Fast Machine Learning for Science Workshop 2022

Monday, 3 October 2022 - Thursday, 6 October 2022

Southern Methodist University

Book of Abstracts

Contents

Fast Dynamical System Modelling Forecasting 1

Quantized Distilled Autoencoder Model on FPGA for Real-Time Crystal Structure Detection in 4D Scanning Transmission Electron Microscopy 1

Rapid Fitting of Band-Excitation Piezoresponse Force Microscopy Using Physics Constrained Unsupervised Neural Networks 1

Extremely Noisy 4D-TEM Strain Mapping Using Cycle Consistent Spatial Transforming Autoencoders 3

Data Driven Weather Forecasting with Rudimentary Observables 3

End-to-End Vertex Finding for the CMS Level-1 Trigger 4

Deployment of ML in changing environments 4

Neural network accelerator for quantum control 5

Online-compatible Unsupervised Non-resonant Anomaly Detection 5

FastML Science Benchmarks: Accelerating Real-Time Scientific Edge Machine Learning 6

Exploring FPGA in-storage computing for Supernova Burst detection in LArTPCs 6

Low-latency Calorimetry Clustering at the LHC with SPVCNN 7

A Deep Learning Approach to Particle Identification for the AMS Electromagnetic Calorimeter 7

Intelligent experiments through real-time AI: Fast Data Processing and Autonomous Detector Control for sPHENIX and future EIC detectors 8

Quantized ONNX (QONNX) 9

Accelerating JEDI-net for jet tagging on FPGAs 10

Harnessing ultrafast ML for new algorithms at the CMS L1 trigger 10

Resource Efficient and Low Latency GNN-based Particle Tracking on FPGA 11

Implementation of a pattern recognition neural network for live reconstruction using AI processors 11

Real-time image processing for high-resolution imaging detectors 12

CryoAI –Prototyping cryogenic chips for machine learning at 22nm 12

Autonomous real-time science-driven follow-up of survey transients 13

Fast recurrent neural networks on FPGAs with hls4ml 13

Low-latency Noise Subtraction of Gravitational Wave Data by DeepClean 14

A novel ML-based method of primary vertex reconstruction in high pile-up condition . . 14

A Normalized Autoencoder for LHC triggers 15

Designing intelligent DAQ systems for radiation instrumentation with hls4ml 15

Large CNN for HLS4ML and Deepcalo 16

Demonstration of Machine Learning-assisted real-time noise regression in LIGO 16

End-to-end acceleration of machine learning in gravitational wave physics 17

Autonomous experiments in scanning probe microscopy: opportunities for rapid inference and decision making 17

Exa.TrkX inference as-a-service 18

Application of deep learning to instability tracking using high-speed video cameras in magnetic confinement fusion 18

FKeras: A Fault Tolerance Library for DNNs 19

Increasing the LHC Computational Power by integrating GPUs as a service 20

Next Generation Coprocessors as a service 20

Interaction Network Autoencoder in the Level-1 Trigger 21

EMD Neural Network Loss for ECON-T ASIC Autoencoder 21

Robust anomaly detection using NuRD 22

Quantized Neural Networks on FPGAs using HAWQ-V3 and hl4ml 22

Design and first test results of a reconfigurable autoencoder on an ASIC for data compression at the HL-LHC 23

In-System Parameter Update and I/O Capture for Machine Learning IP Cores 23

Neural Signal Compression System for a Seizure-Predicting Brain Implant in CMOS 28nm 24

Rapid Generation of Kilonova Light Curves Using Conditional Variational Autoencoder . 25

In-Pixel AI: From Algorithm to Accelerator 25

Detection for Core-collapse Supernova and Fast Data Preprocessing 26

Semi-supervised Graph Neural Networks for Pileup Noise Removal 26

Welcome 27

Machine Learning for Science . . . 27

Efficient Computer Architectures . . . 27

Applications and Opportunities at the Large Hadron Collider . . . 27

Application and Opportunities in Nuclear Physics . . . 27

Benchmarking Fast ML systems . . . 27

Applications and Opportunities for Machine Controls . . . 27

Applications and Opportunities in Fusion Systems . . . 27

Autonomous experiments in scanning probe microscopy: opportunities for rapid inference and decision making . . . 28

Toward Real Time Decision-Making in Material Science Experiments . . . 28

Some current challenges in materials measurements: an industrial perspective . . . 28

Beyond CMOS . . . 28

Applications and Opportunities in Neutrino and DM experiments . . . 28

The need for low latency with billion-year old signals . . . 28

A Machine Learning Software Infrastructure for Gravitational Wave Signal Discovery . . . 28

hls4ml tutorial . . . 29

hls4ml community session . . . 29

Deep Neural Network Algorithms in the CMS Level-1 Trigger . . . 29

Implementing Deep Neural Network Algorithms inside the CMS Level-1 Trigger . . . 29

A Machine Learning Software Infrastructure for Gravitational Wave Signal Discovery . . . 29

Contributed Talks / 1

Fast Dynamical System Modelling Forecasting

Author: Randall Clark[1]

[1] *University of California San Diego Physics Department (PhD Student)*

Corresponding Author: r2clark@ucsd.edu

In our study of (usually chaotic) Dynamical Systems, we invented a method we call Data Driven Forecasting, or DDF, that can take observed data, recreate an approximate form to the original model with a sum of radial basis functions, and rapidly forecast the future behavior of the system. This method is faster than 4th Order Runge-Kutta, so even if the user has knowledge of the data sets original model, this surrogate model could out-speed the standard way of integrating the state of the system forward in time.

Our work includes example applications of DDF to Zebra Finch Neuron data of the voltage in response to external stimuli. DDF learns from the behavior of the voltage from the external stimuli and replicates the behavior of the neuron. With this we have built what we call a "DDF Neuron", this rapid model exhibits the behavior a neuron for a wide array of external stimuli, not just the stimuli that was trained on. Our DDF Neurons advance in time faster than Runge-Kutta methods and learn directly from observed data whereas Runge-Kutta methods must rely on a model that can be large and slow to solve.

DDF can bring rapid time series integration and forecasting to systems with complex models and incomplete observations using time delay embedding to reconstruct the state space. DDF is able to do all this with the data alone and with no knowledge of the underling dynamical model associated with the data.

Contributed Talks / 2

Quantized Distilled Autoencoder Model on FPGA for Real-Time Crystal Structure Detection in 4D Scanning Transmission Electron Microscopy

Authors: Colin Ophus[1]; Javier Mauricio Duarte[2]; Joshua Agar[3]; Jules Muhizi[4]; Nhan Tran[5]; Ryan Forelli[3]; Shuyu Qin[6]

[1] *Lawrence Berkeley National Laboratory*

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

[3] *Lehigh University*

[4] *Fermilab/Harvard University*

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

[6] *Drexel University*

Corresponding Author: rff224@lehigh.edu

https://drive.google.com/file/d/1ldXiChVPOIencQJqCmTaITqLzIdBZ-rg/view?usp=sharing

Contributed Talks / 3

Rapid Fitting of Band-Excitation Piezoresponse Force Microscopy Using Physics Constrained Unsupervised Neural Networks

Authors: Alibek Kaliyev[1]; Joshua Agar[1]

Co-authors: Ryan Forelli [1]; Pedro Sales [2]; Shuyu Qin [3]; Yichen Guo [1]; Olugbodi Oluwafolajinmi ; Seda Memik [4]; Michael Mahoney [5]; Amir Gholami [5]; Rama Vasudevan [6]; Stephen Jesse [6]; Nhan Tran [7]; Philip Coleman Harris [8]; Martin Takáč [1]

[1] *Lehigh University*

[2] *Massachusetts Institute of Technology*

[3] *Drexel University*

[4] *Northwestern University*

[5] *University of California, Berkeley*

[6] *Oak Ridge National Laboratory*

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

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

Corresponding Author: alk224@lehigh.edu

Imaging nanoscale dynamics and response in materials requires imaging techniques with high spatial and temporal resolution. To meet this need, various scanning-probe spectroscopic imaging modes have emerged to understand electrochemical and ionic mobility and dynamics, ferroelectric switching dynamics, and dynamics mechanical responses of materials under external perturbations. These techniques collect large, high-dimensional data that is difficult and time-consuming to analyze. Practically, most analysis happens long after the experiments have been completed. This hinders researchers'abilities to use real-time feedback to conduct experiments on sensitive samples with a creative inquiry.

Machine learning techniques like principle component analysis (PCA), linear and non-linear clustering and non-negative matrix factorization have accelerated analysis. However, these techniques are computationally inefficient, highly dependent on prior estimates, and unable to interpret some complex features physically.

We developed a fully unsupervised deep neural network (DNN) that can be constrained to a known empirical governing equation. This is achieved by training an encoder to predict the model parameters, which are decoded by the underlying empirical expression. As long as the empirical expression is differentiable, it can be trained using stochastic gradient descent.

We evaluate this concept on a benchmark band-excitation piezoresponse force microscopy (BE-PFM) to predict amplitude, phase, cantilever resonance frequency, and dissipation from a simple harmonic oscillator (SHO) model. To extract further insights from piezoelectric hysteresis loops, which were calculated from fit results, it is common to fit the loops to a 9-parameter empirical function that extracts parameters related to the shape of the loop.

We demonstrated several important breakthroughs:

1. Speed –we can train our model to conduct 1.38 million SHO fits in less than 5 minutes and can conduct inference in <3 seconds with a batch size of 1024 on free computing resources (PCIe P100 on Google Colab);
2. Robustness –We demonstrate that SHO and hysteresis loop fit results have narrower and more physically reasonable distributions than least-square fitting (LSQF) results;
3. Signal-to-noise ratio –Our model performs well and provides physically interpretable results on artificially noisy data where well-designed conventional LSQF pipelines fail;
4. Real-time –We conduct quantized-aware training to deploy this model on an FPGA. Simulations predict streaming inference at <50 μs, orders of magnitude faster than the data acquisition and sufficiently fast for real-time control of automated experiments.

This work provides an automated methodology to develop physics-conforming, robust, fast approximation of noisy data with real-time (sub-ms) streaming inference. We demonstrate the efficacy of this methodology on a benchmark BE-PFM dataset. However, the approach broadly applies to spectroscopic fitting when the empirical expression is known. This approach provides a pathway

for real-time interpretation and controls from high-velocity data sources ubiquitous in experimental science.

Contributed Talks / 4

Extremely Noisy 4D-TEM Strain Mapping Using Cycle Consistent Spatial Transforming Autoencoders

Author: Shuyu Qin[None]

Co-author: Joshua Agar [1]

[1] *Drexel University*

Corresponding Author: shq219@lehigh.edu

Atomic-scale imaging of 2D and quantum materials benefits from precisely extracting crystallographic strain, shear, and rotation to understand their mechanical, optical and electronic properties. One powerful technique is 4D-STEM (4-dimensional scanning transmission electron microscopy), where a convergent electron beam is scanned across a sample while measuring the resulting diffraction pattern with a direct electron detector. Extracting the crystallographic strain, shear, and rotation from this data relies either on correlation strain measurement method (e.g., implemented in py4DSTEM) or determining the center of mass (CoM) of the diffraction peaks. These algorithms have limitations. They require manual preprocessing and hyperparameter tuning, are sensitive to signal-to-noise, and generally are difficult to automate. There is no one-size-fits-all algorithm.
Recently, machine learning techniques have been used to assist in analyzing 4D-STEM data, however, these models do not possess the capacity to learn the strain, rotation, or translation instead they just learn an approximation that almost aways tends to be correct as long as the test examples are within the training dataset distribution.
We developed a novel neural network structure –Cycle Consistent Spatial Transforming Autoencoder (CC-ST-AE). This model takes a set of diffraction images and trains a sparse autoencoder to classify an observed diffraction pattern to a dictionary of learned "averaged"diffraction patterns. Secondly, it learns the affine transformation matrix parameters that minimizes the reconstruction error between the dictionary and the input diffraction pattern. Since the affine transformation includes translation, strain, shear, and rotation, we can parsimoniously learn the strain tensor. To ensure the model is physics conforming, we train the model cycle consistently, by ensuring the inverse affine transformation from the dictionary results in the original diffraction pattern.
We validated this model on a number of benchmark tasks including: A Simulated 4D TEM data of WS2 and WSe2 lateral heterostructures (noise free) with a ground truth of the strain, rotation and shear parameters. Secondly, we test this model experimental 4D STEM on 2D-heterostructures of tungsten disulfide (WS2) and tungsten diselenide (WSe2).
This model shows several significant improvements including: 1. When tested on simulated data, the model can recover the ground truth with minimal error. 2. The model can learn the rotation and strain on noisy diffraction patterns where CoM failed, and outperforms correlation strain measurement method. 3. Our model can accommodate large and continuous rotations difficult to obtain with other methods. 4. Our model is more robust to noisy data. 5. Our model can map the strain, shear and rotation; identify dislocation and ripples; and distinguish background and sample area automatically.
Ultimately, this work demonstrates how embedding physical concepts into unsupervised neural networks can simplify, automate, and accelerate analysis pipelines while simultaneously leveraging stochastic averaging that improves robustness on noisy data. This algorithmic concept can be extended to include other physical phenomena (e.g., polarization, sample tilt), can be used in automated experiments, and can be applied to other applications in materials characterization.

Contributed Talks / 5

Data Driven Weather Forecasting with Rudimentary Observables

Author: Henry Abarbanel[1]

Co-author: Luke Fairbanks [1]

[1] *ucsd*

Corresponding Author: 143050fairbanks@gmail.com

Weather forecasting is currently dominated by a handful of centralized institutions running computationally intensive, high-dimensional numerical models over the entire global grid before disseminating results. By leveraging data-driven forecasting principles one may train a machine learning system to use simple measurements such as wind speed, pressure, and temperature to forecast those same observables with reasonable accuracy and less compute. Evidence points to this scheme working for localized measurements rather than needing data from across the globe, enabling a distributed, real-time forecasting system to bolster the traditional predictions, conduct re-analysis, and empower institutional decision-makers.

Contributed Talks / 6

End-to-End Vertex Finding for the CMS Level-1 Trigger

Author: Christopher Edward Brown[1]

Co-authors: Aaron Bundock [2]; Alex Tapper [3]; Benjamin Radburn-Smith [1]; Matthias Komm [4]; Maurizio Pierini [5]; Sioni Paris Summers [5]; Vladimir Loncar [5]

[1] *Imperial College (GB)*

[2] *University of Bristol (GB)*

[3] *Imperial College London*

[4] *Deutsches Elektronen-Synchrotron (DE)*

[5] *CERN*

Corresponding Authors: benjamin.radburn-smith@cern.ch, christopher.eward.brown@cern.ch

The High Luminosity LHC provides a challenging environment for fast trigger algorithms; increased numbers of proton-proton interactions per collision will introduce more background energy in the detectors making triggering on interesting physics signatures more challenging. To help mitigate the effect of this higher background the highest energy interaction in an event can be found and other detector signatures can be associated with it. This primary vertex finding at the CMS Level-1 trigger will be performed within a latency of 250 ns. This work presents an end-to-end neural network based approach to vertex finding and track to vertex association. The network possesses simultaneous knowledge of all stages in the reconstruction chain, which allows for end-to-end optimisation. A quantised and pruned version of the neural network, split into three separate sub networks, is deployed on an FPGA using the hls4ml tools rerun through Xilinx vitis hls to take advantage of optimised pipelining. A custom hls4ml tool for convolutional neural networks that allows fully parallel input is used to ensure the strict latency requirements are met. Hardware demonstration of the network on a prototype Level-1 trigger processing board will also be shown.

Contributed Talks / 7

Deployment of ML in changing environments

Author: Benjamin Radburn-Smith[1]

Co-authors: Alex Tapper [2]; Christopher Edward Brown [1]; Marco Barbone ; Matthias Komm [3]; Sioni Paris Summers [4]

[1] *Imperial College (GB)*

[2] *Imperial College London*

[3] *Deutsches Elektronen-Synchrotron (DE)*

[4] *CERN*

Corresponding Authors: christopher.eward.brown@cern.ch, benjamin.radburn-smith@cern.ch, m.barbone19@imperial.ac.uk

The deployment of fast ML models for on-detector inference is rapidly growing but faces key issues. One such issue is the difference between the training environment and the "real-world"environment in deployment giving unknown errors in inference. Examples of this include training a model on an abundance of well understood simulated data but deploying it on a real and imperfect detector or on a detector in which the performance changes over time that the ML model is unaware of. Various techniques can be employed to mitigate this issue including the use of uncertainty quantification to better understand the inference errors, retraining and redeploying models with new data or the use of continual learning where a model is continually updated with a stream of evolving data. This issue of deploying ML models in changing environments is presented as are the pros and cons of potential solutions.

Contributed Talks / 8

Neural network accelerator for quantum control

Authors: Baris Ozguler[1]; David Xu[2]; Farah Fahim[3]; Gabriel Perdue[3]; Giuseppe Di Guglielmo[3]; Luca Carloni[2]; Nhan Tran[3]

[1] *Fer*

[2] *Columbia University*

[3] *Fermilab*

Corresponding Author: gdg@fnal.gov

Efficient quantum control is necessary for practical quantum computing implementations with current technologies. However, conventional algorithms for determining optimal control parameters are computationally expensive, mainly excluding them from use outside of the simulation. Furthermore, existing hardware solutions structured as lookup tables are imprecise and costly. A more efficient method can be produced by designing a machine learning model to approximate the results of traditional tools. Such a model can then be synthesized into a hardware accelerator for quantum systems. Our study demonstrates a machine learning algorithm for predicting optimal pulse parameters. This algorithm is lightweight enough to fit on a low-resource FPGA and perform inference with a latency of 175ns and pipeline interval of 5ns with gate fidelity greater than 0.99. In the long term, such an accelerator could be used near quantum computing hardware where traditional computers cannot operate, enabling quantum control at a reasonable cost at low latencies without incurring large data bandwidths outside the cryogenic environment.

Contributed Talks / 9

Online-compatible Unsupervised Non-resonant Anomaly Detection

Authors: Ben Nachman[1]; David Shih[None]; Vinicius Massami Mikuni[1]

[1] *Lawrence Berkeley National Lab. (US)*

Corresponding Author: vinicius.massami.mikuni@cern.ch

There is a growing need for anomaly detection methods that can broaden the search for new particles in a model-agnostic manner. Most proposals for new methods focus exclusively on signal sensitivity. However, it is not enough to select anomalous events - there must also be a strategy to provide context to the selected events. We propose the first complete strategy for unsupervised detection of non-resonant anomalies that includes both signal sensitivity and a data-driven method for background estimation. Our technique is built out of two simultaneously-trained autoencoders that are forced to be decorrelated from each other. This method can be deployed offline for non-resonant anomaly detection and is also the first complete online-compatible anomaly detection strategy. We show that our method achieves excellent performance on a variety of signals prepared for the ADC2021 data challenge.

Contributed Talks / 10

FastML Science Benchmarks: Accelerating Real-Time Scientific Edge Machine Learning

Authors: Benjamin Hawks[1]; Christian Herwig[2]; Javier Mauricio Duarte[3]; Jules Muhizi[4]; Nhan Tran[2]; Shvetank Prakash[5]; Vijay Janapa Reddi[5]

[1] *Fermi National Accelerator Lab*

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

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

[4] *Fermilab/Harvard University*

[5] *Harvard University*

Applications of machine learning (ML) are growing by the day for many unique and challenging scientific applications. However, a crucial challenge facing these applications is their need for ultra low-latency and on-detector ML capabilities. Given the slowdown in Moore's law and Dennard scaling, coupled with the rapid advances in scientific instrumentation that is resulting in growing data rates, there is a need for ultra-fast ML at the extreme edge. Fast ML at the edge is essential for reducing and filtering scientific data in real-time to accelerate science experimentation and enable more profound insights. To accelerate real-time scientific edge ML hardware and software solutions, we need well-constrained benchmark tasks with enough specifications to be generically applicable and accessible. These benchmarks can guide the design of future edge ML hardware for scientific applications capable of meeting the nanosecond and microsecond level latency requirements. To this end, we present an initial set of scientific ML benchmarks, covering a variety of ML and embedded system techniques.

Contributed Talks / 11

Exploring FPGA in-storage computing for Supernova Burst detection in LArTPCs

Authors: Benjamin Hawks[1]; Jieran Shen[2]; Jovan Mitrevski[3]; Kate Scholberg[2]; Michael Wang[None]; Nhan Tran[3]; Pengfei Ding[4]; Tejin Cai[5]; Tingjun Yang[3]; Tom Junk[3]

[1] *Fermi National Accelerator Lab*

[2] *Duke University*

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

[4] *Fermi National Accelerator Laboratory*

[5] *York University*

Corresponding Author: bhawks@fnal.gov

Neutrino detectors, such as the Deep Underground Neutrino Experiment (DUNE) "far detector" are usually located deep underground in order to filter background noise. These detectors can be used to observe supernova neutrinos, and serve as a trigger to direct other observers to capture the supernova evolution early for multi-messenger astronomy. The neutrino detectors need to point the other observers to the supernova bursts. Detector data is initially buffered underground. Providing the supernova location only after transferring all that data to the surface for processing would delay the message too long for others to capture the evolution. Therefore, at least some processing needs to be done in the cavern, either to fully point to a supernova, or to select a small subset of data to send to the surface for processing. In order to not burden the processor, we want to exploit "in-storage computation." In particular, we seek to use an accelerator that accesses the data directly from storage for the processing. For our demonstrator, we are using a Xilinx Alveo accelerator, accessing SSD storage using PCIe peer-to-peer transfers. One of the primary tasks that the computational storage system is performing is running a machine learning algorithm to identify regions of interest within LArTPC waveforms. This model was adapted and retrained on simulated DUNE LArTPC data, and further optimized by hand, along using an automated hyperparamter tuning platform determined.ai using the ASHA algorithm. The model is small, taking an input of 200 points of 1D waveform data, and consisting of three 1D convolutional layers with one dense output layer. In total, the model has approximately 21,000 parameters. After training and optimization, it is then converted into FPGA firmware via the hls4ml software package. The hls4ml software package was designed to make deploying optimized NNs on FPGAs and ASICs accessible for domain applications. hls4ml takes ML input from standard tools like Keras or PyTorch and usually produces High Level Synthesis (HLS) code that can be synthesized by for example, Vivado HLS. It was originally written to help the design of the first level triggering system for the CMS detector at CERN. The hls4ml generated HLS is combined with a data parser and run as a kernel in the Vitis accelerator methodology. The hls4ml package provides tunable parameters for various tradeoffs between size and latency. We can also instantiate multiple kernels. We are exploring other processing to also do in the accelerator to best achieve our goal of providing pointing information quickly.

Contributed Talks / 12

Low-latency Calorimetry Clustering at the LHC with SPVCNN

Authors: Alexander Joseph Schuy[1]; Jeffrey Krupa[2]; Philip Coleman Harris[3]; Scott Hauck[None]; Shih-Chieh Hsu[4]; Song Han[2]; William Patrick Mccormack[3]; Zhijian Liu[2]

[1] *University of Washington (US)*

[2] *Massachusetts Institute of Technology*

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

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

Corresponding Author: alexander.joseph.schuy@cern.ch

The search for dark matter and other new physics at the Large Hadron Collider (LHC) involves enormous data collection. Due to this, a high-level trigger system (HLT) must decide which data to keep for long-term storage while maintaining high throughput and on the order of millisecond latency. A central part of the HLT is 3D clustering of low-level detector measurements in the calorimeter. In this work, we show low-latency, high-throughput 3D calorimetry clustering using Sparse Point-Voxel Convolutional Neural Networks (SPVCNN) that can be deployed at-scale to heterogeneous computing systems while maintaining or exceeding the performance of conventional algorithms.

Contributed Talks / 13

A Deep Learning Approach to Particle Identification for the AMS Electromagnetic Calorimeter

Author: Raheem Hashmani[1]

Co-authors: Bilge Demirkoz [1]; Emre Akbaş [1]; Zhili Weng [2]

[1] *Middle East Technical University (TR)*

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

Corresponding Author: raheem.hashmani@cern.ch

The Alpha Magnetic Spectrometer (AMS-02) is a high-precision particle detector onboard the International Space Station containing six different subdetectors. One of these, the Electromagnetic Calorimeter (ECAL), is used to measure the energy of cosmic-ray electrons and positrons and to differentiate these particles from cosmic-ray protons up to TeV energy.

We present a new deep learning approach for particle identification by taking as an input the energy deposition within all the calorimeter cells. By treating the cells as pixels in an image-like format, with effectively 2,592 features, we use various vision-based deep learning models as classifiers and compare their performances. Some of the models selected for training and evaluating range from simple convolutional neural networks (CNN) to more state-of-the-art residual neural networks (ResNet) and convolutional vision transformers (CvT).

The particle identification performance is evaluated using Monte Carlo electron and proton events from 100 GeV to 4 TeV. At 90% electron accuracy, for the entire energy range, the proton rejection power of our CvT model outperforms the CNN and ResNet models by more than a factor of 12 and 10, respectively. This shows promise for future use in the AMS-02 experiment and provides empirical evidence of newer architectures, such as transformers, outperforming CNNs for use in calorimeters.

Contributed Talks / 14

Intelligent experiments through real-time AI: Fast Data Processing and Autonomous Detector Control for sPHENIX and future EIC detectors

Author: Micol Rigatti[1]

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

Corresponding Author: mrigatti@fnal.gov

The upcoming sPHENIX experiment, scheduled to start data taking at the Brookhaven National Laboratory (BNL) Relativistic Heavy Ion Collider in 2023, and the future Electron-Ion Collider (EIC) experiments will employ sophisticated state-of-the-art, high-rate detectors to study high energy heavy ion and electron-ion collisions, respectively. The resulting large volumes of raw data far exceed available DAQ and data storage capacity.
The application of modern computational techniques, in particular recent developments in artificial intelligence (AI) and machine learning (ML), has the potential to address these new challenges and revolutionize our approach to collecting, reconstructing and understanding data, and thereby maximizing the discovery potential. Our proposal seeks to transform future flagship detectors into "intelligent experiments"and "smart data acquisition and control"through the integration of next generation AI/ML hardware, electronics and algorithms into these detector systems.
We propose to develop a selective streaming readout system, built on state-of-the-art readout electronics and computing, to achieve continuous readout and inspection of all essential detector signals. AI algorithms will be used to reduce the raw data volume on the fly by identifying events with heavy flavor jets, through real-time detector control, reconstruction and event selection.
The tracking detectors will all use the BNL-designed Front-End Link eXchange (FELIX) FPGA card as a data aggregator. With the development of advanced deep neural networks, a parallel strategy is needed to ensure that these networks can be designed to operate at low latency and high throughput on the FELIX FPGA cards. Developing and implementing these techniques in the context of scientific low-power, low-latency, or resource constrained use cases is a major goal of this research program.

In this proposal we will use hls4ml to integrate AI models into the streaming system of sPHENIX. hls4ml will take a neural network model generated in a standard ML format (keras/tensorflow/pytorch/onnx) and translate this to an FPGA/ASIC synthesizable high level synthesis code. The generated code in high level synthesis language (HLS) is C++ based.
The challenge in creating an optimal digital design is to balance available resources with achieving the power, latency, throughput goals of the target application. In this talk, we will present details of the implementation and the latest progress of this project.

Contributed Talks / 15

Quantized ONNX (QONNX)

Authors: Alessandro Pappalardo[1]; Benjamin Hawks[2]; Hendrik Borras[3]; Javier Mauricio Duarte[4]; Jovan Mitrevski[5]; Jules Muhizi[6]; Matthew Trahms[7]; Michaela Blott[1]; Nhan Tran[5]; Scott Hauck[None]; Shih-Chieh Hsu[8]; Sioni Paris Summers[9]; Vladimir Loncar[None]; Yaman Umuroglu[None]

[1] *AMD Adaptive and Embedded Computing Group (AECG) Labs*

[2] *Fermi National Accelerator Lab*

[3] *Uni Heidelberg*

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

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

[6] *Fermilab/Harvard University*

[7] *UW ACME Lab*

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

[9] *CERN*

Corresponding Author: jovan.mitrevski@cern.ch

One of the products of the cooperation between the hls4ml and FINN groups is Quantized ONNX (QONNX), a simple but flexible method to represent uniform quantization in ONNX. Its goal is to provide a high-level representation that can be targeted by training frameworks while minimizing reliance on implementation-specific details. It should also be lightweight, only adding a small number of operators. QONNX accomplishes this by being in the fused quantize-dequantize (QDQ) style. The main operator is the Quant operator, which takes a bitwidth, scale, and zero-offset to quantize an input tensor and then immediately dequantizes it, undoing the scale and zero offset. The resulting values are (quantized) floating point numbers, which can be used by standard ONNX operators. There is also a BipolarQuant operator, which is like the regular Quant operator but specialized for binary quantization. Finally there is a Trunc operator to truncate the least significant bits. Currently Brevitas, a PyTorch research library for quantization-aware training (QAT), and QKeras, a Keras library for QAT, can produce QONNX. HAWQ support is being added, and is the focus of a separate abstract.

The FINN and hls4ml groups also worked on a common set of utilities to ease the ingestion of QONNX by the FINN and hls4ml software. These utilities simplify the ONNX graphs by doing such things as folding constants, inferring dimensions, making sure nodes are named—commonly referred to as cleaning. FINN and hls4ml also prefer convolution data to be in a channels-last format, so we have a common pass to convert the ONNX graphs to a channels-last format using custom operators. We also have some common optimizers to, for example, change Gemm operators to lower level MatMul and Add operators so that FINN and hls4ml do not need to handle Gemm explicitly.

We will also present how hls4ml ingests QONNX. Given the high-level nature of QONNX, a direct implementation, dequantizing right after quantizing, does not map well to hardware. Instead, hls4ml makes use of optimizers to convert the abstract graph to something that can be more easily implemented on an FPGA or ASIC. In particular, the scale and zero-point in a quantization and in dequantization, if not one and zero respectively, are logically stripped from the quantization operation, resulting in three operations: scale and offset, quantization, and unscale and de-offset. The

unscaling can then often be propagated down across linear operations like matrix multiplies or convolutions, to produce quantized dense or convolution layers. As an optimization, for power-of-two scales and zero offsets, we can offload the scale propagation to the HLS compiler by using fixed precision numbers, and for quantized constant weights, we can merge the scale/offset and quantization into the weights, only leaving an unscale and de-offset node if needed.

We also introduce a QONNX model zoo to share quantized neural networks in the QONNX format.

Contributed Talks / 16

Accelerating JEDI-net for jet tagging on FPGAs

Authors: Zhiqiang Que[1]; Alexander D Tapper[1]; Wayne Luk[1]

[1] *Imperial College London*

Corresponding Author: z.que@imperial.ac.uk

This work proposes a novel reconfigurable architecture for reducing the latency of JEDI-net, a Graph Neural Network (GNN) based algorithm for jet tagging in particle physics, which achieves state-of-the-art accuracy. Accelerating JEDI-net is challenging since low latency is requited to potentially deploy the network on the online event selection systems at the CERN Large Hadron Collider. This presentation proposes a custom code transformation with strength reduction for matrix multiplication operations which avoids the costly multiplication of the adjacency matrix with the input feature matrix. It exploits sparsity patterns as well as binary adjacency matrices, and avoids irregular memory access, leading to a reduction in latency and improvement of hardware efficiency. We also introduce an outer-product based matrix multiplication approach which is enhanced by the strength reduction for low-latency design. Furthermore, a customizable template for this architecture has been designed and open-sourced, which enables the generation of low-latency FPGA designs with efficient resource utilization using high-level synthesis tools. Evaluation results show that our FPGA implementation is up to 9.5 times faster and consumes up to 6.5 times less power than a GPU implementation. Moreover, the throughput and latency of our FPGA design is sufficiently high to enable deployment of JEDI-net in a sub-microsecond, real-time collider trigger system, enabling it to benefit from improved accuracy.

Contributed Talks / 17

Harnessing ultrafast ML for new algorithms at the CMS L1 trigger

Author: Daniel Diaz[1]

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

Corresponding Author: daniel.cipriano.diaz@cern.ch

In the high luminosity LHC (HL-LHC) era, the CMS detector will be subject to an unprecedented level of simultaneous proton-proton interactions (pile-up) that complicate the reconstruction process. Mitigation of the effects of pile-up is of prime importance. In preparation for this, the detector will be upgraded, providing more granularity and more information than we have had before. In addition to the pile-up mitigation, we can use these upgrades to enable and improve the physics strategy at the Level-1 (L1) trigger. With the inclusion of FPGA boards with greater resources in the L1 upgrade, and new codesign tools like hls4ml for easily converting neural networks into FPGA firmware, we now have the ability to deploy ultrafast machine learning algorithms at L1. This talk will describe plans to use machine learning techniques at L1 to perform anomaly detection, long-lived particle detection, and better estimate the missing transverse momentum.

Contributed Talks / 18

Resource Efficient and Low Latency GNN-based Particle Tracking on FPGA

Authors: Bo-Cheng Lai[None]; Shi-Yu Huang[None]

Co-authors: Abdelrahman Elabd ; Javier Duarte [1]; Jin-Xuan Hu ; Mark Neubauer [2]; Markus Atkinson [2]; Scott Hauck ; Shih-Chieh Hsu [3]; Vesal Razavimaleki [4]

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

[2] *Univ. Illinois at Urbana Champaign (US)*

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

[4] *Univ. Illinois at Urbana-Champaign (US)*

Corresponding Authors: ariel56899@gmail.com, bclai@nycu.edu.tw

Charged particle tracking is important in high-energy particle physics. For CERN Large Hadron Collider (LHC), tracking algorithms are used to identify the trajectories of charged particles created in the collisions. The existing tracking algorithms are typically based on the combinatorial Kalman filter where the complexity increases quadratically with the number of hits. The poor scalability issue will be exacerbated when the beam intensities are expected to increase dramatically. Therefore, new tracking algorithms based on Graph Neural Networks (GNNs) are introduced to enhance the scalability of particle tracking tasks. These GNN algorithms are implemented on Field Programmable Gate Arrays (FPGAs) to meet the strict latency requirement of fast particle tracking. However, the previous design on Xilinx Virtex UltraScale+ VU9P FPGA can only accommodate a small GNN (28 nodes / 56 edges) due to the significant resource requirement of complex graph processing patterns. A collision event (660 nodes / 1320 edges) needs to be partitioned into smaller sub-graphs to fit the GNN processing to VU9P FPGA. Dividing a collision event into smaller sub-graphs could cause a higher possibility of missing important trajectories between sub-graphs.
In this work, we introduce a resource efficient and low latency architecture to accelerate large GNN processing on FPGA. This design leverages the GNN processing patterns and trajectory data properties to significantly improve the parallelism and computation throughput. We propose a highly parallel architecture with configurable parameters for users to adjust latency, resource utilization, and parallelism. A customized data allocation is used to address the irregular processing patterns and attain high processing parallelism. We further exploit the properties of trajectories between inner and outer detector layers, and reduce the unnecessary dependencies and edges in the graph.
The design is synthesized using hls4ml and implemented on Xilinx Virtex UltraScale+ VU9P FPGA. The proposed design can support a graph of size 660 nodes and 1560 edges with Initialization Interval of 200 ns.

Contributed Talks / 19

Implementation of a pattern recognition neural network for live reconstruction using AI processors

Authors: Jochen Kaminski[1]; Klaus Desch[2]; Michael Lupberger[2]; Patrick Schwaebig[None]

[1] *University of Bonn (DE)*

[2] *University of Bonn*

Corresponding Author: schwaebig@physik.uni-bonn.de

For years, data rates generated by modern particle detectors and the corresponding readout electronics exceeded by far the limits of bandwidth and data storage space available in many experiments. Using fast triggers to discard uninteresting and irrelevant events is a solution used to this day. FPGAs, ASICs or even directly the readout chip are programmed or designed to apply a fixed set of rules based on low level parameters for an event pre-selection. However with detector technology

progressing quickly and newly devised experiments demanding ever-increasing particle collision rates new ways for triggering have to be considered.

One of these is live track reconstruction for triggering meaning that high level information like particle momentum is exracted from the raw data and directly used for triggering. Up until now this approach was rarely possible due to a conflict between processing time and the required trigger latency. With the emergence of novel fast and highly parallelized processors, targeted mainly at AI inference, attempts to sufficiently accelerate tracking algorithms become viable. The Xilinx Versal AI Series Adaptive Compute Acceleration Platform (ACAP) is one such technology and combines traditional FPGA and CPU resources with dedicated AI cores and a network on chip for fast memory access. Despite being available for some years this technology is still largely unexplored for particle physics cases.

In this talk a Versal ACAP implementation of a neural network for pattern recognition for a dark photon experiment at the ELSA accelerator in Bonn, Germany will be shown as an example application and the expected performance will be discussed.

Contributed Talks / 20

Real-time image processing for high-resolution imaging detectors

Authors: Georgia Karagiorgi[None]; Giuseppe Di Guglielmo[1]; Luca Carloni[2]; Nicholas Alexander Kasseinov[3]

[1] *Fermilab*

[2] *Columbia University*

[3] *Columbia University (US)*

Corresponding Author: georgia@nevis.columbia.edu

Modern-day particle and astro-particle physics experiments call for detectors with increasingly higher imaging resolution to be deployed in often inaccessible, remote locations, e.g., deep underground or in-flight on balloons or satellites. The inherent limitations in available on-detector power and computational resources, combined with the need to operate these detectors continually, thus producing an exorbitant amount of data, calls for fast, efficient, and accurate data processing to filter out usually rare features of interest from the data, and save it for further, offline processing and physics analysis. Real-time data processing using machine learning algorithms such as convolutional neural networks provides a promising solution to this challenge. This talk reviews ongoing R&D to demonstrate such capability for the case of the future Deep Underground Neutrino Experiment (DUNE).

Contributed Talks / 21

CryoAI –Prototyping cryogenic chips for machine learning at 22nm

Authors: Chinar Syal[1]; Davide Giri[2]; Farah Fahim[3]; Giuseppe Di Guglielmo[3]; Joseph Zuckerman[2]; Luca Carloni[2]; Maico Cassel Dos Santos[2]; Manuel Valentin[4]; Nhan Tran[1]; Seda Memik[4]

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

[2] *Columbia University*

[3] *Fermilab*

[4] *Northwestern University*

Corresponding Author: manuelvalentin2028@u.northwestern.edu

We present our design experience of a prototype System-on-Chip (SoC) for machine learning applications that run in a cryogenic environment to evaluate the performance of the digital backend flow. We combined two established open-source projects (ESP and HLS4ML) into a new system-level design flow to build and program the SoC. In the modular tile-based architecture, we integrated a low-power 32-bit RISC-V microcontroller (Ibex), 200KB SRAM-based scratchpad, and an 18K-parameter neural-network accelerator. The network is an autoencoder working on audio recordings and trained on industrial use cases for the early detection of failures in machines like slide rails, fans, or pumps. For the hls4ml translation, we optimized the reference architecture using quantization and model compression techniques with minimal AUC performance reduction. This project is also an early evaluation of Siemens Catapult as an HLS backend for hls4ml. Finally, we fabricated the SoC in a 22nm technology and are currently testing it.

Contributed Talks / 22

Autonomous real-time science-driven follow-up of survey transients

Authors: Christoffer Fremling[1]; Matthew Graham[2]; Michael Coughlin[3]; Niharika Sravan[None]

[1] *Caltech*

[2] *California Institute of Technology*

[3] *University of Minnesota*

Corresponding Author: niharika.sravan@gmail.com

Astronomical surveys continue to provide unprecedented insights into the time-variable Universe and will remain the source of groundbreaking discoveries for years to come. However, their data throughput has overwhelmed the ability to manually synthesize alerts for devising and coordinating necessary follow-up with limited resources. The advent of Rubin Observatory, with alert volumes an order of magnitude higher at otherwise sparse cadence, presents an urgent need to overhaul existing human-centered protocols in favor of machine-directed infrastructure for conducting science inference and optimally planning expensive follow-up observations. We present the first implementation of autonomous real-time science-driven follow-up using value iteration to perform sequential experiment design. We demonstrate it for strategizing photometric augmentation of Zwicky Transient Facility Type Ia supernova light-curves given the goal of minimizing SALT2 parameter uncertainties. We find a median improvement of 2-6% for SALT2 parameters and 3-11% for photometric redshift with 2-7 additional data points in g, r and/or i compared to random augmentation. The augmentations are automatically strategized to complete gaps and for resolving phases with high constraining power (e.g. around peaks). We suggest that such a technique can deliver higher impact during the era of Rubin Observatory for precision cosmology at high redshift and can serve as the foundation for the development of general-purpose resource allocation systems.

Contributed Talks / 23

Fast recurrent neural networks on FPGAs with hls4ml

Authors: Aaron Wang[None]; Caterina Vernieri[1]; Chaitanya Paikara[2]; Dylan Sheldon Rankin[3]; Elham E Khoda[4]; Michael Aaron Kagan[1]; Philip Coleman Harris[3]; Rafael Teixeira De Lima[1]; Richa Rao[2]; Scott Hauck[None]; Shih-Chieh Hsu[5]; Sioni Paris Summers[6]; Vladimir Loncar[None]

[1] *SLAC National Accelerator Laboratory (US)*

[2] *University of Washington*

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

[4] *University of Washington (US)*

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

[6] *CERN*

Corresponding Author: elham.e.khoda@cern.ch

Recurrent neural networks have been shown to be effective architectures for many tasks in high energy physics, and thus have been widely adopted. Their use in low-latency environments has, however, been limited as a result of the difficulties of implementing recurrent architectures on field-programmable gate arrays (FPGAs). In this paper we present an implementation of two types of recurrent neural network layers- long short-term memory and gated recurrent unit- within the hls4ml 1 framework. We demonstrate that our implementation is capable of producing effective designs for both small and large models, and can be customized to meet specific design requirements for inference latencies and FPGA resources. We show the performance and synthesized designs for multiple neural networks, many of which are trained specifically for jet identification tasks at the CERN Large Hadron Collider.

1 J. Duarte et al., "Fast inference of deep neural networks in FPGAs for particle physics", JINST 13 (2018) P07027, arXiv:1804.06913

Contributed Talks / 24

Low-latency Noise Subtraction of Gravitational Wave Data by DeepClean

Author: Chia-Jui Chou[1]

[1] *National Yang Ming Chiao Tung University*

Corresponding Author: agoodmanjerry@gmail.com

DeepClean is the technique using deep learning to clean the environmental noises in the gravitational wave strain data. The signals from the witness censors recording the environmental noises are used to produce the noises coupled to the strain data. After training the DeepClean model, the online cleaning in low latency is conducted by the Inference-as-a-Service model. The plans of implementing the online DeepClean in LIGO and KAGRA for the coming O4 observation will be introduced.

Contributed Talks / 25

A novel ML-based method of primary vertex reconstruction in high pile-up condition

Authors: Alexander Joseph Schuy[1]; Haoran Zhao[1]; Ke Li[1]; Philip Coleman Harris[2]; Scott Hauck[None]; Shih-Chieh Hsu[3]; Song Han[4]; Zhijian Liu[4]

[1] *University of Washington (US)*

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

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

[4] *Massachusetts Institute of Technology*

Corresponding Author: haoran.zhao@cern.ch

The High-Luminosity LHC (HL-LHC) is expected to reach a peak instantaneous luminosity of $7.5 \times 10^{34}\mathrm{cm}^{-2}\mathrm{s}^{-1}$ at a center-of-mass energy of $\sqrt{s} = 14\mathrm{TeV}$. This leads to an extremely high density environment with up to 200 interactions per proton-proton bunch crossing. Under these conditions,

event reconstruction represents a major challenge for experiments due to the high pileup vertices present.
To tackle the dense environment, we adapted a novel ML-based method named Sparse Point-Voxel Convolution Neural Network(SPVCNN), the current state-of-the-art techniques in computer vision, which leverages point-based method and space voxelization to categorize tracks into primary vertices. Then SPVCNN is trained and tested on the samples generated by ACTS.
In this talk, the performance of SPVCNN vertexing will be presented, as well as the comparison with the conventional Adaptive Multi-Vertex Finding(AMVF) algorithm used in ATLAS.

Contributed Talks / 26

A Normalized Autoencoder for LHC triggers

Authors: Barry Dillon[1]; Luigi Favaro[None]; Michael Kramer[2]; Peter Rangi Sorrenson[3]; Tilman Plehn[None]

[1] *University of Heidelberg*

[2] *Rheinisch Westfaelische Tech. Hoch. (DE)*

[3] *Universität Heidelberg*

Corresponding Author: luigi.favaro@stud.uni-heidelberg.de

The main goal for the upcoming LHC runs is still to discover BSM physics. It will require analyses able to probe regions not linked to specific models but generally identified as beyond the Standard Model. Autoencoders are the typical choice for fast anomaly detection models. However, they have shown to misidentify anomalies of low complexity signals over background events. I will present an energy-based Autoencoder called Normalized AE, a density-based high-performance anomaly search algorithm. I will discuss how NAE is able to symmetrically tag QCD and top jets and I will discuss the possibility of implementing NAEs on FPGAs for the LHC L1 trigger.

Contributed Talks / 27

Designing intelligent DAQ systems for radiation instrumentation with hls4ml

Author: Audrey Corbeil Therrien[1]

Co-authors: Berthié Gouin-Ferland [1]; Charles-Étienne Granger [1]; Hamza Ezzaoui Rahali [1]; Mohammad Mehdi Rahimifar [1]

[1] *Université de Sherbrooke*

Corresponding Author: audrey.corbeil.therrien@usherbrooke.ca

As detector technologies improve, the increase in resolution, number of channels and overall size create immense bandwidth challenges for the data acquisition system, long data center compute times and growing data storage costs. Much of the raw data does not contain useful information and can be significantly reduced with veto and compression systems as well as online analysis.
The improvements in artificial intelligence, particularly the many flavours of machine learning, adds a powerful tool to data acquisition strategies by providing embedded analysis, reducing the data at the source. However large and deep neural network are still compute intensive and one of the most important challenges ML designers face is minimizing the model size without losing the precision and accuracy required for a scientific application. The combination of signal processing algorithms and compression algorithms with machine learning can improve the latency and accuracy of edge systems by reducing the width and depth of the model.
Using this strategy, we develop hardware compatible signal processing and machine learning systems for various radiation detectors. We will present two current use cases:

1) The CookieBox, an attosecond angular streaking detector used for X-ray pulse shape recovery generating ˜800 GB/s. This system requires microsecond latency to apply a veto on downstream detectors. We designed both a fully connected neural network and a convolutional neural network (CNN) each combined with a non-uniform data quantizer. These networks achieve 86 % accuracy in 8 µs for the fully connected network and 88 % accuracy in 23 µs for the CNN (including data transfer times) on a ZYNQ XC7Z02.

2) The billion pixel X-ray camera for use in synchrotrons, XFEL facilities and pulsed power facilities generating up to 15 TB/s. Data is compressed using a trained neural network to accelerate the ISTA algorithm (Learned ISTA) and combined with a DEFLATE compression to achieve 83:1 compression. The network compressed each 6x6 pixel patch in less than 2 µs when implemented on a ZYNQ XC7Z02.

For both systems complementary, additional hardware modules were designed and integrated with hls4ml to implement the complete data analysis on FPGA. Several other ongoing projects are currently benefitting from the methods developed with these systems, including medical imaging and dark matter search.

Contributed Talks / 28

Large CNN for HLS4ML and Deepcalo

Authors: Alexander Joseph Schuy[1]; Bo-Cheng Lai[None]; ChiJui Chen[None]; Dylan Sheldon Rankin[2]; Lin-Chi Yang[None]; Philip Coleman Harris[2]; Scott Hauck[None]; Shih-Chieh Hsu[3]; Yan-Lun Huang[None]; Ziang Yin[1]

[1] *University of Washington (US)*

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

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

Corresponding Authors: hisky1256@gmail.com, gary0710172.ee07@nctu.edu.tw, kugelblitz.ee05@gmail.com

Convolutional neural networks (CNN) have been widely applied in a tremendous of applications that involve image processing, including particle physics. Deepcalo is a package designed for developing CNNs using ATLAS data at CERN, targeting tasks like energy regression of electrons and photons. Although it has been shown that CNNs used in Deepcalo can handle the task smoothly, the extensive computation resources and high-power consumption lead it hard to perform real-time inference during the experiment. As a result, it is limited in software simulation usage.

To accelerate the inference time and lower the power consumption, we implement those CNNs on FPGAs (Field Programmable Gate Arrays) with HLS4ML. HLS4ML is an automated tool for deploying machine-learning models on FPGAs, targeting ultra-low latency using fully-on-chip architecture. Based on HLS C++ codes by Dr. Dylan Rankin, we extend the HLS4ML library for supporting an automatic large CNNs conversion. In this work, we introduce a deeply-optimized workflow for implementing large CNNs on FPGAs. Implemented on an AlveoU50 FPGA running at 200 MHz, the accelerator infers with 0.039 of IQR75 loss in 0.6 ms.

Contributed Talks / 29

Demonstration of Machine Learning-assisted real-time noise regression in LIGO

Author: Muhammed Saleem Cholayil[1]

Co-authors: Alec Gunny ; Michael Coughlin [1]; Dylan Sheldon Rankin [2]; Erik Katsavounidis [3]; Philip Coleman Harris [2]

[1] *University of Minnesota*

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

[3] *MIT*

Corresponding Author: mcholayi@umn.edu

Gravitational wave (GW) detectors such as advanced LIGO, advanced Virgo, and KAGRA are high-precision instruments that record the strain signals from transient astrophysical sources such as merging binary black holes. The sensitivities of these detectors are often limited by instrumental and environmental noise that couple non-linearly to the GW strain. Noise regression algorithms running as close as possible to real-time are therefore important for maximizing the science outcomes of these interferometers. DeepClean is a deep learning convolutional neural network algorithm for the subtraction of non-linear and non-stationary noise from GW strain data. DeepClean computes the noise contamination with the help of auxiliary witness sensors that record those instrumental and environmental random processes. We deploy DeepClean as a low-latency noise-regression algorithm for LIGO data and demonstrate the performance in terms of latency, signal-to-noise ratio, and astrophysical parameter estimation.

Contributed Talks / 30

End-to-end acceleration of machine learning in gravitational wave physics

Author: Alec Gunny[1]

Co-authors: Dylan Sheldon Rankin [2]; Eric Anton Moreno [3]; Erik Katsavounidis [4]; Ethan Marx [1]; Michael Coughlin [5]; Philip Coleman Harris [2]; Ryan Raikman [1]; Saleem Muhammed [5]; William Benoit [5]

[1] *Massachusetts Institute of Technology*

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

[3] *Massachusetts Institute of Technology (US)*

[4] *MIT*

[5] *University of Minnesota*

Corresponding Author: alecg@mit.edu

While applications of deep learning (DL) to gravitational wave (GW) physics are becoming increasingly common, very few have reached the maturity to be deployed in truly automated services. This is symptomatic of a larger gap between the existing tool sets for both GW physics and DL, neither of which has historically been developed or optimized for use with the other. This has led to suboptimal training code which is forced to tradeoff between speed and data robustness, divergent methods for analyzing the efficacy of trained models, and difficulties in deploying and distributing models within the traditional GW computing environment. Taken together, these challenges combine to create experimental pipelines which take longer to iterate upon and which produce results that are both less conclusive and less reproducible. We present here a set of libraries, ml4gw and hermes, aimed at bridging some of these gaps and allowing for the development of DL-powered GW physics applications which are faster, more intuitive, and better able to leverage the powerful modeling techniques available in the GW literature.

31

Autonomous experiments in scanning probe microscopy: opportunities for rapid inference and decision making

Author: Rama Vasudevan[1]

Co-authors: Sergei Kalinin [2]; Yongtao Liu [1]; Maxim Ziatdinov [1]; Sai Mani Valleti ; Stephen Jesse [1]; Jan-Chi Yang [3]

[1] *Oak Ridge National Laboratory*
[2] *University of Tennessee*
[3] *National Cheng Kung University*

Corresponding Author: vasudevanrk@ornl.gov

The rise of robotics, automation and the creation of various levels of abstraction have by now enabled automated experiments on a range of scientific instruments ranging from chemical robots for molecular synthesis, to electron and scanning probe microscopes that can be programmed to enable automated and autonomous experiments with a view towards physics discovery.

In this talk, I will briefly outline automated and autonomous experiments as it pertains to scanning probe microscopy, here at the Center for Nanophase Materials Sciences. It will be shown that microscopy in general is an ideal playground for the development, testing and deployment of machine learning methods, form both a hardware and algorithmic viewpoint. Typical automated setups and needs for Fast ML will be discussed in the context of problems such as using reinforcement learning inline on the microscope for tuning domain wall functionality in ferroelectrics. We posit that the correct deployment of algorithms and simulations at the edge, on HPC and at the cluster level, with workflow tools and connectivity, will be critical in realizing truly autonomous microscopy platforms for physics discovery. This work was supported by the Center for Nanophase Materials Sciences (CNMS), which is a US Department of Energy, Office of Science User Facility at Oak Ridge National Laboratory.

Contributed Talks / 32

Exa.TrkX inference as-a-service

Authors: Alina Lazar[1]; Shih-Chieh Hsu[2]; Xiangyang Ju[3]; Yongbin Feng[4]

[1] *Youngstown State University*
[2] *University of Washington Seattle (US)*
[3] *Lawrence Berkeley National Lab. (US)*
[4] *Fermi National Accelerator Lab. (US)*

Corresponding Author: yongbin.feng@cern.ch

Particle tracking plays a crucial role in many particle physics experiments, e.g, the Large Hadron Collider. Yet, it is also one of the most time-consuming components in the whole particle reconstruction chain. The Exa.TrkX group has developed in recent years a promising and well-performed machine-learning-based pipeline that carries out the track finding, which is the most computationally expensive part of particle tracking. An important research direction is to accelerate the pipeline, via software-based approaches such as model pruning, tensor operation fusion, reduced precision, quantization, and hardware-based approaches such as usages of different coprocessors, such as GPUs, TPUs, and FPGAs.
In this talk, we will introduce our implementation of Exa.TrkX inference as-a-service through NVIDIA Triton Inference servers. Clients read data and send track-finding inference requests to (remote) servers; servers run the inference pipeline on different types of coprocessors and return outputs to clients. The pipeline running on the server side includes three discrete deep learning models and two CUDA-based domain algorithms. Compared with normal local inferences, this approach allows us more freedom to easily utilize different types of coprocessors more efficiently, while maintaining similar throughputs and latency. We will discuss in detail different server configurations explored in order to achieve this.

Contributed Talks / 33

Application of deep learning to instability tracking using high-speed video cameras in magnetic confinement fusion

Authors: Yumou Wei[1]; Jeffrey Levesque[1]; Christopher Hansen[2]

[1] *Columbia University*

[2] *University of Washington*

Corresponding Author: yw2714@columbia.edu

High-speed cameras have broadly been used to monitor plasma-wall interactions and to study spatial features of the plasma edge inside magnetic confinement fusion experiments. Depending on plasma parameters and photon energy sensitivity, a 2D imaging system can also be used to track the phase and amplitude of long-wavelength instability modes 1. Such cameras can be used in devices where there is reduced diagnostic access around the experiment. Using deep-learning-based algorithms, streaming cameras could be used in real-time mode control applications similar to using standard magnetic sensors. Such algorithms will require an inference latency on the order of microseconds, so careful deployment strategies will be necessary. Developments of this control routine on the High Beta Tokamak –Extended Pulse (HBT-EP) device will be presented.

1 Angelini, et al., Plasma Phys Contr Fusion, 57, 045008 (2015).

Supported by U.S. DOE Grants DE-FG02-86ER53222

Contributed Talks / 34

FKeras: A Fault Tolerance Library for DNNs

Author: Olivia Weng[None]

Co-authors: Andres Meza [1]; Benjamin Hawks [2]; Quinlan Bock [3]; Javier Mauricio Duarte [4]; Nhan Tran [5]; Ryan Kastner

[1] *UC San Diego*

[2] *Fermi National Accelerator Lab*

[3] *Fermilab National Accelerator Laboratory*

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

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

Corresponding Author: oweng@ucsd.edu

Many scientific applications require NNs to operate correctly in safety-critical or high radiation environments, including automated driving, space, and high energy physics. For example, physicists at the Large Hadron Collider (LHC) seek to deploy an autoencoder in a high radiation environment to filter their experimental data, which is collected at a high data rate (˜40TB/s). This is challenging because the autoencoder must operate efficiently, within 200 ns, in a resource-constrained setting to process all the data as well as correctly amid high radiation. As such, the autoencoder's hardware must be both efficient and robust.

However, efficiency and robustness are often in conflict with each other. Robust hardware methods like triple modular redundancy protect against faults by increasing resources by 200%, in turn reducing efficiency 1. To address these opposing demands, we must understand the fault tolerance inherent to NNs. NNs have many redundant parameters, suggesting we do not need to introduce a blanket redundancy in the hardware —the common practice—when it is already present in the software. To identify where this redundancy exists in a NN, we present FKeras, an open-source tool that measures the fault tolerance of NNs at the bit level. Once we identify which parts of the NN are insensitive to radiation faults, we need not protect them, reducing the resources spent on robust hardware.

FKeras takes a fine-grained, bottom-up approach to evaluate the fault tolerance of NNs at the bit-level. The user can evaluate both floating point and quantized NNs, for which previous work had little support. Since FKeras builds on top of QKeras, a quantized NN library, users can easily adjust

quantization settings as well as fault injection settings (like bit error rate, bit error location, transient versus permanent fault, etc.) during training and/or inference. Prior work [1-3] introduced tools to evaluate NN robustness; however, they are too coarse-grained or are closed source, precluding researchers from fully understanding the robustness of NNs. They also have limited quantization support. FKeras is open-sourced, allowing researchers to easily evaluate quantized NNs at the bit-level. Having a bit-level understanding is paramount when every bit counts, especially in resource-constrained settings at the extreme edge. FKeras is a first step towards providing an open-source way of identifying which bits must be protected and which do not.

We would like to extend FKeras to statically identify which bits are insensitive to faults, without simulation to save time. At the workshop, we look forward to discussing and better understanding the fault tolerance needs of science. We will keep these needs in mind as we continue to build FKeras, with the goal of better supporting the scientific community.

1 Bertoa et al. "Fault Tolerant Neural Network Accelerators with Selective TMR." IEEE D&T'22.
[2] Chen et al. "Tensorfi: A flexible fault injection framework for tensorflow applications." ISSRE' 20.
[3] Mahmoud et al. "Pytorchfi: A runtime perturbation tool for dnns." DSN-W'20.

Contributed Talks / 35

Increasing the LHC Computational Power by integrating GPUs as a service

Authors: Elham E Khoda[1]; Javier Mauricio Duarte[2]; Kevin Pedro[3]; Miaoyuan Liu[4]; Nhan Tran[3]; Nirmal Thomas[None]; Philip Coleman Harris[5]; Raghav Kansal[2]; Shih-Chieh Hsu[6]; Simon Rothman[5]; Stefan Piperov[4]; William Patrick Mccormack[5]; Yongbin Feng[3]

[1] *University of Washington (US)*

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

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

[4] *Purdue University (US)*

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

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

Corresponding Authors: william.patrick.mc.cormack.iii@cern.ch, yongbin.feng@cern.ch

Over the past several years, machine learning algorithms at the Large Hadron Collider have become increasingly more prevalent. Because of their highly parallelized design, Machine Learning-based algorithms can be sped up dramatically when using coprocessors, such as GPUs. With increasing computational demands coming from future LHC upgrades, there is a need to enhance the overall computational power of the next generation of LHC reconstruction. In this talk, we demonstrate a strategy to port deep learning algorithms to GPUs efficiently. By exploiting the as-a-service paradigm to port algorithms to GPU, we are able to optimally use GPU resources, allowing for a path towards efficient GPU adoption at the LHC as more algorithms become parallelizable. In this talk, we present this path and demonstrate an end-to-end workflow with current reconstruction using the Compact Muon Solenoid.

Contributed Talks / 36

Next Generation Coprocessors as a service

Authors: Dylan Sheldon Rankin[1]; Elham E Khoda[2]; Javier Mauricio Duarte[3]; Miaoyuan Liu[4]; Nhan Tran[5]; Nirmal Thomas[None]; Philip Coleman Harris[1]; Shih-Chieh Hsu[6]; Simon Rothman[1]; Stefan Piperov[4]; William Patrick Mccormack[1]; Yongbin Feng[5]; Kevin Pedro[5]

[1] *Massachusetts Inst. of Technology (US)*
[2] *University of Washington (US)*
[3] *Univ. of California San Diego (US)*
[4] *Purdue University (US)*
[5] *Fermi National Accelerator Lab. (US)*
[6] *University of Washington Seattle (US)*

Corresponding Authors: yongbin.feng@cern.ch, william.patrick.mc.cormack.iii@cern.ch

In the as-as-service paradigm, we offload coprocessors to servers to run dedicated algorithms at high rates. The use of as-a-service allows us to balance computation loads leading to a dynamically resource-efficient system. Furthermore, as-a-service enables the integration of new types of coprocessors easily and quickly. In this talk, we present next generation studies using as-a-service computing, and we show the most recent performance of Intelligence Processing Units (IPUs), FPGAs, and how parallelized rule-based algorithms can also be implemented as-a-service quickly. We also show how we can optimize as-a-service to take into account network efficient inference strategies, including ragged batching. Finally, we propose a set of benchmarks that present real challenges and can enable us to understand how the future as-a-service landscape will evolve and how it can be used in recent scientific developments.

Contributed Talks / 37

Interaction Network Autoencoder in the Level-1 Trigger

Author: Sukanya Krishna[1]

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

Corresponding Author: sskrishn@ucsd.edu

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.

Contributed Talks / 38

EMD Neural Network Loss for ECON-T ASIC Autoencoder

Author: Rohan Shenoy[1]

Co-author: Javier Mauricio Duarte [1]

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

Corresponding Author: rohan.shenoy2911@gmail.com

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.

Contributed Talks / 39

Robust anomaly detection using NuRD

Authors: Aahlad Puli[1]; Abhijith Gandrakota[2]; Lily Zhang[1]

[1] *New York University*

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

Corresponding Author: abhijith.gandrakota@cern.ch

Anomaly Detection algorithms, when used as triggering mechanisms in experiments like the LHC, can help make data collection more precise by predominantly capturing events on interest. To ensure the triggering events are of interest, these detection algorithms should be robust against nuisance kinematic variables and detector conditions. To achieve this robustness, popular detection models, built via autoencoders for example, have to go through a decorrelation stage, where the anomaly thresholds for the scores are decorrelated with the nuisances; this post-training procedure sacrifices detection accuracy. We propose a class of robust anomaly detection technique that accounts for nuisances in the prediction, called Nuisance-Randomized Distillation (NuRD). Our nuisance-aware anomaly detection methods we build with NuRD have shorter inference times than autoencoder-based methods and do not require the extra decorrelation step (and therefore do not suffer the associated accuracy loss).

Contributed Talks / 40

Quantized Neural Networks on FPGAs using HAWQ-V3 and hl4ml

Author: Javier Ignacio Campos[1]

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

Corresponding Author: javier.ignacio.campos@cern.ch

Neural networks have been shown to be helpful in identifying events of interest in particle physics. However, to be used for live trigger decisions, they must meet demandingly low latencies and resource utilization for deployment on Field Programmable Gate Arrays (FPGAs). HAWQ-V3, a Hessian-based quantization-aware training framework, and hls4ml, an FPGA firmware implementation package, address these issues. HAWQ-V3 is a training framework enabling ultra-low and mixed-precision quantization. It introduced an approach to determining the relative quantization precision of each layer based on the layer's Hessian spectrum. More recently, it implements a computational graph with only integer addition, multiplication, and bit-shifting. We present a neural network classifier implemented with HAWQ-V3 for high-pT jets from simulations of LHC proton-proton collisions. We then introduce an extension for HAWQ-V3 to translate our classifier into the Quantized ONNX (QONNX) intermediate representation format, an extension of the Open Neural

Network Exchange (ONNX) format, supporting arbitrary-precision and low-precision neural networks. We demonstrate how the conversion of HAWQ-V3 models leverages the PyTorch Just-in-Time compiler to trace and translate models to QONNX operators. We then proceed to hls4ml to create firmware implementation of our quantized neural network and review its estimated latency and resource utilization for an FPGA.

Contributed Talks / 41

Design and first test results of a reconfigurable autoencoder on an ASIC for data compression at the HL-LHC

Authors: Alpana Shenai[1]; Chinar Syal[1]; Christian Herwig[1]; Cristina Ana Mantilla Suarez[1]; Cristinel Veniamin Gingu[2]; Danny Noonan[1]; Davide Braga[3]; Duje Coko[4]; Farah Fahim[5]; Giuseppe Di Guglielmo[5]; James Hoff[1]; Javier Mauricio Duarte[6]; Jennifer Ngadiuba[7]; Jim Hirschauer[1]; Jon Wilson[8]; Ka Hei Martin Kwok[1]; Llovizna Miranda[9]; Manuel Valentin[9]; Matteo Lupi[10]; Maurizio Pierini[10]; Nhan Tran[1]; Pamela Klabbers[11]; Paul Michael Rubinov[1]; Philip Coleman Harris[12]; Ralph Owen Wickwire[5]; Seda Memik[9]; Sioni Paris Summers[10]; Vladimir Loncar[None]; Xiaoran Wang[1]; Yingyi Luo[9]

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

[2] *Fermi National Accelerator laboratory*

[3] *FERMILAB*

[4] *University of Split. Fac.of Elect. Eng., Mech. Eng. and Nav.Architect. (HR)*

[5] *Fermilab*

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

[7] *FNAL*

[8] *Baylor University (US)*

[9] *Northwestern University*

[10] *CERN*

[11] *Fermi National Accelerator Laboratory*

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

Corresponding Author: daniel.christopher.noonan@cern.ch

The High Granularity Calorimeter (HGCAL) is a new subdetector of the CMS experiment in development as part of the upgrades for the High Luminosity LHC. The HGCAL readout system includes the Endcap Trigger Concentrator (ECON-T) ASIC, responsible for algorithmically reducing the immense data volume associated with the trigger patch of this six-million channel "imaging"calorimeter. To accomplish the data reduction, a reconfigurable autoencoder algorithm has been implemented in the ECON-T. The design, optimization, and implementation of this neural network encoder and first test results of the functionality within the ECON-T ASIC prototype are presented.

Contributed Talks / 42

In-System Parameter Update and I/O Capture for Machine Learning IP Cores

Author: Brett McMillian[1]

[1] *Crossfield Technology LLC*

Corresponding Author: brett.mcmillian@crossfieldtech.com

Crossfield Technology LLC is teaming with Fermi National Accelerator Laboratory (Fermilab) for a Department of Energy Small Business Innovation Research (SBIR) Phase II program to develop a

framework that remotely updates weights and biases in a High Level Synthesis (HLS) for machine learning (HLS4ML) IP core running on an Arria 10 SoC FPGA. The framework can also capture the inputs and outputs of the HLS4ML IP core and drive the inputs from software. The capabilities are easy to integrate into existing Quartus projects with minimal user configuration and enable users to rapidly test new algorithm parameters and acquire feedback.

Crossfield's framework consists of FPGA IP cores and software that runs under embedded Linux on the dual-core ARM Cortex-A9 Hard Processing System (HPS). The cores connect to the lightweight HPS to FPGA (LWH2F), FPGA to SDRAM (F2SDRAM), and FPGA to HPS IRQ (F2HIRQ) bridges, and predefined LWH2F address mappings ensure interoperability with the Linux device driver. The device driver software handles the setup of Direct Memory Access regions and control signals for the IP cores, and user applications access it as a character device.

During the second half of the Phase II program, Crossfield will be working with Fermilab to integrate the framework with an Experimental Physics and Industrial Control System (EPICS) Input/Output Controller (IOC) application. Crossfield is developing a graphical desktop application that will communicate with the embedded IOC application to provide remote access to the embedded framework. Our goal is to enable users to update weights and biases, capture ML inputs and outputs, and drive ML inputs in the FPGA fabric from across the local area network.

While the initial implementation focuses on the Arria 10 SoC, the design is portable to other Intel and Xilinx SoC FPGAs. Crossfield can provide flexible licensing options to meet customer needs and budgets. We also offer design services to assist customers with integration and customization of the software package, custom FPGA hardware design and custom board support package development.

Contributed Talks / 43

Neural Signal Compression System for a Seizure-Predicting Brain Implant in CMOS 28nm

Author: William Lemaire[None]

Co-authors: Berthié Gouin-Ferland ; Vincent Gauthier ; Esmaeil Ranjbar Koleibi ; Montassar Dridi ; Maher Benhouria ; Konin Koua ; Takwa Omrani ; Sébastien Roy ; Réjean Fontaine

Corresponding Author: william.lemaire@usherbrooke.ca

Recent advances in neuroscience tools allow recording brain signals with a large number of electrode channels. These tools allow to further develop an understanding of neural diseases and develop novel treatments for intractable conditions such as drug-resistant epilepsy or blindness induced by age-related macular degeneration.

Designing implantable systems with a high electrode count is challenging due to the large data rate for wireless transmission and the extremely limited power budget to avoid tissue damage. To mitigate this issue, we develop a neural recording ASIC in 28-nm CMOS with embedded neural spike compression which takes advantage of the sparse nature of neural coding. The compression system includes a spike detector, a dimensionality reduction algorithm, and a quantization algorithm.

To determine the optimal compromise between compression ratio, signal quality, and hardware resources utilization, we compare different dimensionality reduction algorithms (autoencoders, principal component analysis) and vector quantization algorithms (quantized neural networks, tree-structured vector quantizers, lattice quantizers). We implement the algorithms using Mentor Catapult and we synthesize with Cadence Genus to get power and area measurements. We evaluate the signal-to-noise and distortion ratio of the reconstructed signal, the synthesized area, and power. Finally, we compare the compression methods in terms of the spike sorting accuracy.

Successfully embedding this compression system into an ASIC allows to significantly increase the number of electrodes in a wireless system. This is a first step toward the development of a seizure-forecasting system for patients with refractory epilepsy.

Contributed Talks / 44

Rapid Generation of Kilonova Light Curves Using Conditional Variational Autoencoder

Author: Surojit Saha[1]

[1] *Institute of Astronomy, National Tsing Hua University, Taiwan*

Corresponding Author: surojitsaha@gapp.nthu.edu.tw

The discovery of the optical counterpart, along with the gravitational waves from GW170817, of the first binary neutron star merger, opened up a new era for multi-messenger astrophysics. The optical counterpart, designated as a kilonova (KN), has immense potential to reveal the nature of compact binary merging systems. Ejecta properties from the merging system provide important information about the total binary mass, the mass ratio, system geometry, and the equation of state of the merging system. In this study, a neural network has been applied to learn the optical light curves of the KN associated with GW170817 using data from Kasen model and we generate the light curves based on different ejecta properties such as lanthanide fraction, ejecta velocity and ejecta mass. For training the autoencoder, we use simulated KN light curves, where each light curve depends on ejecta mass, ejecta velocity and lanthanide fraction of the ejecta. We generated the light curves using our basic autoencoder code, which, as expected, is quite in agreement with the original light curves. Next, in order to verify the model built from the autoencoder, we apply denoising autoencoder to separate the noisy data from the real data. This process establishes the accuracy and robustness of the code. Following this we, develop conditional variational autoencoder (CVAE), which is for generating light curves based on the physical parameter of our choice. This flexibility was absent in the initial stages. Using the conditional variational autoencoder on simulated data and completing the training process, we generate light curves based on physical parameter of our choice. We have verified that, for a physical parameter present in the simulated data, the generated light curve for the same physical parameter is quite accurate with the original input light curve. This confirms that the code can now generate light curves for any random feasible physical parameter with satisfying accuracy. The timeline for generating the light curves using CVAE is very small, due to which this technique has the ability to replace time consuming and resource-draining simulations. Using the CVAE, we can look into the extremum detection limit associated with a KN model. Since there are several other factors that influences the KN, CVAE trained with simulated data from model with more detailed inclusion of physical parameters could give a more insight into the physics of KN. Currently, the CVAE is being trained on a different KN model and test on a separate data from another different model. This allows us to verify the variational aspect of the CVAE and get a more general look into the different KN light curves. The merit of this approach lies in its rapid generation of light curves based on desired parameters and at the same time encompass all the possible light curves related to KN.

Contributed Talks / 45

In-Pixel AI: From Algorithm to Accelerator

Authors: Priyanka Dilip[1]; Manuel Valentin[2]; Danny Noonan[3]; Giuseppe Di Guglielmo[4]; Seda Memik[2]; Nhan Tran[3]; Farah Fahim[4]

[1] *Stanford University / Fermilab*

[2] *Northwestern University*

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

[4] *Fermilab*

Corresponding Author: priyankadilip7@gmail.com

Ptychography is a technique for imaging an object through reconstruction of the diffraction of coherent photons. Through measuring these diffraction patterns across the whole of the object, small

scale structures can be reconstructed. In-pixel detectors used for these measurements, the maximum frame rate is often limited by the rate at which data can be transferred off of the device. In this talk, we will present an implementation for lossy data compression through a neural network Autoencoder and Principal Component Analysis integrated into a pixel detector. The 50x compression, together with placing the digital backend in parallel with the pixel array, is used to address major tradeoffs in area, latency, and congestion. The flow from algorithm specification in a high-level language, to High-Level Synthesis into hardware implementation in a 65nm technology, will be detailed. The improvements from these machine learning-based data compression will be analyzed in comparison with full readout and zero-suppression, also implemented in the same technology.

Contributed Talks / 46

Detection for Core-collapse Supernova and Fast Data Preprocessing

Author: Andy Chen[1]

[1] *National Yang Ming Chiao Tung University (NYCU)*

Corresponding Author: ra13phoenix@icloud.com

We want to use the WaveNet model for the detection of the gravitational wave signals from Core-collapse supernovas. The model is trained by the 3-D simulated core-collapse supernova waveforms injected into the background of Advanced LIGO detectors. The goal is to increase the efficiency of the model training and the hyperparameter tuning.

Contributed Talks / 47

Semi-supervised Graph Neural Networks for Pileup Noise Removal

Authors: Garyfallia Paspalaki[1]; Miaoyuan Liu[1]; Nhan Tran[2]; Pan Li[None]; Shikun Liu[None]; Tianchun Li[None]; Yongbin Feng[2]

[1] *Purdue University (US)*

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

Corresponding Author: liu2112@purdue.edu

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 of 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 distribution gap exists.

48

Welcome

Corresponding Authors: alessio.deiana@cern.ch, allison.renae.mc.carn@cern.ch

49

Machine Learning for Science

50

Efficient Computer Architectures

51

Applications and Opportunities at the Large Hadron Collider

Corresponding Author: maximilian.j.swiatlowski@cern.ch

52

Application and Opportunities in Nuclear Physics

Corresponding Author: jhuang@bnl.gov

53

Benchmarking Fast ML systems

54

Applications and Opportunities for Machine Controls

Corresponding Author: verena.kain@cern.ch

55

Applications and Opportunities in Fusion Systems

56

Autonomous experiments in scanning probe microscopy: opportunities for rapid inference and decision making

57

Toward Real Time Decision-Making in Material Science Experiments

58

Some current challenges in materials measurements: an industrial perspective

59

Beyond CMOS

60

Applications and Opportunities in Neutrino and DM experiments

Corresponding Authors: jonathan.asaadi@uta.edu, jonathan.asaadi@gmail.com

61

The need for low latency with billion-year old signals

Corresponding Author: mjg@caltech.edu

62

A Machine Learning Software Infrastructure for Gravitational Wave Signal Discovery

Authors: Alec Gunny[None]; Dylan Sheldon Rankin[1]; Eric Anton Moreno[2]; Erik Katsavounidis[3]; Ethan Marx[3]; Michael Coughlin[4]; Muhammed Saleem Cholayil[4]; Philip Coleman Harris[1]; Ryan Raikman[5]

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

[2] *Massachusetts Institute of Technology (US)*

[3] *MIT*

[4] *University of Minnesota*

[5] *Massachusetts Institute of Technology*

Corresponding Author: emarx@mit.edu

Machine Learning methods for gravitational wave signal discovery have shown a lot of promise. However, the current literature contains huge variability in algorithm design choices, from architecture selection to dataset engineering, making it very difficult to disentangle which design choices ultimately lead to better model performance. We present here an end to end software infrastructure for fast evaluation of ML algorithms aimed at detecting signals from compact binary coalescences (CBC's). The fully automated infrastructure allows us to alter any component of our algorithm, and quickly determine if these changes led to model improvements.

63

hls4ml tutorial

Corresponding Authors: sioni.paris.summers@cern.ch, elham.e.khoda@cern.ch

64

hls4ml community session

Contributed Talks / 67

Deep Neural Network Algorithms in the CMS Level-1 Trigger

Corresponding Author: aaportel@ucsd.edu

Contributed Talks / 68

Implementing Deep Neural Network Algorithms inside the CMS Level-1 Trigger

Corresponding Author: dhoang@mit.edu

Contributed Talks / 69

A Machine Learning Software Infrastructure for Gravitational Wave Signal Discovery

Corresponding Author: emarx@mit.edu

Machine Learning methods for gravitational wave signal discovery have shown a lot of promise. However, the current literature contains huge variability in algorithm design choices, from architecture selection to dataset engineering, making it very difficult to disentangle which design choices ultimately lead to better model performance. We present here an end to end software infrastructure for fast evaluation of ML algorithms aimed at detecting signals from compact binary coalescences (CBC's). The fully automated infrastructure allows us to alter any component of our algorithm, and quickly determine if these changes led to model improvements.