presenter: ANAND, Advaith (University of Washington (US))

Session Classification: Morning Plenary

Contribution ID: **31** Type: **poster**

Model Logging for FPGA Deployment

Monday 14 October 2024 17:50 (5 minutes)

Deploying lightweight models on FPGAs requires robust workflows for tracking, saving, and transferring model information, and ensuring that this information adheres to FAIR (Findable, Accessible, Interoperable, and Reproducible) principles. We present a Python package that automates the identification and documentation of key metadata for machine learning models developed in PyTorch or TensorFlow. This tool captures the model architecture, Python environment, and system hardware details, integrating this information with the code, visualizations, and checkpoints. The entire package is then uploaded to DataFed, a federated scientific data management system chosen for its enforcement of FAIR principles, making model data easily discoverable and shareable across collaborations.

Additionally, provenance data is embedded to explicitly track the progression of each model, ensuring traceability from the original code through each checkpoint. This is particularly advantageous for small, fast models deployed on FPGAs, where iteration speed and accountability are critical. By automating these processes, the package ensures that FPGA-based machine learning systems remain efficient, reproducible, and optimized for performance, all while adhering to open science standards for data management and collaboration.

Author: GODDY, Julian

Co-author: Dr AGAR, Joshua (Drexel University)

Presenter: GODDY, Julian

Session Classification: Poster Session / Reception

Contribution ID: **32** Type: **poster**

Detecting the Invisible: Electron Hit Localization in High-Resolution TEM Images Using Deep Learning on FPGAs

Monday 14 October 2024 17:25 (5 minutes)

Field-Programmable Gate Arrays (FPGAs) are increasingly becoming pivotal in the advancement of artificial intelligence (AI) and deep learning applications. Their unique architecture allows for customizable hardware acceleration, which is instrumental in handling the intensive computational demands of modern AI algorithms.

Transmission Electron Microscopy (TEM) provides exceptional high-resolution imaging capabilities essential for exploring materials at the atomic scale. However, real-time analysis of TEM data poses significant computational challenges due to the sheer volume and complexity of the generated images. We present a deep learning approach for object detection that accurately locates electron hits within high-resolution TEM images. Our model is trained on a curated dataset collected from TEM experiments, focusing on images containing up to three electron hits. Despite the initial limitation in electron count per image, the model demonstrates robust performance in accurately identifying and localizing electron events.

To bridge the gap between high computational demands and real-time processing requirements, we deploy the trained TensorFlow model onto Field-Programmable Gate Arrays (FPGAs). This deployment leverages the parallel processing capabilities of FPGAs, significantly accelerating inference times and enabling on-the-fly data analysis during TEM operations. The integration of our deep learning model with FPGA hardware showcases a scalable solution for real-time electron hit detection, potentially extending to images with higher electron counts in future work.

Our approach not only enhances the efficiency of TEM data analysis but also opens avenues for dynamic experimentation where immediate feedback is crucial. This fusion of high-resolution data acquisition with accelerated deep learning inference sets a new precedent for real-time computational microscopy.

Author: APPIAH OSEI, Derrick

Presenter: APPIAH OSEI, Derrick

Session Classification: Poster Session / Reception

Contribution ID: **33** Type: **poster**

Online track reconstruction with graph neural networks on FPGAs for the ATLAS experiment

Monday 14 October 2024 18:10 (5 minutes)

The next phase of high energy particle physics research at CERN will involve the High-Luminosity Large Hadron Collider (HL-LHC). In preparation for this phase, the ATLAS Trigger and Data AcQuisition (TDAQ) system will undergo upgrades to the online software tracking capabilities. Studies are underway to assess a heterogeneous computing farm deploying GPUs and/or FPGAs, together with the use of modern machine learning algorithms such as Graph Neural Networks (GNNs). We present a study on the reconstruction of tracks in the new all-silicon ATLAS Inner Tracker using GNNs on FPGAs for the Event Filter system. We explore each of the steps in a GNN-based tracking pipeline: graph construction, edge classification using an interaction network, and segmentation of the graph into track candidates. We investigate optimizations of the GNN approach that aim to minimize FPGA resources utilization and maximize throughput while retaining high track reconstruction efficiency and low fake rates required for the ATLAS Event Filter tracking system. These studies include model hyperparameter tuning, model pruning and quantization-aware training, and sequential processing of regions of the detector as graphs.

Author: BURLESON, Jared (University of Illinois at Urbana-Champaign)

Presenter: BURLESON, Jared (University of Illinois at Urbana-Champaign)

Session Classification: Poster Session / Reception

Contribution ID: **34** Type: **not specified**

talk1

Session Classification: Afternoon Plenary

Contribution ID: **35**

Type: **not specified**

talk1

Session Classification: Afternoon Plenary

Contribution ID: **36** Type: **not specified**

Postbac Research Experience: Multimodal Classification of Astronomical Transients

Monday 14 October 2024 14:05 (10 minutes)

The Zwicky Transient Facility is capable of triggering hundreds of thousands of alerts a night for potential astronomical transients. Quickly classifying these objects is critical for determining candidates for follow up observations. Machine learning already plays a role in the transient classification pipeline, and has shown success training on image time series and photometric time series. For the start of my A3D3 Postbac at the University of Minnesota, I have been working on a multimodal classifier of astronomical transients using photometric time series data, image time series data, and metadata.

Presenter: BROWN, Alexandra Junell

Session Classification: Afternoon Plenary

Contribution ID: **37** Type: **not specified**

Real-time Anomaly Detection at the CMS L1 Trigger

Monday 14 October 2024 14:54 (13 minutes)

Presenter: QUINNAN, Melissa (Univ. of California San Diego (US))

Session Classification: Afternoon Plenary

Contribution ID: **38** Type: **not specified**

talk placeholder

Session Classification: Afternoon Plenary

Contribution ID: **39** Type: **not specified**

Real-time Anomaly Detection at the CMS L1 Trigger

Presenter: QUINNAN, Melissa (Univ. of California San Diego (US))

Session Classification: Afternoon Plenary

Contribution ID: **40** Type: **talk**

Efficient Multi-modal LLM on the Edge

Monday 14 October 2024 09:10 (25 minutes)

This talk presents efficient multi-modal LLM innovations with algorithm and system co-design. I'll first present VILA, a visual language model deployable on the edge. It is capable of visual in-context learning, multi-image reasoning, video captioning and video QA. Followed by SmoothQuant and AWQ for LLM quantization, which enables VILA deployable on edge devices, bringing new capabilities for mobile vision applications. Second, I'll present StreamingLLM, a KV cache optimization technique for long conversation and QUEST, leveraging sparsity for KV cache compression.

Author: Dr HAN, Song (MIT)

Presenter: Dr HAN, Song (MIT)

Session Classification: Morning Plenary

Contribution ID: **41** Type: **talk**

Autoencoders for Anomaly Detection and Output Reduction on the Edge (AADORE)

Monday 14 October 2024 15:56 (13 minutes)

Detectors at next-generation high-energy physics experiments face several daunting requirements: high data rates, damaging radiation exposure, and stringent constraints on power, space, and latency. To address these challenges, machine learning (ML) in readout
electronics can be leveraged for smart detector designs, enabling intelligent inference and data reduction at-source. Autoencoders offer a variety of benefits for front-end readout; an on-sensor encoder can perform efficient lossy data compression while simultaneously
providing a latent space representation that can be used for anomaly detection. Results are presented from low-latency and resource-efficient autoencoders for front-end data processing in a futuristic silicon pixel detector. Encoder-based data compression
is found to preserve good performance of off-detector analysis while significantly reducing the off-detector data rate as compared to a similarly sized data filtering approach. Furthermore, the latent space information is found to be a useful discriminator
in the context of real-time sensor defect monitoring. Together these results highlight the multi-faceted utility of autoencoder-based front-end readout schemes, and motivate their consideration in future detector designs.

Author: YUE, Alexander

Presenter: YUE, Alexander

Session Classification: Afternoon Plenary

Contribution ID: **42** Type: **talk**

Postbac research experience: light from black holes and other multi-messenger astronomy

Monday 14 October 2024 13:45 (10 minutes)

In astronomy, the successful identification of electromagnetic counterparts to gravitational wave signals unlocks unique science that is otherwise impossible with siloed observations. Efforts in multi-messenger astronomy stand to become increasingly fruitful
but also more complex over the next decade as new instruments provide exponentially larger data streams. Not only does this work require filtering large data streams, but also it calls for prompt follow-up which can catch rapidly fading targets. I present
my work on the follow-up infrastructure for these quickly evolving and in some cases largely unmodeled signals, which has focused on enabling fast response necessary to collect these multi-messenger observations. I also share my experience doing research as
a postbac, and how this opportunity has impacted my career.

Author: NOLAN, Kira

Presenter: NOLAN, Kira

Session Classification: Afternoon Plenary

Contribution ID: **43** Type: **talk**

Attention at Silicon Speed: Towards Efficient Transformers on FPGAs

Monday 14 October 2024 13:55 (10 minutes)

The High-Luminosity Large Hadron Collider (HL-LHC), anticipated to begin operations in 2029, will generate data at an astounding rate on the order of 100 terabits per second. To efficiently process and filter these data, the Compact Muon Solenoid (CMS) experiment
relies on the extremely low-latency Level-1 trigger, which uses Field-Programmable Gate Arrays (FPGAs). My project focuses on further optimizing this process by adapting efficient transformer models for implementation on FPGAs using the hls4ml package. This
talk will highlight my contributions, ongoing work to optimize FPGA implementations of transformer models, and my experiences as an A3D3 Postbaccalaureate Fellow.

Author: FLYNN, Rian (Purdue University (US))

Presenter: FLYNN, Rian (Purdue University (US))

Session Classification: Afternoon Plenary

Contribution ID: **44** Type: **talk**

Trainee-led: Overview of HAC

Monday 14 October 2024 10:45 (15 minutes)

Algorithm and Hardware Co-Development (HAC) is a key focus area within A3D3, supporting the institute's mission to build accelerated AI solutions for scientific discovery. Our team develops AI algorithms to address significant challenges in data-driven research, including data irregularity, label scarcity, and the complexity of understanding AI models, while also providing efficient hardware implementations to optimize these algorithms. In this presentation, I will provide an overview of the developments in HAC, including our efforts to meet strict computational requirements such as low latency and high throughput. I will also discuss our progress in developing hardware design automation tools, which enable domain experts to more easily integrate their AI models into hardware platforms. Together, these innovations form the foundation for supporting various fields, from high-energy physics to neuroscience.

Author: MIAO, Siqi

Presenter: MIAO, Siqi

Session Classification: Morning Plenary

Contribution ID: **45** Type: **talk**

Distribution Shifts in Graph Machine Learning and Graph Domain Adaptation

Monday 14 October 2024 09:35 (25 minutes)

Graphs have been widely applied to model intricate relationships among entities. The application of Graph Machine Learning (GML) to enhance prediction capabilities for graph-structured data is prevalent in several scientific disciplines, such as particle physics, material science, and biology. However, applications in these domains often present challenges due to changes in data distributions due to the label collection process they employ. Specifically, the data used for model training often comes from the thoroughly investigated regimes, whose distributions often do not align well with the under-explored regime of scientific interest. Furthermore, the interconnected nature of entities in a graph presents an additional level of complexity, making current distributionally robust methods suboptimal when being applied to graph data.

This presentation will focus on our recent studies on GML under distribution shifts. Our studies are motivated by the observation of the data distribution shift in particle physics. We propose a method named graph structure alignment. The key idea of our approach is to estimate and quantify shifts in entity connection patterns from the training phase to real-world evaluation. Consequently, the influence of neighboring entities on a central node can be appropriately calibrated based on prior estimations, serving to mitigate the distribution shift in graph data.

Author: LI, Pan

Presenter: LI, Pan

Session Classification: Morning Plenary

Contribution ID: **46** Type: **not specified**

Dr. Zhijian Liu

Monday 14 October 2024 11:45 (5 minutes)

Presenter: LIU, Zhijian (Massachusetts Institute of Technology)

Session Classification: Morning Parallels

Contribution ID: **47** Type: **not specified**

Dr. Julia Gonski

Monday 14 October 2024 11:50 (5 minutes)

Presenter: GONSKI, Julia Lynne (SLAC National Accelerator Laboratory (US))

Session Classification: Morning Parallels

Contribution ID: **48** Type: **not specified**

Dr. Dylan Rankin

Monday 14 October 2024 11:55 (5 minutes)

Presenter: RANKIN, Dylan Sheldon (University of Pennsylvania (US))

Session Classification: Morning Parallels

Contribution ID: **49** Type: **not specified**

Panel Discusion

Monday 14 October 2024 12:00 (45 minutes)

Presenters: RANKIN, Dylan Sheldon (University of Pennsylvania (US)); GONSKI, Julia Lynne (SLAC National Accelerator Laboratory (US)); LIU, Zhijian (Massachusetts Institute of Technology)

Session Classification: Morning Parallels