Fast Machine Learning for Science Workshop

Report of Contributions

https://indico.cern.ch/e/fml2020

Contribution ID: **1** Type: **Talk**

Matrix Element Regression with Deep Neural Networks – breaking the CPU barrier

Tuesday, 1 December 2020 15:02 (6 minutes)

The Matrix Element Method (MEM) is a powerful method to extract information from measured events at collider experiments. Compared to multivariate techniques built on large sets of experimental data, the MEM does not rely on an examples-based learning phase but directly exploits our knowledge of the physics processes. This comes at a price, both in term of complexity and computing time since the required multi-dimensional integral of a rapidly varying function needs to be evaluated for every event and physics process considered. This can be mitigated by optimizing the integration, as is done in the MoMEMta package, but the computing time remains a concern, and often makes the use of the MEM in full-scale analysis unpractical or impossible. We investigate in this paper the use of a Deep Neural Network (DNN) built by regression of the MEM integral as an ansatz for analysis, especially in the search for new physics.

Primary authors: BURY, Florian (UCLouvain - CP3); DELAERE, Christophe

Presenter: BURY, Florian (UCLouvain - CP3)

Contribution ID: **2** Type: **Talk**

Application of a neural network based technique for track identification in Nuclear Track Detectors (NTD)

Tuesday, 1 December 2020 14:44 (6 minutes)

Nuclear Track Detectors (NTDs) have been in use for decades, mainly as detectors of heavily ionizing particles. Existence of natural thresholds of detection makes them an ideal choice as detectors in the search for rare, heavily ionizing hypothesized particles (e.g. Monopoles, Strangelets etc.) against a large low-Z background in cosmic rays as well as particle accelerators. But identification of particle tracks in NTDs presents a significant challenge, with conventional image analysis software coming up short, requiring the intervention of human experts. This makes the job of scanning NTDs a painstakingly slow process, prone to human errors. In recent years, the use of Machine Learning techniques has opened up the possibilities of new advances in image analysis. In this work, we have taken a technique combining sequential application of convolution and de-convolution previously developed by us and further upgraded it with the use of Artificial Neural Network. This has further reduced the need for manual intervention, is producing better results than commercially available software and is promising to dramatically speed up the scanning process, thereby facilitating the more widespread adaptation of NTDs.

Primary author: Dr MAULIK, Atanu (University of Alberta (Now at INFN, Bologna))

Co-authors: Dr PALODHI, Kanik (University of Calcutta, Kolkata); Mr CHATTERJEE, Joydeep (University of Calcutta, Kolkata); Dr BHATTACHARYYA, Rupamoy (Bose Institute, Kolkata)

Presenter: Dr PALODHI, Kanik (University of Calcutta, Kolkata)

Contribution ID: **3** Type: **Talk**

Accelerating Graph Neural Networks on FPGAs for Particle Track Reconstruction using OpenCL and hls4ml

Wednesday, 2 December 2020 15:30 (8 minutes)

Current charged particle tracking algorithms at the CERN Large Hadron Collider (LHC) scale quadratically or worse with increasing number of overlapping proton-proton collisions in an event (pileup). As the LHC moves into its high-luminosity phase, pileup is expected to increase to an average of 200 overlapping collisions, highlighting the need for new algorithmic strategies. Recent work has shown that graph neural networks (GNNs) are well-suited to classifying segments of tracks. The real-time data filter at the LHC (L1 trigger) requires sub-microsecond latencies that can only be met by devices like field-programmable gate arrays (FPGAs). Accelerating neural networks on FPGAs facilitates energy efficient data-processing on large datasets with execution times that meet the L1 trigger latency requirements.

In this talk, we present two complementary FPGA implementations of an interaction network, a type of GNN, using OpenCL, an open-source framework for writing programs that execute across heterogenous acceleration platforms, and hls4ml, an open-source compiler of machine learning models into firmware. The OpenCL implementation adopts a CPU-plus-FPGA coprocessing approach where the CPU host program manages the application and all computational operations are accelerated using dedicated kernels deployed to the FPGA and take advantage of the FPGA hardware architecture to parallelize operations. The hls4ml implementation utilizes Xilinx high-level-synthesis tools to convert the GNN model to FPGA firmware making it suitable for both FPGA-only and co-processing applications. We will present comparisons of the two implementations in terms of their resource usage, latency, and tracking performance on the publicly-available TrackML benchmark dataset.

Primary authors: HEINTZ, Aneesh (Cornell University (US)); RAZAVIMALEKI, Vesal (Univ. of California San Diego (US)); DUARTE, Javier Mauricio (Univ. of California San Diego (US)); DEZOORT, Gage (Princeton University (US)); OJALVO, Isobel (Princeton University (US)); THAIS, Savannah Jennifer (Princeton University (US)); ATKINSON, Markus Julian (Univ. Illinois at Urbana Champaign (US)); NEUBAUER, Mark (Univ. Illinois at Urbana Champaign (US)); GRAY, Lindsey (Fermi National Accelerator Lab. (US))

Presenter: HEINTZ, Aneesh (Cornell University (US))

Contribution ID: **4** Type: **Talk**

A Quartus backend for hls4ml: deploying low-latency Neural Networks on Intel FPGAs

Monday, 30 November 2020 15:38 (6 minutes)

We describe the new Quartus backend of hls4ml, designed to deploy Neural Networks on Intel FPGAs. We list the supported network components and layer architectures (dense, binary/ternary, and convolutional neural networks) and evaluate its performance on a benchmark problem previously considered to develop the Vivado backend of hls4ml. We also introduce the support for recurrent layers and introduce a new asynchronous calling model to increase performance for larger models. In addition to that, we also demonstrate the use of this new model to optimize large-sparse networks.

Primary author: JAVED, Hamza (Pakistan Institute of Engin. and (PK))

Presenter: JAVED, Hamza (Pakistan Institute of Engin. and (PK))

Contribution ID: **5** Type: **Talk**

Autoencoders for anomaly detection in real-time at the LHC

Monday, 30 November 2020 14:40 (6 minutes)

At the LHC, data are collected at 40 MHz but only 1 kHz of data can be stored for physics studies. A typical LHC experiment operates a real-time selection system, that has to decide if an event should be stored or discarded. The first stage of this system, the L1 trigger, runs on custom electronic boards, mounting FPGAs. A L1 algorithm needs to operate within O(1μsec) latency. In this system, we aim to operate an unsupervised algorithm designed to identify outliers. Possibly highlighting the occurrence of new phenomena in LHC collisions. To this purpose, we design an autoencoder processing particle four momenta and we exploit hls4ml to deploy the model on an FPGA and evaluate its resource consumption and latency in various configurations.

Primary author: GOVORKOVA, Katya (CERN)

Presenter: GOVORKOVA, Katya (CERN)

Contribution ID: **6** Type: **Talk**

Ultra Low-latency, Low-area Inference Accelerators using Heterogeneous Deep Quantization with QKeras and hls4ml

Tuesday, 1 December 2020 14:52 (8 minutes)

While the quest for more accurate solutions is pushing deep learning research towards larger and more complex algorithms, edge devices demand efficient inference i.e. reduction in model size, speed and energy consumption. A technique to limit model size is quantization, i.e. using fewer bits to represent weights and biases. Such an approach usually results in a decline in performance. In this CERN-Google collaboration, we introduce a novel method for designing optimally heterogeneously quantized versions of deep neural network models for minimum-energy, high-accuracy, nanosecond inference and fully automated deployment on-chip. With a per-layer, per-parameter type automatic quantization procedure, sampling from a large base of quantizers, model energy consumption and size are minimized while high accuracy is maintained. This is crucial for the event selection procedure in proton-proton collisions at the CERN Large Hadron Collider, where resources are limited and a latency of O(1) micro second is required. Nanosecond inference and a resource consumption reduced by a factor of 50 when implemented on FPGA hardware is achieved.

Primary authors: AARRESTAD, Thea (CERN); SUMMERS, Sioni Paris (CERN); POL, Adrian Alan (CERN)

Presenter: AARRESTAD, Thea (CERN)

Contribution ID: **7** Type: **Talk**

Large and compressed Convolutional Neural Networks on FPGAs with hls4ml

Monday, 30 November 2020 14:58 (6 minutes)

We present ultra low-latency Deep Neural Networks with large convolutional layers on FPGAs using the hls4ml library. Taking benchmark models trained on public datasets, we discuss various options to reduce the model size and, consequently, the FPGA resource consumption: pruning, quantization to fixed precision, and extreme quantization down to binary or ternary precision. We demonstrate how inference latencies of O(10) micro seconds can be obtained while high accuracy is maintained

Primary author: LONCAR, Vladimir (CERN)

Presenter: LONCAR, Vladimir (CERN)

Contribution ID: **8** Type: **Talk**

Anomaly Detection with Spiking Neural Networks on Neuromorphic Chips

Wednesday, 2 December 2020 15:10 (8 minutes)

We describe anomaly detection applications on Neuromorphic Chips, exploiting Spiking Neural Networks on the Intel Loihi chip. We describe different workflows to train models directly on Loihi or to convert Neural Networks to Spiking Neural Networks. As a benchmark, we consider the problem of Gravitational Wave detection without a-priori assumption of the wave profile. We discuss baseline models and compare their reach to that of Spiking Neural Networks.

Primary authors: MORENO, Eric Anton (California Institute of Technology (US)); BORZYSZKOWSKI, Bartlomiej Pawel

Presenter: BORZYSZKOWSKI, Bartlomiej Pawel

Contribution ID: **9** Type: **Talk**

Adversarial mixture density network for particle reconstruction: a case study in collider simulation

Tuesday, 1 December 2020 14:36 (6 minutes)

An adversarial mixture density network (AMDN) with gaussian kernels is used to simulate muon reconstruction in the setup of collider detectors. The network is trained on events generated using Madgraph5, Pythia8 and the Delphes3 fast detector simulation implementation for the Compact Muon Solenoid (CMS). It is observed that the network can reproduce relevant kinematic distributions with a very good level of agreement, and at the same time the underlying correlations between reconstructed variables. Without prior collider-specific constraints,the trained network also acquires the azimuthal symmetry, a key feature in CMS simulation. While popular generative models, such as generataive adversarial networks (GANs), demonstrates wide success in various research areas, our work demonstrates that an alternative algorithmic approach more specific to Monte Carlo simulation in collider physics can be favourable and help tackle the increasing computing demands from simulation in collider experiments.

Primary author: LO, Kin Ho (University of Florida (US))

Presenter: LO, Kin Ho (University of Florida (US))

Contribution ID: **10** Type: **Talk**

A OneAPI backend of hls4ml to speed up Neural Network inference on CPUs

Monday, 30 November 2020 15:30 (6 minutes)

A recent effort to explore a neural network inference in FPGAs using High-Level Synthesis language (HLS), focusing on low-latency applications in triggering subsystems of the LHC, resulted in a framework called hls4ml. Deep Learning model converted to HLS using the hls4ml framework can be executed on CPUs, but have subpar performance. We present an extension of hls4ml using the new Intel oneAPI toolkit that converts deep learning models into high-performance Data Parallel C++ optimized for Intel x86 CPUs. We show that inference time on Intel CPUs is improved hundreds of times over previous HLS-based implementation, and several times over unmodified Keras/TensorFlow.

Primary author: LONCAR, Vladimir (CERN)

Presenter: LONCAR, Vladimir (CERN)

Contribution ID: **11** Type: **Talk**

muon detection using deep learning, applied to CONNIE events

Tuesday, 1 December 2020 15:38 (6 minutes)

The CONNIE experiment (Coherent Neutrino-Nucleus Interaction Experiment) is a collaboration from some countries in South America, EEUU and Switzerland . The data collected during the CONNIE experiment can be used to search for time variations of particles arriving at the detectors with periodic and stochastic nature. This experiment uses 12 high resistivity CCDs (Charge-Coupled Devices) placed in the vicinity of the Angra dos Reis nuclear reactor (Planta Almirante Alvaro Alberto, Rio de Janeiro, Brazil), with the purpose of detecting the antineutrinos generated in the reactor by measuring low-energy recoils from coherent elastic scattering (CEνNS). The sensors have recorded images of particles during the last 2 years in 3 hour expositions, where the majority of particles in the images are muon and beta particles that are considered as background. This work uses a deep learning algorithm to classify and detect muon particles in the images in order to remove them from the images for the purpose of neutrino studies, and also to build a time series that can be used as a stability monitor of the detection system.

Primary author: Mr BERNAL, Javier (Facultad De Ingenieria UNA)

Co-authors: Dr STALDER, Diego (Facultad de ingenieria UNA); Dr MOLINA, Jorge (Facultad de Ingenieria UNA)

Presenter: Mr BERNAL, Javier (Facultad De Ingenieria UNA)

Contribution ID: **12** Type: **Talk**

Quantifying DNA Damage in Comet Assay images using Neural Networks

Monday, 30 November 2020 14:30 (8 minutes)

Proton therapy for cancer treatment is a rapidly growing field and increasing evidence suggests it induces more complex DNA damage than photon therapy. Accurate comparison between the two treatments requires quantification of the damage caused, one method being the comet assay. The program outlined here is based on neural network architecture and aims to speed up analysis of comet assay images and provide accurate, quantified assessment of the DNA damage levels apparent in them.

The comet assay is an established technique in which DNA fragments are spread out under the influence of an electric field, producing a comet-like object. The elongation and intensity of the comet tail (consisting of DNA fragments) indicate the level of damage incurred. Many methods to measure this damage exist, using a variety of algorithms. These can be time consuming, so often only a small fraction of the comets available in an image are analysed. The automatic analysis presented here aims to improve this.

Object detection and localisation, implemented by a Mask-RCNN neural network, are used to perform instance segmentation of the comets. The identified comet instances are then saved as masks, which when overlaid onto the original image, provide pixel coordinates of the identified comets. A minimum accuracy of 90% has been achieved by the model in identifying comets in an image. The model has been trained via transfer learning from Microsoft's extensive COCO model, which is based on over 200,000 labelled images. This has significantly reduced both training time and also the number of images required for training (less than 70 images have been used here).

To supplement the training and testing of the network a Monte Carlo model is being developed in order to create simulated comet assay images.

Primary authors: DHINSEY, Selina; Prof. GREENSHAW, Tim (University of Liverpool, Physics Department); Dr PARSONS, Jason (University of Liverpool, Molecular and Clinical Cancer Medicine); Prof. WELSCH, Carsten (University of Liverpool, Physics)

Presenter: DHINSEY, Selina

Contribution ID: **13** Type: **Talk**

Development of ML FPGA filter for particle identification with transition radiation detector.

Wednesday, 2 December 2020 14:50 (8 minutes)

Transition Radiation Detectors (TRD) have the attractive features of being able to separate particles by their gamma factor. A new TRD development, based on a GEM technology, is being carried out as a R&D project for the future Electron Ion Collider (EIC) and for upgrade of the GlueX experiment. This detector combines a high precision GEM tracker with TRD functionality and optimized for electron identification.
Modern concepts of trigger-less readout and data streaming will produce a very large data volume to be read from detectors. From a resource standpoint, it appears strongly advantageous to perform both the pre-processing of data and data reduction at earlier stages of a data acquisition. Following this trend, we began to develop an FPGA based Machine Learning algorithm for a real-time particle identification with GEMTRD. This research is important for streaming readout systems being developed now at JLab for EIC. The report will describe first steps in the development of ML-FPGA filter for GEMTRD.

Primary authors: DICKOVER, Cody (Jefferson Lab); FANELLI, Cristiano (MIT); LAWRENCE, David (Jefferson Lab); ROMANOV, Dmitry; BARBOSA, Fernando (Thomas Jefferson National Accelerator Facility); BELFORE, Lee (ODU); JOKHOVETS, Lioubov (FZJ); FURLETOV, Sergey (Jefferson Lab); FURLETOVA, Yulia (Jefferson Lab)

Presenter: FURLETOV, Sergey (Jefferson Lab)

Contribution ID: **14** Type: **Talk**

Convolutional Neural Network Fast Inference Deployment on FPGAs

Monday, 30 November 2020 15:06 (6 minutes)

From self-driving cars to particle physics, the uses of convolutional neural networks are plentiful. To greatly decrease inference latency, CNNs and other deep learning architectures can be deployed to hardware compute environments in the form of Field Programmable Gate Arrays (FPGAs). The open source package HLS4ML is leveraged to complete model conversion and RTL synthesis. The work presented here describes methods with which the generated Verilog/VHDL can be further optimized to yield further latency reductions and smaller hardware resource requirements.

Primary author: REIS, Andrew Harmon (Southern Methodist University (US))

Presenter: REIS, Andrew Harmon (Southern Methodist University (US))

Contribution ID: **15** Type: **Talk**

Making ML easier at CERN with Kubeflow

Tuesday, 1 December 2020 14:10 (8 minutes)

Different groups at CERN have been focusing on changing existing workflows and processes to rely on machine learning, covering trigger farms, fast simulation, anomaly detection, reinforcement learning, etc.

To help end users in these tasks a service must hide the underlying infrastructure complexity and integrate well with existing identity and storage services, as well as easing the tasks of data preparation, model training, serving, among others.

In this talk we present a new solution available at CERN based on Kubeflow, the ML platform running on top of Kubernetes. We describe how the underlying resources - CPUs and GPUs - are offered to the end user hiding the complex details that allow the service to scale horizontally, and shared with the goal of optimizing resource usage. We present how existing on-premise capacity can be extended to external resources (public clouds) without users realizing, and for use cases where on-demand usage is cost effective such as covering for peak periods.

In the second part of the talk we cover the complete ML lifecycle. Examples will include quick code development and iteration using notebooks; submission of analysis pipelines allowing workloads to easily scale out, and including the direct conversion of a notebook to a pipeline; distributed model training with submission via both a web interface and an API; hyper-parameter tuning support with multiple search algorithms available; and finally model storage and serving.

Primary authors: BRITO DA ROCHA, Ricardo (CERN); GOLUBOVIC, Dejan (CERN)

Presenter: GOLUBOVIC, Dejan (CERN)

Contribution ID: **16** Type: **Talk**

Using an Optical Processing Unit for tracking and calorimetry at the LHC

Tuesday, 1 December 2020 14:20 (6 minutes)

Experiments at HL-LHC and beyond will have ever higher read-out rate. It is then essential to explore new hardware paradigms for large scale computations. We have considered the Optical Processing Unit (OPU) from LightOn https://lighton.ai , which is an analog device to multiply a binary 1 mega pixel image by a (fixed) 1E6x1E6 random matrix, resulting in a mega pixel image, at a 2kHz rate. It could be used for the whole branch of Machine Learning using random matrix in particular for dimensionality reduction. In this talk, we have explored the potential of OPU for two typical HEP use cases:

1) "Tracking": high energy proton collisions at the LHC yield billions of records with typically 100,000 3D points corresponding to the trajectory of 10.000 particles. Using two datasets from previous tracking challenges, we investigate the OPU potential to solve similar or related problems in high-energy physics, in terms of dimensionality reduction, data representation, and preliminary results.

2) "Calorimeter Event classification": high energy proton collision at the Large Hadron Collider have been simulated, each collision being recorded as an image representing the energy flux in the detector. The task is to train a classifier to separate signal from the background. The OPU allows fast end-to-end classification without building intermediate objects (like jets). This technique is presented, compared with more classical particle physics approaches.

Primary authors: ROUSSEAU, David (IJCLab-Orsay); Dr BASARA, Laurent (LAL/LRI, Université Paris Saclay); GHOSH, Aishik (Université Paris-Saclay (FR)); BISWAS, Biswajit (Centre National de la Recherche Scientifique (FR))

Presenter: ROUSSEAU, David (IJCLab-Orsay)

Contribution ID: **17** Type: **Talk**

Design of a reconfigurable autoencoder algorithm for detector front-end ASICs

Monday, 30 November 2020 14:48 (8 minutes)

The next generation of particle detectors will feature unprecedented readout rates and require optimizing lossy data compression and transmission from front-end application-specific integrated circuits (ASICs) to the off-detector trigger processing logic. Typically, channel aggregation and thresholding are applied, removing information useful for particle reconstruction in the process. A new approach to this challenge is directly embedding machine learning (ML) algorithms in ASICs on the detector front-end to allow intelligent data compression before transmission. We present an algorithm optimized for the High-Granularity Endcap Calorimeter (HGCal) installed in the CMS Experiment for the high-luminosity upgrade to the Large Hadron Collider. We trained a neural-network (NN) autoencoder to achieve optimal compression fidelity for physics reconstruction while respecting hardware constraints on internal parameter precisions, computational (circuit) complexity, and area footprint. The autoencoder improves over non-ML algorithms in reconstructing low-energy signals in high-occupancy environments. Quantization-aware training is performed using qKeras and is implemented in RTL using the hls4ml compiler tool. Finally, we discuss our solution's flexibility, wherein sensors may be individually tuned to optimize performance across the full detector and over the range of expected run conditions during the detector's lifetime.

Primary authors: BLANCO VALENTÍN, Manu (Northwestern University); DI GUGLIELMO, Giuseppe (Columbia University); FAHIM, Farah (Fermilab); HAWKS, Benjamin (Fermi National Accelerator Laboratory); HERWIG, Christian (Fermi National Accelerator Lab. (US)); HIRSCHAUER, Jim (Fermi National Accelerator Lab. (US)); KWOK, Ka Hei Martin (Brown University (US)); LUO, Yingyi (Northwestern University); NOONAN, Daniel (Florida Institute of Technology (US)); OGRENCI MEMIK, Seda (Northwestern University); TRAN, Nhan Viet (Fermi National Accelerator Lab. (US))

Presenter: DI GUGLIELMO, Giuseppe (Columbia University)

Contribution ID: **18** Type: **Talk**

Convolutional Neural Networks for real-time processing of ATLAS Liquid-Argon Calorimeter signals with FPGAs

Monday, 30 November 2020 15:14 (6 minutes)

Physicists use the Large Hadron Collider (LHC) at CERN/Geneva to create proton-proton (pp) collisions to study rare particle-physics processes at high energies. Within the Phase-II upgrade, the LHC and the particle detectors will be prepared for high luminosity operation, starting in 2027. One challenge is the high level of signal pile-up caused by up to 200 simultaneous pp collisions. Moreover, in the case of the Liquid-Argon (LAr) Calorimeters of the ATLAS detector, the signals of up to 25 subsequent collisions overlap, which further increases the difficulty to reconstruct the energy deposit in the detector.

In order to cope with this, the readout electronics of the ATLAS LAr Calorimeters will be upgraded, which will allow a real-time processing of the full sequence of digitized pulses sampled at 40 MHz. Conventional signal processing applies an optimal filter to reconstruct the energy of the detector hits. However, the high level of pile-up and a new trigger scheme requires a more advanced signal reconstruction method.

We have developed a dilated convolutional neural network (CNN) which improves the efficiency to identify significant energy deposits above a given noise threshold and which reduces the number of incorrectly identified hits when compared to an optimal filter. Since the implementation target of the CNN is a Field Programmable Gate Array (FPGA), the number of parameters and the mathematical operations are well controlled. A second network structure aims at reconstructing the hit energy, using the information of the hit identification network. The CNN training data are generated by a dedicated simulation program, called AREUS, which provides realistic signal sequences including all noise sources.

Moreover, we implemented the CNN structure in firmware in an automated way, translating the CNN training output file into VHDL, targeting an INTEL Stratix-10 FPGA. Linearized sigmoid activation functions are tested and compared to the full-precision calculation. Very good agreement between FPGA and computer based calculations is observed. We also analyzed the FPGA resource usage and the maximum frequency at which the algorithm can be executed.

The presentation will summarize the latest performance results obtained with the CNN approach and the most recent prototype implementations in FPGA firmware.

Primary authors: BERTHOLD, Anne-Sophie (Technische Universitaet Dresden (DE)); FRITZSCHE, Nick (Technische Universitaet Dresden (DE)); Mr HENTGES, Rainer (Technische Universitaet Dresden (DE)); VOIGT, Johann Christoph (Technische Universitaet Dresden (DE)); STRAESSNER, Arno (Technische Universitaet Dresden (DE))

Presenters: BERTHOLD, Anne-Sophie (Technische Universitaet Dresden (DE)); FRITZSCHE, Nick (Technische Universitaet Dresden (DE))

Contribution ID: **19** Type: **Talk**

FastCaloGAN: a tool for fast simulation of the ATLAS calorimeter system with Generative Adversarial Networks

Monday, 30 November 2020 15:22 (6 minutes)

Building on the recent success of deep learning algorithms, Generative Adversarial Networks (GANs) are exploited for modelling the response of the ATLAS detector calorimeter to different particle types and simulating calorimeter showers for photons, electrons and pions over a range of energies (between 256˜MeV and 4˜TeV) in the full detector η range. The properties of showers in single-particle events and of jets in di-jets events are compared with full detector simulation performed by GEANT4. The good performance of FastCaloGAN demonstrates the potential of GANs to perform a fast calorimeter simulation for the ATLAS experiment.

Primary author: FAUCCI GIANNELLI, Michele (INFN e Universita Roma Tor Vergata (IT))

Presenter: FAUCCI GIANNELLI, Michele (INFN e Universita Roma Tor Vergata (IT))

Contribution ID: **20** Type: **Talk**

Level 1 trigger track quality machine learning models on FPGAs for the Phase 2 upgrade of the CMS experiment

Tuesday, 1 December 2020 14:28 (6 minutes)

In 2026, the LHC will be upgraded to the HL-LHC which will provide up to 10 times as many proton-proton collisions per bunch crossing. In order to keep up with the increase in data rates, the CMS collaboration is updating the Level 1 Trigger system to run particle selection and reconstruction algorithms on FPGAs in real-time with the data collection system. One such particle algorithm measures the quality of the reconstructed tracks to classify them as "real" or "fake" reconstructed tracks. In this work, we develop supervised machine learning algorithms for track quality classification and test these models on simulated FPGAs using the HLS4ML and Conifer open-source packages.

Primary author: SAVARD, Claire (University of Colorado Boulder (US))

Presenter: SAVARD, Claire (University of Colorado Boulder (US))

Contribution ID: **21** Type: **Talk**

An Early Exploration into the Interplay between Quantization and Pruning of Neural Networks

Tuesday, 1 December 2020 15:10 (8 minutes)

Machine Learning (ML) is already being used as a powerful tool in High Energy Physics, but the typically high computational cost associated with running ML workloads is often a bottleneck in data processing pipelines. Even on high performance hardware such as Field Programmable Gate Arrays (FPGAs) and Application Specific Integrated Circuits (ASICs) the speed and size of these models are often heavily constrained by available hardware resources. Various model optimization techniques, such as pruning and quantization, have been used in an attempt to alleviate the high performance costs, but often not together, or without fully understanding the interplay of the two techniques. We attempt to explore this interplay between quantization and pruning in order to better understand how they interact. Targeting FPGAs and ASICs, we attempt to determine how to yield the best performance with both quantization and pruning. In this presentation, we explore these techniques by optimizing the HLS4ML 3 Hidden Layer Jet Substructure tagging model, finding that we can successfully optimize the model down to 3-5% of its original size while retaining comparable performance to the original network. Finally, we discuss some next steps into understanding how the different optimization techniques affect the model internally, beyond standard performance metrics.

Primary authors: Mr HAWKS, Benjamin (Fermi National Accelerator Laboratory); TRAN, Nhan Viet (Fermi National Accelerator Lab. (US)); DUARTE, Javier Mauricio (Univ. of California San Diego (US))

Presenter: Mr HAWKS, Benjamin (Fermi National Accelerator Laboratory)

Contribution ID: **22** Type: **Talk**

Real-time Artificial Intelligence for Accelerator Control: A Study at the Fermilab Booster

Tuesday, 1 December 2020 15:20 (6 minutes)

We describe a method for precisely regulating the gradient magnet power supply (GMPS) at the Fermilab Booster accelerator complex using a neural network (NN). We demonstrate preliminary results by training a surrogate machine-learning model on real accelerator data, and using the surrogate model in turn to train the NN for its regulation task. We additionally show how the neural networks that will be deployed for control purposes may be compiled to execute on field-programmable gate arrays (FPGAs). This capability is important for operational stability in complicated environments such as an accelerator facility.

Primary authors: HERWIG, Christian (Fermi National Accelerator Lab. (US)); DUARTE, Javier Mauricio (Univ. of California San Diego (US)); TRAN, Nhan Viet (Fermi National Accelerator Lab. (US)); QUINTERO PARRA, Andres Felipe (Fermi National Accelerator Lab. (US)); ST. JOHN, Jason (FNAL); KAFKES, Diane (FNAL); PELLICO, William (FNAL); PERDUE, Gabriel (FNAL); SCHUPBACH, Brian (FNAL); SEIYA, Kiyomi (FNAL); HUANG, Yunzhi (PNNL); SCHRAM, Malachi (PNNL); KELLER, Rachael (Columbia)

Presenter: HERWIG, Christian (Fermi National Accelerator Lab. (US))

Contribution ID: **23** Type: **not specified**

Neutrino Astrophysics

Presenter: SCHOLBERG, Kate (Duke University)

Session Classification: Talks

Contribution ID: **24** Type: **not specified**

Welcome and Orientation

Monday, 30 November 2020 10:00 (5 minutes)

Presenter: DEIANA, Allison Mccarn (Southern Methodist University (US))

Contribution ID: **25** Type: **not specified**

Welcome from SMU

Monday, 30 November 2020 10:05 (5 minutes)

Presenter: Dr CLOWARD, Karisa (Southern Methodist University)

Contribution ID: **26** Type: **not specified**

Overview of Workshop

Monday, 30 November 2020 10:10 (5 minutes)

Presenter: TRAN, Nhan (Fermi National Accelerator Lab. (US))

Contribution ID: **27** Type: **not specified**

Neutrino Astrophysics

Wednesday, 2 December 2020 10:40 (30 minutes)

Presenters: SCHOLBERG, Kate (Duke University); SCHOLBERG, Kate (Duke University)

Contribution ID: **28** Type: **not specified**

Trends in Computer Architectures

Monday, 30 November 2020 10:20 (30 minutes)

Presenter: BLOTT, Michaela (Xilinx Research)

Contribution ID: **29** Type: **not specified**

Accelerator-based Neutrinos

Wednesday, 2 December 2020 10:00 (30 minutes)

Presenter: HEWES, Jeremy Edmund (University of Cincinnati (US))

Contribution ID: **30** Type: **not specified**

Efficient Neural Network Training and Inference

Monday, 30 November 2020 11:15 (30 minutes)

Presenter: GHOLAMI, Amir

Contribution ID: **31** Type: **not specified**

Cosmology

Wednesday, 2 December 2020 11:35 (30 minutes)

Presenter: NTAMPAKA, Michelle (STSCI)

Contribution ID: **32** Type: **not specified**

Imaging: Electron Microscopy

Monday, 30 November 2020 12:55 (30 minutes)

Presenter: AGAR, Josh (Lehigh University)

Contribution ID: **33** Type: **not specified**

Deep Learning Acceleration of Progress in Fusion Energy Research

Tuesday, 1 December 2020 13:15 (30 minutes)

Accelerated progress in delivering accurate predictions in science and industry have been accomplished by engaging advanced statistical methods featuring artificial intelligence/deep learning/machine learning (AI/DL/ML). Associated techniques have enabled new avenues of data-driven discovery in key scientific applications areas such as the quest to deliver Fusion Energy –identified by the 2015 CNN "Moonshots for the 21st Century"televised series as one of 5 prominent grand challenges for the world today. An especially time-urgent and challenging problem facing the development of a fusion energy reactor is the need to reliably predict and avoid large-scale major disruptions in magnetically-confined tokamak systems such as the EUROFUSION Joint European Torus (JET) today and the burning plasma ITER device in the near future – – a ground-breaking $25B international burning plasma experiment with the potential capability to exceed "breakeven" fusion power by a factor of 10 or more with "first plasma"targeted for 2026 in France. Meanwhile, a key challenge is to deliver significantly improved methods of prediction with better than 95% predictive accuracy to provide advanced warning for disruption avoidance/mitigation strategies to be effectively applied before critical damage can be done to ITER

This presentation describes advances in the deployment of deep learning recurrent and convolutional neural networks in Princeton's Deep Learning Code – "FRNN"–that have enabled the rapid analysis of large complex datasets on supercomputing systems that have accelerated progress in predicting tokamak disruptions with unprecedented accuracy and speed (Ref. "NATURE,"(April 26, 2019). This represented the first adaptable predictive DL software trained on leadership class systems to deliver accurate predictions for disruptions across different tokamak devices (DIII-D in the US and JET in the UK) –with the unique capability to carry out efficient "transfer learning"via training on a large data base from one experiment (i.e., DIII-D) and be able to accurately predict disruption onset on an unseen device (i.e., JET) ! Moreover, in recent advances, the FRNN inference engine has recently been deployed in a real-time plasma control system on the DIII-D tokamak facility in San Diego,CA. This opens up exciting avenues for moving from passive disruption prediction to active real-time control with subsequent optimization for reactor scenarios.

Presenter: TANG, Bill (Princeton University)

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

ASICs and Circuits

Wednesday, 2 December 2020 13:15 (30 minutes)

Presenter: DELBRUCK, Tobi (ETH Zurich)

Contribution ID: **35** Type: **not specified**

Health Sensing, Detection, and Monitoring

Tuesday, 1 December 2020 10:40 (30 minutes)

Primary author: SEN, Sougata (Northwestern University)

Co-author: ALSHURAFA, Nabil (Northwestern University)

Presenters: ALSHURAFA, Nabil (Northwestern University); SEN, Sougata (Northwestern University)

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

Electron-Ion Collider

Wednesday, 2 December 2020 13:55 (30 minutes)

Presenter: DIEFENTHALER, Markus (Jefferson Lab)

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

Beyond CMOS

Tuesday, 1 December 2020 11:35 (30 minutes)

Presenter: STRUKOV, Dimitri (UCSB)

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

HPC

Tuesday, 1 December 2020 10:00 (30 minutes)

Presenter: KALESCKY, Robert (Southern Methodist University)

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

Large Hadron Collider

Monday, 30 November 2020 13:35 (30 minutes)

Presenter: NGADIUBA, Jennifer (CERN)

Contribution ID: **40** Type: **Talk**

AI-assisted Tracking Algorithm

Wednesday, 2 December 2020 15:00 (8 minutes)

In this work we describe the development of machine learning models to assist the CLAS12 detector tracking algorithm. Several networks were implemented to assist tracking algorithm to overcome drift chambers inefficiencies using auto-encoders to de-noise wire chamber signals and corruption detection.A classifier network was used to identify track candidates from numerous combinatorial segments using different types of networks including: Convolutional Neural Networks (CNN), Multi-Layer Perceptron (MLP) and Extremely Randomized Trees (ERT). The final implementation provided an accuracy >99%. The implementation of AI assisted tracking into the CLAS12 reconstruction workflow and provided code speedup of up to 4 times.

Primary author: GAVALIAN, Gagik (Jefferson Lab)

Presenter: GAVALIAN, Gagik (Jefferson Lab)

Contribution ID: **41** Type: **Talk**

Deep Learning based acceleration of Gravitational Waves

Wednesday, 2 December 2020 15:20 (8 minutes)

In gravitational-wave detectors, regression techniques are applied to remove noise artifacts in order to improve the ability to observe and extract information from astrophysics signals. We present a deep learning-based noise regression method called DeepClean that can subtract linear and nonlinear noise in gravitational-wave data from the Advanced LIGO detectors. We also discuss our work toward a new computing model in gravitational-wave data analysis where GPU and FPGA acceleration on machine learning inference can be deployed on an as-a-service basis. We use DeepClean as a use-case for exploring such computing models in order to achieve real-time capabilities and overall flexibility such models provide.

Primary authors: GUNNY, Alec; HARRIS, Philip Coleman (Massachusetts Inst. of Technology (US))

Presenter: GUNNY, Alec

Contribution ID: **42** Type: **Talk**

SONIC: Coprocessors as a service for deep learning inference in high energy physics

Wednesday, 2 December 2020 15:40 (6 minutes)

In the next decade, the demands for computing in large scientific experiments are expected to grow tremendously. During the same time period, CPU performance increases will be limited. At the CERN Large Hadron Collider (LHC), these two issues will confront one another as the collider is upgraded for high luminosity running. Alternative processors such as graphics processing units (GPUs) can resolve this confrontation provided that algorithms can be sufficiently accelerated. In many cases, algorithmic speedups are found to be largest through the adoption of deep learning algorithms. We present a comprehensive exploration of the use of GPU-based hardware acceleration for deep learning inference within the data reconstruction workflow of high energy physics. We present several realistic examples and discuss a strategy for the seamless integration of coprocessors so that the LHC can maintain, if not exceed, its current performance throughout its running.

Primary authors: KRUPA, Jeffrey (Massachusetts Institute of Technology); HARRIS, Philip Coleman (Massachusetts Inst. of Technology (US))

Presenter: RANKIN, Dylan Sheldon (Massachusetts Inst. of Technology (US))

Contribution ID: **43** Type: **Talk**

GPU-accelerated machine learning inference as a service for computing in neutrino experiments

Tuesday, 1 December 2020 15:46 (6 minutes)

Machine learning algorithms are becoming increasingly prevalent and performant in the reconstruction of events in accelerator-based neutrino experiments. These sophisticated algorithms can be computationally expensive. At the same time, the data volumes of such experiments are rapidly increasing. The demand to process billions of neutrino events with many machine learning algorithm inferences creates a computing challenge. We explore a computing model in which heterogeneous computing with GPU coprocessors is made available as a web service. The coprocessors can be efficiently and elastically deployed to provide the right amount of computing for a given processing task. With our approach, Services for Optimized Network Inference on Coprocessors (SONIC), we integrate GPU acceleration specifically for the ProtoDUNE-SP reconstruction chain without disrupting the native computing workflow. With our integrated framework, we accelerate the most time-consuming task, track and particle shower hit identification, by a factor of 17. This results in a factor of 2.7 reduction in the total processing time when compared with CPU-only production. For this particular task, only 1 GPU is required for every 68 CPU threads, providing a cost-effective solution.

Primary authors: HARRIS, Philip Coleman (Massachusetts Inst. of Technology (US)); WANG, Mike

Presenter: WANG, Mike

Contribution ID: **44** Type: **Talk**

Building the tools to run large scale machine learning with FPGAs with two new approaches: AIGEAN and FAAST

Tuesday, 1 December 2020 15:28 (8 minutes)

FPGA programming is becoming easier as the vendors begin to provide environments, such as for machine learning (ML), that enable programming at higher levels of abstraction.The vendor platforms target FPGAs in a single host server.To scale to larger systems of FPGAs requires communication through the hosts, which has a significant impact on performance. We demonstrate the deployment of ML algorithms on single FPGAs through FAAST a newly developed FPGA based infrastructure framework. We also present a new Framework, AIGEAN, to run multiple FPGA and CPU heterogeneous system that can leverage direct FPGA-to-FPGA communication links. AIgean and FAAST, take as input an ML algorithm created with a standard ML framework and a specification of the available FPGA and CPU resources. The outputs are software and hardware cores that can compute one or more ML layers. These layers can be distributed across a heterogeneous cluster of CPUs and FPGAs for execution. As part of this work we present an optimized FPGA implementation of a CNNs. We show that in some cases FPGAs can exceed the performance of other accelerators, including GPUs.

Primary authors: RANKIN, Dylan Sheldon (Massachusetts Inst. of Technology (US)); HARRIS, Philip Coleman (Massachusetts Inst. of Technology (US))

Presenter: TARAFDAR, Naif (University of Toronto)

Contribution ID: **45** Type: **not specified**

Acknowledgments

Wednesday, 2 December 2020 11:18 (2 minutes)

Presenter: HARRIS, Philip Coleman (Massachusetts Inst. of Technology (US))

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

hls4ml Tutorial

Thursday, 3 December 2020 09:00 (3 hours)

Presenter: SUMMERS, Sioni Paris (CERN)

Session Classification: Tutorial Session

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

hls4ml Tutorial

Thursday, 3 December 2020 13:00 (3 hours)

Presenter: SUMMERS, Sioni Paris (CERN)

Session Classification: Tutorial Session