Discussion

Amir Farbin

Where is ML needed?

- Traditionally ML Techniques in HEP
    - Applied to Particle/Object Identification
    - Signal/Background separation
    - Here, ML maximizes reach of existing data/detector… equivalent to additional integral luminosity.
        - There is lots of interesting work here… and potential for big impact.
- Now we hope ML can help address looming computing problems
    - Reconstruction
        - LArTPC- Algorithmic Approach very difficult
        - HL-LHC Tracking- Pattern Recognition blows up due to combinatorics
    - Simulation
        - LHC Calorimetry- Large Fraction of ATLAS CPU goes into shower simulation.

Working Groups

- There is a variety of areas where new techniques can potentially improve performance or speed of an existing algorithm, or even address difficult unsolved problem.
    - Seems to me that Deep Learning is the predominant approach…
- Because of rules of our Experiments, the majority of this work has been done with self made datasets.
    - Reminiscent of search techniques papers…
    - Dataset often made public.
    - Mainly proof of principle or feasibility studies…
- Many of the most demanding problems are reconstruction/simulation related, requiring large full Geant4-based simulations.
    - These can be rather cumbersome to create, store, and manipulate.
    - Often multiple problems can be explored with a single dataset.
    - Perhaps also provide real / test-beam data…
- In most cases, we are far away from solutions that are sophisticated/realistic enough for real data.
    - Move from proof of principle towards understanding
    - From toys to realistic detectors/environment.

Working Groups

- We want to go from feasibilities studies on simplified or toy samples to more realistic (simulation and environment) and systematically understand and compare architectures.
    - Physicists have to *feel like* they understand what is happening…
    - We need to probe the “blackbox”…
- We would like to facilitate/accelerate this work by creating Working Groups around public datasets and associated problems.
    - Engage colleagues in other experiments.
    - Collaborate with Data Science experts.
    - Gradually develop solutions that we can take back to our experiments.
- Goal would be to generate technique papers documenting the work.
    - Journals…
    - Conference submissions? (e.g. NIPS)?
- Umbrella organization? e.g. this workshop series?

Working Groups

- There are analogies to the Snowmass exercises.
    - Benchmark public datasets.
    - Working groups.
    - Papers…
- How could we structure such groups?
    - By dataset? Calo, LArTPC, Tracking, Jets, …
    - By problem? Image Classification, Energy Regression, Generative Model-based Simulation, …
    - Area? Calorimetry, Tracking, LHC experiments, neutrino experiments, LArTPC detectors, …
    - By paper?
- What are the specific Working Groups?
- Do we select “conveners”? Other roles? Dataset team? Software team? “Champions”? Liaisons to Experiments? …
- Meetings: how often? Different types for each WG?
    - DS@HEP Specific Discussion Forums?
- Is there an umbrella organization? Associated with this workshop series?

Attribution Policies

- Datasets should be cited.
- Do we formalize rules for the author list? Like a collaboration?
  - Different levels of contribution?
- Can this activity co-exist with parallel work in the experiments?

Datasets

- One dataset, or classes of datasets? E.g. LCD vs Calorimetry…
- Challenge? Private test dataset, solution submission, leader tables?
- Documentation:
    - Dataset papers? Yes… Journal?
    - Tutorials?
    - Problem formulation?
        - HiggsML is a good example.
        - Background for non-physicists.
        - Benchmarks? Algorithm Performance and Timing?
            - Baseline? Target performance?

Storage/Distribution

- Some of these datasets are large.
  - Likely will have multiple versions. Versioning scheme?
    - Data Challenges?
    - Break into blocks?
- Very large datasets might require more reliable transfer tools?
- Store at labs? EOS/dCache.
- Dataset repositories?
- Clouds?

Software

- Tutorial and simple examples for beginners?
- Simulation software/instructions?
- Data IO? Many of these will require specialized tools to efficiently read.
- Baseline? e.g. tracking in ACTS?
- Common code/framework?
- Current best result? Pretrained models?
- Tools to be brought back to experiments?

Collaborative Technologies

- Webpage?
- Slack channel? Trello board?
- Conferencing software?

Expanding the Community

- We need to advertise…
    - Conferences/workshops (e.g. DPF)… inside/outside HEP.
        - Advertise datasets/problems/tutorials
        - Target results
- People in ML community are looking for datasets and interesting problems.
- Outreach… e.g. LArTPC HandScan.
- Challenges, e.g. Kaggle?

Resources

- For people to contribute meaningfully to some of these projects they will need access to GPUs.
    - For some development, just GPU/person is sufficient.
    - But for the studies we want large resources for hyper parameter scans.
- How do we address this?
    - “Private” Clusters
    - Labs
    - HPCs
    - Cloud- would allow broader access…
        - Who pays?
        - Data should be centrally stored… expensive.
- Short-term vs long-term plans…

Future Workshops

- Alternate between General and Expert meetings… (and Europe/US?)
  - General: Nov 2015 CERN
  - Experts: July 2016 Simons Foundation (NYC)
  - General: May 2017 FNAL
  - Experts: ??? Europe.
- If working groups workout, our meetings may start looking more like collaboration meetings.
- Is this a good model?