Contribution ID: **199** Type: **presentation**

trackML : the Kaggle HEP tracking challenge

Thursday 12 July 2018 15:00 (15 minutes)

The High-Luminosity LHC will see pileup level reaching 200, which will greatly increase the complexity the tracking component of the event reconstruction.
To reach out to Computer Science specialists, a Tracking Machine Learning challenge (trackML) is being set up on Kaggle for the first 2018 semester by a team of ATLAS, CMS and LHCb physicists tracking experts and Computer Scientists, building on the experience of the successful Higgs Machine Learning challenge in 2014. A dataset consisting of an accurate simulation of a LHC experiment tracker has been created, listing for each event the measured 3D points, and the list of 3D points associated to a true track. The data set is large to allow the training of data hungry Machine Learning methods : the orders of magnitude are : 100.000 events, 1 billion tracks, 100 GigaByte. Typical CPU time spent by traditional track reconstruction algorithms is 100s per event. No limit on the training resources will be imposed. The participants to the challenge should find the tracks in an additional test dataset, which means building the list of 3D points belonging to each track (deriving the track parameters is not the topic of the challenge). The emphasis is to expose innovative approaches, rather than hyper-optimising known approaches. The challenge will be run in two phases:
1. During the Accuracy phase (March to June 2018), a metric reflecting the accuracy of the model at finding the proper point association that matters to most physics analysis will allow to elect the programs that could be good candidate at replacing the existing algorithms. The metric is based on the overall fraction of points associated to a good track (a good track being a track where more than 50% of the points come from the same ground truth tracks) has been shown to be well behaved and robust
2. The Throughput phase (July to October 2018) will focus on optimising the inference speed on one CPU core, starting from the collection of algorithms exposed in the first phase. The training speed will remain unconstrained. We aim with this second phase at finding new implementation of algorithms for faster execution, at the cost of minimal accuracy loss.

This talk will summarize the findings of the Accuracy phase, where multiple algorithms will have competed, and new approaches from Machine Learning that have been exposed. The various merits of the different algorithms will be discussed not only the accuracy, but also the detailed performance aspects, efficiency, fake rates as a function of the track parameters and the track density. The talk will also help advertise the second phase of the challenge.

Primary authors: ROUSSEAU, David (LAL-Orsay, FR); YILMAZ, Yetkin (Laboratoire Leprince-Ringuet, France); Dr VLIMANT, Jean-Roch (California Institute of Technology (US)); GUYON, Isabelle; INNOCENTE, Vincenzo (CERN); SALZBURGER, Andreas (CERN); AMROUCHE, Sabrina (RSA - Universite de Geneve (CH)); GOLLING, Tobias (Universite de Geneve (CH)); KIEHN, Moritz (Universite de Geneve (CH)); CALAFIURA, Paolo (Lawrence Berkeley National Lab. (US)); FARRELL, Steven Andrew (Lawrence Berkeley National Lab. (US)); GRAY, Heather (LBNL); GLIGOROV, Vladimir (Centre National de la Recherche Scientifique (FR)); GERMAIN, Cecile (Universite Paris Sud); HUSHCHYN, Mikhail (Yandex School of Data Analysis (RU)); USTYUZHANIN, Andrey (Yandex School of Data Analysis (RU))

Presenter: KIEHN, Moritz (Universite de Geneve (CH))

Session Classification: T6 - Machine learning and physics analysis

Track Classification: Track 6 –Machine learning and physics analysis