2nd IML Machine Learning Workshop

Report of Contributions

https://indico.cern.ch/e/IML2018

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

Research on EOS File storage Strategy Based on Access Characteristics Using Machine Learning Method

Machine learning has been an attractive topic in high-energy physics field for many years. For example, machine learning algorithms devoted to the reconstruction of particle tracks or jets in high energy physics experiments. EOS is an open source parallel distributed file system. It has been generally used in large scale cluster computing for both physics and user use cases at IHEP, like LHAASO and CEPC. The EOS design has included a split of hot and cold storage by defining groups. It records the frequency of file visits in the past period. Nevertheless, EOS cannot pick the most valuable data for users. Some files are frequently accessed since their creation and they'd better be placed at hot storage. Some files are accessed by a single user and used not so often. They'd better be placed at cold storage. Different files may have common access characteristics. For example, system log files are continuously written. Their pointers always seek for last byte of the files. We assume that other file clusters have such unique read or write characteristics too. So all files could be divided into several categories based on different access characteristics.We used some clustering algorithms, like Birch and Mini-batch K-means, to mine users'file access mode. We communicated with EOS users about file categories. After that, a supervised Machine Learning classification was built. Given a file, it searched this file's access log from EOS FST, compared its access features with different file categories, and predicted which category it belonged to. We wrote MapReduce tasks to process file access characteristics and saved them in HBase. In this paper we discussed two classify algorithms: RandomForest and LSTM. Then several data storage strategies were designed for each file category. The strategies included EOS storage group selection, the number of file copies and redundancy level. In addition, EOS's file scheduler would be redesigned as a plug-in to support file classification and category forecast. This paper will also present several test cases and LHAASO's sample files. More functional and performance tests are in progress.

Intended contribution length

20 minutes

Authors: CHENG, Zhenjing (INSTITUE OF HIGH ENERGY PHYSICS); Mr HU, Qingbao (IHEP); LI, Haibo (Institute of High Energy Physics Chinese Academy of Science); CHENG, Yaodong (IHEP); CHEN, Gang (INSTITUTE OF HIGH ENERGY PHYSICS)

Presenter: CHENG, Zhenjing (INSTITUE OF HIGH ENERGY PHYSICS)

Session Classification: Session 1

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

Training Deep Learning Models on Many-Core Processors

Leveraging on our previous work on developing DNN-based classification models for Higss events [1], we turn to CNN-based classification models for muon events. Using Intel Knights Landing (KNL) processors, we present performance metrics on training convolutional neural networks (CNNs) on multiple KNL computing nodes for the task of muon identification (i.e “high Pt” or “low Pt”). This work is an improvement over previous similarly tasked workload of using deep neural networks (DNNs) for higgs identification (i.e. “higgs” or “background”).

Intended contribution length

20 minutes

Authors: Mr OJIKA, David Nonso (University of Florida (US)); Mr VASISHTA, Akash (University of Florida); Ms PRASAD, Chandana (University of Florida); Mr JIANG, Chao (University of Florida)

Co-authors: Dr HERMAN, Lam (University of Florida); Dr ACOSTA, Darin (University of Florida); Dr GORDON-ROSS, Ann (University of Florida)

Presenter: Mr OJIKA, David Nonso (University of Florida (US))

Session Classification: Session 5

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

Drones: Making faster and smarter decisions with software triggers

Wednesday, April 11, 2018 4:55 PM (20 minutes)

Data collection rates in high energy physics (HEP), particularly those at the Large Hadron Collider (LHC) are a continuing challenge and require large amounts of computing power to handle. For example, at LHCb an event rate of 1 MHz is processed in a software-based trigger. The purpose of this trigger is to reduce the output data rate to manageable levels, which amounts to a reduction from 60 GB per second to an output data rate of 0.6 GB per second. Machine learning (ML) is becoming an evermore important tool in the data reduction, be it with the identification of interesting event topologies, or the distinction between individual particle species. For the case of LHCb data-taking, over 600 unique signatures are searched for in parallel in real time, each with its own set of requirements. However, only a handful at present make use of machine learning, despite the large ecosystem. Often the reason for this is the relative difficulty in the application of a preferred ML classifier to the C++/Python combination of event selection frameworks. One way to overcome this is to use an approximate network known as a drone that can learn the features of your preferred form and can be executed in an easily parallelisable way. We present the uses and advantages of such an approach.

Intended contribution length

20 minutes

Authors: BENSON, Sean (Nikhef National institute for subatomic physics (NL)); GIZDOV, Konstantin (The University of Edinburgh (GB))

Presenter: BENSON, Sean (Nikhef National institute for subatomic physics (NL))

Session Classification: Session 6

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

What is the machine learning.

Tuesday, April 10, 2018 4:30 PM (20 minutes)

Applications of machine learning tools to problems of physical interest are often criticized for producing sensitivity at the expense of transparency. In this talk, I explore a procedure for identifying combinations of variables – aided by physical intuition – that can discriminate signal from background. Weights are introduced to smooth away the features in a given variable(s). New networks are then trained on this modified data. Observed decreases in sensitivity diagnose the variable's discriminating power. Planing also allows the investigation of the linear versus non-linear nature of the boundaries between signal and background. I will demonstrate these features in both an easy to understand toy model and an idealized LHC resonance scenario.

Intended contribution length

20 minutes

Authors: CHANG, Spencer (University of Oregon); COHEN, Tim (University of Oregon); OSTDIEK, Bryan (University of Oregon)

Presenter: OSTDIEK, Bryan (University of Oregon)

Session Classification: Session 4

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

Generating high-level physics variables based on Monte Carlo simulated ttH events using Wasserstein GANs

Developing and building an analysis in high energy particle physics requires a large amount of simulated events. Simulations at the LHC are usually complex and computationally intensive due to sophisticated detector architectures. In this context, Generative Adversarial Networks (GANs) have recently caught a wide interest. GANs can learn to generate complex data distributions and produce samples up to 5 orders of magnitude faster than well-established simulations.
The recently introduced Wasserstein GAN (WGAN) further improves and stabilizes the training process of generative models. In this talk we present the results of a WGAN trained to produce a set of high-level physics variables in the context of top-quark pair associated Higgs boson production (ttH). In contrast to other GAN applications presented in the literature this high-dimensional data has no simple visual representation. We demonstrate how the quality of our generated data can be evaluated using the already trained WGAN model itself as well as a correlation score based on the Fisher transformation.
For benchmarking purposes we introduce a simple discrimination task between ttH and its primary irreducible background. In this setup we train two separate WGANs, one for the signal and one for background events. The performance of a discriminator based on these generated samples is compared to a network trained on the original simulated events.

Intended contribution length

20 minutes

Authors: ERDMANN, Martin (Rheinisch Westfaelische Tech. Hoch. (DE)); RATH, Yannik Alexander (RWTH Aachen University (DE)); RIEGER, Marcel (RWTH Aachen University (DE)); SCHMIDT, David Josef (Rheinisch Westfaelische Tech. Hoch. (DE))

Presenter: SCHMIDT, David Josef (Rheinisch Westfaelische Tech. Hoch. (DE))

Session Classification: Session 3

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

Identifying the relevant dependencies of the neural network response on characteristics of the input space

Tuesday, April 10, 2018 4:55 PM (20 minutes)

The use of neural networks in physics analyses poses new challenges for the estimation of systematic uncertainties. Since the key to a proper estimation of uncertainties is the precise understanding of the algorithm, novel methods for the detailed study of the trained neural network are valuable.
This talk presents an approach to identify those characteristics of the neural network inputs that are most relevant for the response and therefore provides essential information to determine the systematic uncertainties.

Intended contribution length

20 minutes

Authors: WUNSCH, Stefan (KIT - Karlsruhe Institute of Technology (DE)); FRIESE, Raphael Marius (KIT - Karlsruhe Institute of Technology (DE)); WOLF, Roger (KIT - Karlsruhe Institute of Technology (DE)); QUAST, Gunter (KIT - Karlsruhe Institute of Technology (DE))

Presenter: WUNSCH, Stefan (KIT - Karlsruhe Institute of Technology (DE))

Session Classification: Session 4

Contribution ID: **8** Type: **not specified**

Conditional Wasserstein GANs for fast simulation of electromagnetic showers in a CMS HGCAL prototype

The increased instantaneous luminosity at HL-LHC will raise the computing requirements for event reconstruction and analysis for current LHC-based experiments, hence limiting the available resources for the simulation of particles traversing matter. Developments of the performance of state-of-the-art simulation frameworks such as Geant4 are proceeding but are unlikely to fully compensate for this trend.
Generative adversarial neural networks (GANs) have already been shown to provide promising fast simulation models which would speed up the computation time by multiple orders of magnitude. Instead of assuming a simplified calorimeter, we have studied the generation of electron-induced showers in a current prototype of the CMS High Granularity Calorimeter (HGCAL) upgrade project. This prototype calorimeter is made of seven 6-inch and 128-channels hexagonal silicon pad sensors interspersed with absorbers. The setup already includes many of the features required for the challenging HGCAL upgrade.
Our generative model is trained adapting the concept of the Wasserstein distance. Furthermore, conditioning on the binned energy of the incident electrons and on their continuously distributed impact position is integrated implementing two auxiliary regression networks which provide additional terms to the loss function.
In this talk, we present the status of our study. In particular, we show the chosen network architectures, demonstrate the training procedure with the Wasserstein loss and the successful inclusion of the physical constraints. Finally, we provide comparisons of high level observables between simulations obtained with Geant4 and with our generative model.

Intended contribution length

20 minutes

Author: QUAST, Thorben (Rheinisch Westfaelische Tech. Hoch. (DE))

Co-authors: Dr SICKING, Eva (CERN); Prof. ERDMANN, Martin (RWTH Aachen University); Mr GLOMBITZA, Jonas (RWTH Aachen University)

Presenter: QUAST, Thorben (Rheinisch Westfaelische Tech. Hoch. (DE))

Session Classification: Session 3

Contribution ID: **9** Type: **not specified**

Refining particle detector simulations using the Wasserstein distance in adversarial networks

Machine learning models, especially deep neural networks produce appropriate predictions when working on a test set similar to the training set. In physics research machine learning models are usually designed to be used for data application but trained on simulations. Therefore, differences between simulations and data can cause substantial uncertainties in the application.
Here we attempt to reduce these differences by adapting the Wasserstein GAN (WGAN) concept. WGANs have recently been introduced as technological progress in the field of generative models by avoiding mode collapsing, ensuring adequate gradients and providing a meaningful loss metric. We adapt the WGAN concept and apply it within a method to reduce data-simulation mismatches by refining the simulated data. For our investigations we used a calorimeter measuring spatially distributed signal patterns induced by cosmic rays.
We demonstrate that training a deep network with the refined simulated signals leads to a more precise energy reconstruction of events compared to training a network with the simulated signals which differ from data signals. Details can be found in arXiv:1802.03325.

Intended contribution length

20 minutes

Authors: Prof. ERDMANN, Martin; Mr GEIGER, Lukas; Mr GLOMBITZA, Jonas; Mr SCHMIDT, David Josef

Presenter: Mr GLOMBITZA, Jonas

Session Classification: Session 3

Contribution ID: **10** Type: **not specified**

Convolutional Neural Network for Track Seed Filtering at the CMS HLT

Monday, April 9, 2018 11:50 AM (20 minutes)

Collider will constantly bring nominal luminosity increase, with the ultimate goal of reaching a peak luminosity of $5 \cdot 10^{34} cm^{-2} s^{-1}$ for ATLAS and CMS experiments planned for the High Luminosity LHC (HL-LHC) upgrade. This rise in luminosity will directly result in an increased number of simultaneous proton collisions (pileup), up to 200, that will pose new challenges for the CMS detector and, specifically, for track reconstruction in the Silicon Pixel Tracker.

One of the first steps of the track finding workflow is the creation of track seeds, i.e. compatible pairs of hits from different detector layers, that are subsequently fed to to higher level pattern recognition steps. However the set of compatible hit pairs is highly affected by combinatorial background resulting in the next steps of the tracking algorithm to process a significant fraction of fake doublets.

A possible way of reducing this effect is taking into account the shape of the hit pixel cluster to check the compatibility between two hits. To each doublet is attached a collection of two images built with the ADC levels of the pixels forming the hit cluster. Thus the task of fake rejection can be seen as an image classification problem for which Convolutional Neural Networks (CNNs) have been widely proven to provide reliable results.

In this work we present our studies on CNNs applications to the filtering of track pixel seeds. We will show the results obtained for simulated event reconstructed in CMS detector, focussing on the estimation of efficiency and fake rejection performances of our CNN classifier.

Intended contribution length

20 minutes

Author: Mr DI FLORIO, Adriano (Universita e INFN, Bari (IT))

Presenter: Mr DI FLORIO, Adriano (Universita e INFN, Bari (IT))

Session Classification: Session 1

Contribution ID: **11** Type: **not specified**

Direct Learning of Systematics-Aware Summary Statistics

Wednesday, April 11, 2018 3:05 PM (20 minutes)

Complex machine learning tools, such as deep neural networks and gradient boosting algorithms, are increasingly being used to construct powerful discriminative features for High Energy Physics analyses. These methods are typically trained with simulated or auxiliary data samples by optimising some classification or regression surrogate objective. The learned feature representations are then used to build a sample-based statistical model to perform inference (e.g. interval estimation or hypothesis testing) over a set of parameters of interest. However, the effectiveness of the mentioned approach can be reduced by the presence of known uncertainties that cause differences between training and experimental data, included in the statistical model via nuisance parameters. This work presents an end-to-end algorithm, which leverages on existing deep learning technologies but directly aims to produce inference-optimal sample-summary statistics. By including the statistical model and a differentiable approximation of the effect of nuisance parameters in the computational graph, loss functions derived form the observed Fisher information are directly optimised by stochastic gradient descent. This new technique leads to summary statistics that are aware of the known uncertainties and maximise the information that can be inferred about the parameters of interest object of a experimental measurement.

Intended contribution length

20 minutes

Author: DE CASTRO MANZANO, Pablo (Universita e INFN, Padova (IT))

Co-author: DORIGO, Tommaso (Universita e INFN, Padova (IT))

Presenter: DE CASTRO MANZANO, Pablo (Universita e INFN, Padova (IT))

Session Classification: Session 6

Contribution ID: **12** Type: **not specified**

Multivariate Analysis Techniques for charm reconstruction with ALICE

Wednesday, April 11, 2018 3:30 PM (20 minutes)

ALICE is the experiment at the LHC dedicated to heavy-ion collisions. One of the key tools to investigate the strongly-interacting medium (Quark-Gluon Plasma, QGP) formed in heavy-ion collisions is the measurement of open-charm particle production. In particular, charmed baryons, such as ΛC, provide essential information for the understanding of charm thermalisation and hadronisation in the QGP. Data from proton-proton and proton-Pb collisions are needed as a reference for interpreting the results in Pb-Pb collisions, as well as to study charm hadronisation into baryons "in-vacuum". The relatively short lifetime of the ΛC baryon, cτ˜60μm, makes the reconstruction of its decay a challenging task that profits from the excellent performance of ALICE in terms of secondary vertex reconstruction and particle identification. The application of multivariateanalysis (MVA) techniques through Boosted Decision Trees can facilitate the separation of the ΛC signal from the background, and as such be a complementary approach to the more standard technique based on topological and kinematical cuts. In this contribution, the analysis and results of the ΛC -baryon production with MVA in pp collisions at √s = 7 TeV and in p-Pb collisions at √sNN = 5.02 TeV will be shown.

Intended contribution length

20 minutes

Author: ZAMPOLLI, Chiara (CERN)

Presenter: ZAMPOLLI, Chiara (CERN)

Session Classification: Session 6

Contribution ID: **13** Type: **not specified**

Event Categorization using Deep Neural Networks for ttH (H→bb) with the CMS Experiment

Wednesday, April 11, 2018 9:55 AM (20 minutes)

The analysis of top-quark pair associated Higgs boson production enables a direct measurement of the top-Higgs Yukawa coupling. In ttH (H→bb) analyses, multiple event categories are commonly used in order to simultaneously constrain signal and background contributions during a fit to data. A typical approach is to categorize events according to both their jet and b-tag multiplicities. The performance of this procedure is limited by the b-tagging efficiency and diminishes for events with high b-tag multiplicity such as in ttH (H→bb).
Machine learning algorithms provide an alternative method of event categorization. A promising choice for this kind of multi-class classification applications are deep neural networks (DNNs). In this talk, we present a categorization scheme using DNNs that is based on the underlying physics processes of events in the semi-leptonic ttH (H→bb) decay channel. Furthermore, we discuss different methods employed for improving the network's categorization performance.

Intended contribution length

20 minutes

Authors: Prof. ERDMANN, Martin (RWTH Aachen University); RATH, Yannik Alexander (RWTH Aachen University (DE)); RIEGER, Marcel (RWTH Aachen University (DE))

Presenter: RIEGER, Marcel (RWTH Aachen University (DE))

Session Classification: Session 5

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

Generative Models for Fast Cluster Simulations in the TPC for the ALICE Experiment

Tuesday, April 10, 2018 10:00 AM (20 minutes)

Simulating detector response for the Monte Carlo-generated collisions is a key component of every high-energy physics experiment. The methods used currently for this purpose provide high-fidelity results, but their precision comes at a price of high computational cost. In this work, we present a proof-of-concept solution for simulating the responses of detector clusters to particle collisions, using the real-life example of the TPC detector in the ALICE experiment at CERN. An essential component of the proposed solution is a generative model that allows to simulate synthetic data points that bear high similarity to the real data. Leveraging recent advancements in machine learning, we propose to use state-of-the-art generative models, namely Variational Autoencoders (VAE) and Generative Adversarial Networks (GAN), that prove their usefulness and efficiency in the context of computer vision and image processing.
The main advantage offered by those methods is a significant speed up in the execution time, reaching up to the factor of 103 with respect to the Geant 3. Nevertheless, this computational speedup comes at a price of a lower simulation quality and in this work we show quantitative and qualitative proofs of those limitations of generative models. We also propose further steps that will allow to improve the quality of the models and lead to their deployment in production environment of the TPC detector.

Intended contribution length

30 minutes

Authors: DEJA, Kamil Rafal (Warsaw University of Technology (PL)); TRZCINSKI, Tomasz Piotr (Warsaw University of Technology (PL)); GRACZYKOWSKI, Lukasz Kamil (Warsaw University of Technology (PL))

Presenter: DEJA, Kamil Rafal (Warsaw University of Technology (PL))

Session Classification: Session 3

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

Classification of decays involving variable decay chains with convolutional architectures

Wednesday, April 11, 2018 4:30 PM (20 minutes)

Vidyo contribution

We present a technique to perform classification of decays that exhibit decay chains involving a variable number of particles, which include a broad class of B meson decays sensitive to new physics. The utility of such decays as a probe of the Standard Model is dependent upon accurate determination of the decay rate, which is challenged by the combinatorial background arising in high-multiplicity decay modes. In our model, each particle in the decay event is represented as a fixed-dimensional vector of feature attributes, forming an $n \times k$ representation of the event, where n is the number of particles in the event and k is the dimensionality of the feature vector. A convolutional architecture is used to capture dependencies between the embedded particle representations and perform the final classification. The proposed model performs outperforms standard machine learning approaches based on Monte Carlo studies across a range of variable final-state decays with the Belle II detector.

Intended contribution length

20 minutes

Author: TAN, Justin (The University of Melbourne, Belle II)

Presenter: TAN, Justin (The University of Melbourne, Belle II)

Session Classification: Session 6

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

Fisher information metrics for binary classifier evaluation and training

Wednesday, April 11, 2018 9:05 AM (20 minutes)

Different evaluation metrics for binary classifiers are appropriate to different scientific domains and even to different problems within the same domain. This presentation focuses on the optimisation of event selection to minimise statistical errors in HEP parameter estimation, a problem that is best analysed in terms of the maximisation of Fisher information about the measured parameters. After describing a general formalism to derive evaluation metrics based on Fisher information, three more specific metrics are introduced for the measurements of signal cross sections in counting experiments (FIP1) or distribution fits (FIP2) and for the measurements of other parameters from distribution fits (FIP3). The FIP2 metric is particularly interesting because it can be derived from any ROC curve, provided that prevalence is also known. In addition to its relation to measurement errors when used as an evaluation criterion (which makes it more interesting that the ROC AUC), a further advantage of the FIP2 metric is that it can also be directly used for training decision trees (instead of the Shannon entropy or Gini coefficient). Preliminary results based on the Python sklearn framework are presented. The problem of overtraining for these classifiers is also briefly discussed, in terms of the difference of the FIP2 metric on the validation and training set, and of their difference from the theoretical limit. Finally, the expected Fisher information gain from completely random branch splits in the decision tree and its possible relevance in reducing overtraining is analysed.

Intended contribution length

20 minutes

Author: VALASSI, Andrea (CERN)

Presenter: VALASSI, Andrea (CERN)

Session Classification: Session 5

Contribution ID: **17** Type: **not specified**

A Deep Learning tool for fast simulation

Tuesday, April 10, 2018 11:50 AM (20 minutes)

Machine Learning techniques have been used in different applications by the HEP community: in this talk, we discuss the case of detector simulation. The need for simulated events, expected in the future for LHC experiments and their High Luminosity upgrades, is increasing dramatically and requires new fast simulation solutions. We will describe an R&D activity, aimed at providing a configurable tool capable of training a neural network to reproduce the detector response and replace standard Monte Carlo simulation. This represents a generic approach in the sense that such a network could be designed and trained to simulate any kind of detector and, eventually, the whole data processing chain in order to get, directly in one step, the final reconstructed quantities, in just a small fraction of time. We will present the first application of three-dimensional convolutional Generative Adversarial Networks to the simulation of high granularity electromagnetic calorimeters. We will describe detailed validation studies comparing our results to standard Monte Carlo simulation, showing, in particular, the very good agreement we obtain for high level physics quantities and detailed calorimeter response. We will show the computing resources needed to train such networks and the implementation of a distributed adversarial training strategy (based on data parallelism). Finally we will discuss how we plan to generalize our model in order to simulate a whole class of calorimeters, opening the way to a generic machine learning based fast simulation approach.

Intended contribution length

30 minutes

Authors: Dr VALLECORSA, Sofia (Gangneung-Wonju National University (KR)); KHATTAK, Gul Rukh (University of Peshawar (PK)); CARMINATI, Federico (CERN); Dr VLIMANT, Jean-Roch (California Institute of Technology (US))

Presenter: KHATTAK, Gul Rukh (University of Peshawar (PK))

Session Classification: Session 3

Contribution ID: **18** Type: **not specified**

Particle identification at LHCb: new calibration techniques and machine learning classification algorithms

Tuesday, April 10, 2018 5:45 PM (20 minutes)

Particle identification (PID) plays a crucial role in LHCb analyses. Combining information from LHCb subdetectors allows one to distinguish between various species of long-lived charged and neutral particles. PID performance directly affects the sensitivity of most LHCb measurements. Advanced multivariate approaches are used at LHCb to obtain the best PID performance and control systematic uncertainties. This talk highlights recent developments in PID that use innovative machine learning techniques, as well as novel data-driven approaches which ensure that PID performance is well reproduced in simulation.

Intended contribution length

20 minutes

Presenter: LUCIO MARTINEZ, Miriam (Universidade de Santiago de Compostela (ES))

Session Classification: Session 4

Contribution ID: **19** Type: **not specified**

Deep neural network-based multi-class boosted object tagger in ATLAS

A deep neural network-based multi-class boosted object tagger is developed in the context of a search for pair production of heavy vector-like quarks with hadronic final states in ATLAS. The four classes of the tagger are W/Z (V)-boson, Higgs-boson, top-quark and background jets. As the unambiguous identification of the origin of the jet is essential for this search, an identification algorithm using this four-class deep neural network is designed to allow for this. In this analysis jets from boosted objects are reconstructed with a variable cone size by re-clustering calibrated small radius jets. Both lower level information (properties of the constituent small radius jets) and the higher level features of the variable radius reclustered jets are used to train the deep neural network. By using only this information as input to the tagger, the systematic uncertainties of the tagger are obtained by a propagation of the small radius jet uncertainties. The identification algorithm developed for this final state and its performance in Monte Carlo simulation will be presented.

Intended contribution length

20 minutes

Author: AKILLI, Ece (Universite de Geneve (CH))

Presenter: AKILLI, Ece (Universite de Geneve (CH))

Session Classification: Session 5

Contribution ID: **20** Type: **not specified**

Recursive Neural Networks in Quark/Gluon Tagging

Wednesday, April 11, 2018 11:25 AM (20 minutes)

Vidyo contribution

Based on the natural tree-like structure of jet sequential clustering, the recursive neural networks (RecNNs) embed jet clustering history recursively as in natural language processing. We explore the performance of RecNN in quark/gluon discrimination. The results show that RecNNs work better than the baseline BDT by a few percent in gluon rejection at the working point of 50\% quark acceptance. We also experimented on some relevant aspects which might influence the performance of networks. It shows that even only particle flow identification as input feature without any extra information on momentum or angular position is already giving a fairly good result, which indicates that most of the information for q/g discrimination is already included in the tree-structure itself.

Intended contribution length

20 minutes

Author: CHENG, Taoli (University of Chinese Academy of Sciences)

Presenter: CHENG, Taoli (University of Chinese Academy of Sciences)

Session Classification: Session 5

Contribution ID: **21** Type: **not specified**

Machine learning in the LHCb tracking

The LHCb experiment at CERN operates a high precision and robust tracking system to reach its physics goals, including precise measurements of CP-violation phenomena in the heavy flavour quark sector and searches for New Physics beyond the Standard Model. Since Run2, the experiment has put in place a new trigger strategy with a real-time reconstruction, alignment and calibration, imposing strong constraints to the execution of the track reconstruction in terms of timing and throughput. In order to face these constraints, fast machine learning techniques are used in the reconstruction algorithms: they allow to reject fake tracks at an early stage, making the execution faster and additionally providing tracks of better quality. The latest example in this direction is provided by the so-called downstream algorithm, reconstructing tracks from long-lived particles, i.e. particles decaying after the vertex detector. Adopted since 2015 in Run II data taking, its performance, dependent on the purity of the reconstructed track samples, has been improved in 2016 by using two filters, based on a binned Boosted Decision Tree (bBDT) and a neural network.

In this talk, the computational intelligence aspects of the track reconstruction in LHCb will be discussed, with a focus on the adaptation of the employed machine learning techniques to the real-time high level trigger environment.

Intended contribution length

20 minutes

Authors: POLCI, Francesco (Centre National de la Recherche Scientifique (FR)); DZIURDA, Agnieszka (CERN); GRILLO, Lucia (University of Manchester (GB)); DUJANY, Giulio (Universite Pierre et Marie Curie et Universite Denis Diderot ()

Presenters: POLCI, Francesco (Centre National de la Recherche Scientifique (FR)); DZIURDA, Agnieszka (CERN); GRILLO, Lucia (University of Manchester (GB)); DUJANY, Giulio (Universite Pierre et Marie Curie et Universite Denis Diderot ()

Session Classification: Session 1

Contribution ID: **22** Type: **not specified**

Fast calorimeter simulation in LHCb

Tuesday, April 10, 2018 11:25 AM (20 minutes)

Fast calorimeter simulation in LHCb

In HEP experiments CPU resources required by MC simulations are constantly growing and become a very large fraction of the total computing power (greater than 75%). At the same time the pace of performance improvements from technology is slowing down, so the only solution is a more efficient use of resources. Efforts are ongoing in the LHC experiments to provide multiple options for simulating events in a faster way when higher statistics is needed. A key of the success for this strategy is the possibility of enabling fast simulation options in a common framework with minimal action by the final user. In this talk we will describe the solution adopted in Gauss, the LHCb simulation software framework, to selectively exclude particles from being simulated by the Geant4 toolkit and to insert the corresponding hits generated in a faster way. The approach, integrated within the Geant4 toolkit, has been applied to the LHCb calorimeter but it could also be used for other subdetectors. The hits generation can be carried out by any external tool, e.g. by a static library of showers or more complex machine-learning techniques. In LHCb generative models, which are nowadays widely used for computer vision and image processing are being investigated in order to accelerate the generation of showers in the calorimeter. These models are based on maximizing the likelihood between reference samples and those produced by a generator. The two main approaches are Generative Adversarial Networks (GAN), that takes into account an explicit description of the reference, and Variational Autoencoders (VAE), that uses latent variables to describe them. We will present how GAN approach can be applied to the LHCb calorimeter simulation, its advantages and drawbacks.

Intended contribution length

20 minutes

Authors: RATNIKOV, Fedor (Yandex School of Data Analysis (RU)); LHCB COLLABORATION

Presenter: ZAKHAROV, Egor

Session Classification: Session 3

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

Studies to mitigate difference between real data and simulation for jet tagging

Tuesday, April 10, 2018 5:20 PM (20 minutes)

The aim of the studies presented is to improve the performance of jet flavour tagging on real data while still exploiting a simulated dataset for the learning of the main classification task. In the presentation we explore "off the shelf"domain adaptation techniques as well as customised additions to them. The latter improves the calibration of the tagger, potentially leading to smaller systematic uncertainties. The studies are performed with simplified simulations for the case of b-jet tagging. The presentation will include first results as well as discuss pitfalls that we discovered during our research.

Intended contribution length

20 minutes

Authors: STOYE, Markus (CERN); VERZETTI, Mauro (CERN); KIESELER, Jan (CERN); MARTELLI, Arabella (Imperial College (GB)); BUCHMULLER, Oliver (Imperial College (GB))

Presenters: STOYE, Markus (CERN); VERZETTI, Mauro (CERN); KIESELER, Jan (CERN); MARTELLI, Arabella (Imperial College (GB)); BUCHMULLER, Oliver (Imperial College (GB))

Session Classification: Session 4

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

DeepJet: a deep-learned multiclass jet-tagger for slim and fat jets

Monday, April 9, 2018 11:00 AM (20 minutes)

We present a customized neural network architecture for both, slim and fat jet tagging. It is based on the idea to keep the concept of physics objects, like particle flow particles, as a core element of the network architecture. The deep learning algorithm works for most of the common jet classes, i.e. b, c, usd and gluon jets for slim jets and W, Z, H, QCD and top classes for fat jets. The developed architecture promising gains in performance as shown in simulation of the CMS collaboration. Currently the tagger is under test in real data in the CMS experiment.

Intended contribution length

20 minutes

Authors: VERZETTI, Mauro (CERN); KIESELER, Jan (CERN); STOYE, Markus (CERN); QU, Huilin (Univ. of California Santa Barbara (US)); GOUSKOS, Loukas (Univ. of California Santa Barbara (US))

Presenters: VERZETTI, Mauro (CERN); KIESELER, Jan (CERN); STOYE, Markus (CERN); QU, Huilin (Univ. of California Santa Barbara (US)); GOUSKOS, Loukas (Univ. of California Santa Barbara (US))

Session Classification: Session 1

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

DeepJet: a portable ML environment for HEP

Wednesday, April 11, 2018 9:30 AM (20 minutes)

In this presentation we will detail the evolution of the DeepJet python environment. Initially envisaged to support the development of the namesake jet flavour tagger in CMS, DeepJet has grown to encompass multiple purposes within the collaboration. The presentation will describe the major features the environment sports: simple out-of-memory training with a multi-treaded approach to maximally exploit the hardware acceleration, simple and streamlined I/O to help bookkeeping of the developments, and finally docker image distribution, to simplify the deployment of the whole ecosystem on multiple datacenters. The talk will also cover future development, mainly aimed at improving user experience.

Intended contribution length

20 minutes

Authors: MEHTA, Swapneel Sundeep (Dwarkadas J Sanghvi College of Engineering (IN)); Mr MEHTA, swapneel (IT/DB Group); VERZETTI, Mauro (CERN); KIESELER, Jan (CERN); STOYE, Markus (CERN)

Presenters: MEHTA, Swapneel Sundeep (Dwarkadas J Sanghvi College of Engineering (IN)); Mr MEHTA, swapneel (IT/DB Group)

Session Classification: Session 5

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

Adversarial Tuning of Perturbative Parameters in Non-Differentiable Physics Simulators

Tuesday, April 10, 2018 11:00 AM (20 minutes)

In this contribution, we present a method for tuning perturbative parameters in Monte Carlo simulation using a classifier loss in high dimensions. We use an LSTM trained on the radiation pattern inside jets to learn the parameters of the final state shower in the Pythia Monte Carlo generator. This represents a step forward compared to unidimensional distributional template-matching methods.

Intended contribution length

20 minutes

Authors: DE OLIVEIRA, Luke Percival; PAGANINI, Michela (Yale University (US)); NACHMAN, Ben (University of California Berkeley (US)); MRENNA, Steve (Fermi National Accelerator Lab. (US)); SHIMMIN, Chase Owen (Yale University (US)); TIPTON, Paul Louis (Physics Department - Yale University)

Presenter: PAGANINI, Michela (Yale University (US))

Session Classification: Session 3

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

Machine learning in jet physics

Wednesday, April 11, 2018 11:00 AM (20 minutes)

High energy collider experiments produce several petabytes of data every year. Given the magnitude and complexity of the raw data, machine learning algorithms provide the best available platform to transform and analyse these data to obtain valuable insights to understand Standard Model and Beyond Standard Model theories. These collider experiments produce both quark and gluon initiated hadronic jets as the core components. Deep learning techniques enable us to classify quark/gluon jets through image recognition and help us to differentiate signals and backgrounds in Beyond Standard Model searches at LHC. We are currently working on quark/gluon jet classification and progressing in our studies to find the bias between event generators using domain adversarial neural networks (DANN). We also plan to investigate top tagging, weak supervision on mixed samples in high energy physics, utilizing transfer learning from simulated data to real experimental data.

Intended contribution length

20 minutes

Author: NARAYANA VARMA, Sreedevi (King's College London)

Presenter: NARAYANA VARMA, Sreedevi (King's College London)

Session Classification: Session 5

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

HL-LHC tracking challenge

Monday, April 9, 2018 11:25 AM (20 minutes)

At HL-LHC, the seven-fold increase of multiplicity wrt 2018 conditions poses a severe challenge to ATLAS and CMS tracking experiments. Both experiment are revamping their tracking detector, and are optimizing their software. But are there not new algorithms developed outside HEP which could be invoked: for example MCTS, LSTM, clustering, CNN, geometric deep learning and more? We organize on the Kaggle platform a data science competition to stimulate both the ML and HEP communities to renew core tracking algorithms in preparation of the next generation of particle detectors at the LHC.

In a nutshell : one event has 100.000 3D points ; how to associate the points onto 10.000 unknown approximately helicoidal trajectories ? avoiding combinatorial explosion ? you have a few seconds. But we do give you 100.000 events to train on.

We ran ttbar+200 minimum bias event into ACTS a simplified (yet accurate) simulation of a generic LHC silicon detectors, and wrote out the reconstructed hits, with matching truth. We devised an accuracy metric which capture with one number the quality of an algorithm (high efficiency/low fake rate).
The challenge will run in two phases: the first on accuracy (no stringent limit on CPU time), starting in April 2018, and the second (starting in the summer 2018) on the throughput, for a similar accuracy.

Intended contribution length

20 minutes

Authors: VLIMANT, Jean-Roch (California Institute of Technology (US)); ROUSSEAU, David (LAL-Orsay, FR)

Presenter: VLIMANT, Jean-Roch (California Institute of Technology (US))

Session Classification: Session 1

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

DarkMachines

Monday, April 9, 2018 9:20 AM (10 minutes)

http://www.darkmachines.org/

Intended contribution length

Author: DORIGO, Tommaso (Universita e INFN, Padova (IT))

Presenter: DORIGO, Tommaso (Universita e INFN, Padova (IT))

Session Classification: Session 1

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

Joint Wasserstein GAN contribution

Tuesday, April 10, 2018 9:05 AM (45 minutes)

This is a merger of three individual contributions:
- https://indico.cern.ch/event/668017/contributions/2947026/
- https://indico.cern.ch/event/668017/contributions/2947027/
- https://indico.cern.ch/event/668017/contributions/2947028/

Intended contribution length

Authors: ERDMANN, Martin (Rheinisch Westfaelische Tech. Hoch. (DE)); RATH, Yannik Alexander (RWTH Aachen University (DE)); RIEGER, Marcel (RWTH Aachen University (DE)); SCHMIDT, David Josef (Rheinisch Westfaelische Tech. Hoch. (DE)); QUAST, Thorben (Rheinisch Westfaelische Tech. Hoch. (DE)); SICKING, Eva (CERN); GLOMBITZA, Jonas (Rheinisch-Westfaelische Tech. Hoch. (DE)); Mr GEIGER, Lukas

Presenters: SCHMIDT, David Josef (Rheinisch Westfaelische Tech. Hoch. (DE)); QUAST, Thorben (Rheinisch Westfaelische Tech. Hoch. (DE)); GLOMBITZA, Jonas (Rheinisch-Westfaelische Tech. Hoch. (DE))

Session Classification: Session 3

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

Welcome

Monday, April 9, 2018 9:00 AM (20 minutes)

Presenters: MONETA, Lorenzo (CERN); STOYE, Markus (CERN); SEYFERT, Paul (CERN); HAAKE, Rudiger (CERN); SCHRAMM, Steven Randolph (Universite de Geneve (CH))

Session Classification: Session 1

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

Daily announcements

Tuesday, April 10, 2018 9:00 AM (5 minutes)

Presenters: MONETA, Lorenzo (CERN); STOYE, Markus (CERN); SEYFERT, Paul (CERN); HAAKE, Rudiger (CERN); SCHRAMM, Steven Randolph (Universite de Geneve (CH))

Session Classification: Session 3

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

Tutorial: Keras/TensorFlow

Tuesday, April 10, 2018 2:00 PM (1 hour)

Intended contribution length

Presenter: WUNSCH, Stefan (KIT - Karlsruhe Institute of Technology (DE))

Session Classification: Session 4

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

Interpreting Deep Neural Networks and their Predictions

Tuesday, April 10, 2018 3:00 PM (1 hour)

Invited talk, http://iphome.hhi.de/samek/

Presenter: SAMEK, Wojciech (Fraunhofer HHI)

Session Classification: Session 4

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

Daily announcements

Wednesday, April 11, 2018 9:00 AM (5 minutes)

Presenters: MONETA, Lorenzo (CERN); STOYE, Markus (CERN); SEYFERT, Paul (CERN); HAAKE, Rudiger (CERN); SCHRAMM, Steven Randolph (Universite de Geneve (CH))

Session Classification: Session 5

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

Close-out and challenge results

Wednesday, April 11, 2018 5:20 PM (20 minutes)

Intended contribution length

Presenters: MONETA, Lorenzo (CERN); STOYE, Markus (CERN); SEYFERT, Paul (CERN); HAAKE, Rudiger (CERN); SCHRAMM, Steven Randolph (Universite de Geneve (CH))

Session Classification: Session 6

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

Tutorial: TMVA

Monday, April 9, 2018 2:00 PM (1 hour)

Intended contribution length

Presenter: MONETA, Lorenzo (CERN)

Session Classification: Session 2

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

Introduction to the industry session

Monday, April 9, 2018 3:00 PM (10 minutes)

Intended contribution length

Presenters: MONETA, Lorenzo (CERN); STOYE, Markus (CERN); SEYFERT, Paul (CERN); HAAKE, Rudiger (CERN); SCHRAMM, Steven Randolph (Universite de Geneve (CH))

Session Classification: Session 2

Contribution ID: **40** Type: **not specified**

Multilevel Optimization for Generative Models, Games and Robotics

Monday, April 9, 2018 3:10 PM (30 minutes)

Intended contribution length

Presenter: PFAU, David (Google DeepMind)

Session Classification: Session 2

Contribution ID: **41** Type: **not specified**

Machine Learning For Enterprises: Beyond Open Source

Monday, April 9, 2018 3:50 PM (30 minutes)

invited talk from IBM analytics

Intended contribution length

Presenter: PUGET, Jean-Francois (IBM Analytics)

Session Classification: Session 2

Contribution ID: **42** Type: **not specified**

Surrogate Models for Fun and Profit

Monday, April 9, 2018 5:00 PM (30 minutes)

Intended contribution length

Presenter: USTYUZHANIN, Andrey (Yandex School of Data Analysis (RU))

Session Classification: Session 2

Contribution ID: **43** Type: **not specified**

Invited industry talk #4

Session Classification: Session 2

Contribution ID: **44** Type: **not specified**

Industry panel

Monday, April 9, 2018 5:40 PM (30 minutes)

Intended contribution length

Presenters: USTYUZHANIN, Andrey (Yandex School of Data Analysis (RU)); PFAU, David (Google DeepMind); PUGET, Jean-Francois (IBM Analytics)

Session Classification: Session 2

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

Overview of ML in HEP

Monday, April 9, 2018 9:35 AM (40 minutes)

Presenter: DE OLIVEIRA, Luke Percival

Session Classification: Session 1

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

TBA

Session Classification: Session 6