TrackML: a Tracking Machine Learning Challenge

Jean-Roch Vlimant (Caltech), Vincenzo Innocente, Andreas Salzburger (CERN), Isabelle Guyon (ChaLearn), Sabrina Amrouche, Tobias Golling, Moritz Kiehn (Geneva University), David Rousseau, Yetkin Yilmaz (LAL-Orsay), Paolo Calafìra, Steven Farrell, Heather Gray (LBNL), Vladimir Vava Gligorov (LPNHE-Paris), Laurent Basara, Cécile Germain, Victor Estrade (LRI-Orsay), Edward Moyse (University of Massachussets), Mikhail Hushchyn, Andrey Ustyuzhanin (Yandex, HSE)
Outline

- Forewords on Tracking and Machine learning
  - Accuracy Phase
  - Throughput Phase
- Outlook
Tracking in a Nutshell

- Particle trajectory bented in a solenoidal magnetic field
- Curvature is a proxy to momentum
- Particle ionize silicon pixel and strip throughout several concentric layers
- **Thousands of sparse hits**
- Lots of hit pollution from low momentum, secondary particles

**Explosion in hit combinatorics** in both seeding and stepping pattern recognition

**Highly computing consuming task** in extracting physics content from LHC data
Tracking with Machine Learning

\[ E = -\frac{1}{2} \left( \sum_{i,j} w_{ij} s_i s_j - \sum_i \varepsilon_i s_i \right). \]

Fig. 4. Tracks in the ALEPH TPC reconstructed with a Hopfield net [13].

https://tinyurl.com/y9swquw6

https://tinyurl.com/yb3v93y9

https://tinyurl.com/y87saehf

https://arxiv.org/abs/1604.01444
Challenge Datasets

- Accurate simulation engine (ACTS) to produce realistic dataset
  - One file with list 3D points
  - Ground truth: one file with point to particle association
  - Ground truth auxiliary: true particle parameter (origin, direction, curvature)
  - Typical events with ~200 parasitic collisions (~10,000 tracks/event, ~100k hits/event)

- The goal of the challenge is to **assemble hits into tracks**
- Large training sample 100k events, 10 billion tracks ~100 GB
Accuracy Phase

April 30 – August 13, 2018

https://www.kaggle.com/c/trackml-particle-identification
Scoring : Accuracy Phase

- At least 50% hits from the same ground truth particle
- At least 50% hits of the ground truth particle in the track
- Sum of weights ($w_{\text{order}} \times w_{pT}$) of truth matched hits

- Score normalized to sum of weights : ideal score is 1
- 100 events used for scoring : precision ~0.1%
Final Leaderboard

https://www.kaggle.com/c/trackml-particle-identification/leaderboard
Physics Performance

Highest score correlates well with the tracking efficiency
Insight on Algorithms 1/2

• First: *Top Quarks*
  ➢ Johan Sokrates is an industrial Mathematics master student
  ➢ Pair seeding, triplet extension, **trajectory following**, track cleaning, all with **machine learning** for quality selection

• Second: *Outrunner*
  ➢ Pei-Lien Chou is a software engineer in image-based deep learning in Tai̇wan
  ➢ **Machine learning** to predict the adjacency matrix

• Third: *Sergey Gorbunov*
  ➢ Sergey Gorbunov is a physicist, expert in tracking
  ➢ Iterative steps, triplet seeding, **trajectory following**
Insight on Algorithms 2/2

- **Jury Innovative prize**
  - Yuval Reina is an electronic engineer and Trian Xylouris is an entrepreneur
  - Marginalized Hough transform with **machine learning classifier**

- **Jury Clustering prize**
  - Jean-François Puget CPMP is a software engineer at IBM. He is both competition and discussion Kaggle grandmaster
  - **DBSCAN clustering** with iterative Hough transform

- **Jury Deep Learning prize**
  - Nicole and Liam Finnie are software engineers
  - DBSCAN seeding, **trajectory following with LSTM**

- **Organization pick**
  - Diogo R. Ferreira is a professor/researcher, focusing on data science and nuclear fusion
  - **Pattern matching**
Throughput Phase

Oct 12 – March 12, 2019

Modifications to the dataset

Same detector geometry with half thinner modules
Beamspot corrected from 5.5 mm to 5.5 cm
Loopers are not simulated
Bug on electron multiple scattering fixed

https://competitions.codalab.org/competitions/20112
Scoring: Throughput Phase

Zero score if time > 600s and accuracy < 0.5

$$\sqrt{\log \left(1 + \frac{600}{\text{time}}\right) \times (\text{accuracy} - 0.5)^2}$$
## Leaderboard

<table>
<thead>
<tr>
<th>#</th>
<th>User</th>
<th>Entries</th>
<th>Date of Last Entry</th>
<th>score ▲</th>
<th>accuracy_mean ▲</th>
<th>accuracy_std ▲</th>
<th>computation time (sec) ▲</th>
<th>computation speed (sec/event) ▲</th>
<th>Duration ▲</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>sgorbuno</td>
<td>9</td>
<td>03/12/19</td>
<td>1.1727 (1)</td>
<td>0.944 (2)</td>
<td>0.00 (14)</td>
<td>28.06 (1)</td>
<td>0.56 (1)</td>
<td>64.00 (1)</td>
</tr>
<tr>
<td>2</td>
<td>fastrack</td>
<td>53</td>
<td>03/12/19</td>
<td>1.1145 (2)</td>
<td>0.944 (1)</td>
<td>0.00 (15)</td>
<td>55.51 (16)</td>
<td>1.11 (16)</td>
<td>91.00 (6)</td>
</tr>
<tr>
<td>3</td>
<td>cloudkitchen</td>
<td>73</td>
<td>03/12/19</td>
<td>0.9007 (3)</td>
<td>0.928 (3)</td>
<td>0.00 (13)</td>
<td>364.00 (18)</td>
<td>7.28 (18)</td>
<td>407.00 (8)</td>
</tr>
<tr>
<td>4</td>
<td>cubus</td>
<td>8</td>
<td>09/13/18</td>
<td>0.7719 (4)</td>
<td>0.895 (4)</td>
<td>0.01 (9)</td>
<td>675.35 (19)</td>
<td>13.51 (19)</td>
<td>724.00 (9)</td>
</tr>
<tr>
<td>5</td>
<td>Taka</td>
<td>11</td>
<td>01/13/19</td>
<td>0.5930 (5)</td>
<td>0.875 (5)</td>
<td>0.01 (12)</td>
<td>2668.50 (23)</td>
<td>53.37 (23)</td>
<td>2758.00 (13)</td>
</tr>
<tr>
<td>6</td>
<td>Vicennial</td>
<td>27</td>
<td>02/24/19</td>
<td>0.5634 (6)</td>
<td>0.815 (6)</td>
<td>0.01 (10)</td>
<td>1270.73 (20)</td>
<td>25.41 (20)</td>
<td>1339.00 (10)</td>
</tr>
<tr>
<td>7</td>
<td>Sharad</td>
<td>57</td>
<td>03/10/19</td>
<td>0.2918 (7)</td>
<td>0.674 (7)</td>
<td>0.02 (4)</td>
<td>1902.20 (22)</td>
<td>38.04 (22)</td>
<td>1986.00 (12)</td>
</tr>
<tr>
<td>8</td>
<td>WeizmannAl</td>
<td>5</td>
<td>03/12/19</td>
<td>0.0000 (8)</td>
<td>0.133 (11)</td>
<td>0.01 (11)</td>
<td>88.08 (17)</td>
<td>1.76 (17)</td>
<td>124.00 (7)</td>
</tr>
<tr>
<td>9</td>
<td>harshakoundinya</td>
<td>2</td>
<td>03/12/19</td>
<td>0.0000 (8)</td>
<td>0.085 (13)</td>
<td>0.01 (6)</td>
<td>49.22 (8)</td>
<td>0.98 (8)</td>
<td>86.00 (3)</td>
</tr>
<tr>
<td>10</td>
<td>iWit</td>
<td>6</td>
<td>03/10/19</td>
<td>0.0000 (8)</td>
<td>0.082 (15)</td>
<td>0.01 (8)</td>
<td>48.23 (3)</td>
<td>0.96 (3)</td>
<td>85.00 (2)</td>
</tr>
<tr>
<td>11</td>
<td>yangguo</td>
<td>1</td>
<td>03/01/19</td>
<td>0.0000 (8)</td>
<td>0.082 (15)</td>
<td>0.01 (8)</td>
<td>48.63 (4)</td>
<td>0.97 (4)</td>
<td>86.00 (3)</td>
</tr>
</tbody>
</table>
Incidentally, best solutions are also best accuracy and best timing. Software will be submitted and analyzed in the coming weeks.
Physics Performance

Highest score correlates well with the tracking efficiency
Insights on Algorithms

- **First: Sgorbuno**
  - Sergey Gorbunov is a physicist, expert in tracking
  - 3rd position at the accuracy phase
  - Iterative steps, triplet seeding, **trajectory following**

- **Second: fastrack**
  - Dmitry Emeliyanov is a physicist
  - Graph representation of neighbors, cellular automaton, track following

- **Third: cloudkitchen**
  - Marcel Kunze is a former physicist,
  - Direct acyclic graph of voxels, pair and triplet classification + Top Quark solution of accuracy phase (**trajectory following**, track cleaning, all with **machine learning** for quality selection)
Summary & Outlooks

- Challenges are hard to prepare
  - Finding the right metric is key.
- Timing the code is hard
  - Reason to use codalab. Kaggle will step up
- Challenges are fun to run and to participate to
  - The TrackML Challenge concluded at the July1-2 Grand Finale at CERN [https://indico.cern.ch/event/813759/](https://indico.cern.ch/event/813759/)
- Understanding the solutions will take time
  - Several proposed algorithms are being implemented in ACTS for benchmarking
- The challenge datasets were used in several papers
  - A consolidated tracking challenge dataset will be made public on CERN open data portal.
Extra Material
A Tracking with Machine Learning Challenge
Previous Challenges

- 2000 teams. Largest competition at the time
- Winners went to DeepMind and OpenAI
  https://www.kaggle.com/c/higgs-boson

- 700 teams.
- Experienced data exploitation
- Some methods learned and re-applied later
  https://www.kaggle.com/c/flavours-of-physics

The organizing team has participated in the organization of both events
The Jury

**Markus Elsing**, CERN senior staff, group leader of the ATLAS computing and software group.

**Frank Gaede** senior physicist at DESY (Germany) is software coordinator for ILD.

**Alison Lowndes** is responsible for NVIDIA's Artificial Intelligence Developer Relations in the Europe, Middle East & Africa region.

**Maurizio Pierini** is a CERN physicist lead of the machine learning for Particle Physics ERC grant.

**Danilo Rezende** is Staff Research Scientist at Google DeepMind.

**Marc Schoenauer** is senior scientist at INRIA-Saclay

**Svyatoslav Voloshynovskyy** is associate professor at University of Geneva.

Reviewed the documentation made public by contestants of the challenge and decided on the level of innovation.

https://sites.google.com/site/trackmlparticle/international-advisory-committee