AI TASK FORCE

Gagik Gavalian, CLAS Collaboration Meeting
July 21, 2020

TEAM:

Gagik Gavalian (PI)
D. Lawrence (PI)
V. Ziegler (core)
T. Britton (core)
N. Baltzell (core)
O. Hansen (core)
C. Fanelli (core)
M. Ungaro (external)
D. Heddle (external)
L. Elouadrhiri (external)

Charge

1. Identify areas in CLAS12 software where AI-supported solution can be implemented
2. Assess existing AI technologies finding the most suited for the different projects
3. Quantify the expected improvement in the area of applications
4. Define a work plan to test the proposed solutions with a time chart and milestones
5. Define the requirement of an education plan to provide basic concepts to the whole Hall-B staff
6. Estimate costs and identify resources needed for the different projects
7. Evaluate synergies with other projects at the lab providing a list of shared resources and common goals

Introduction

The Areas Where CLAS12 can benefit from AI ?

OFFLINE

- **Tracking Applications:** AI can be used to improve track candidate finding . High luminosity track finding efficiency improvements.
- **Clustering:** Electromagnetic Calorimeter clustering. Cluster splitting for two photon separation.
- **Particle Identification:** Improve PID with AI using detector responses.
- **RICH:** Using AI for ring reconstruction

ONLINE

- **Data Reduction:** Reduce data online by choosing events containing electron tracks (Level-3 trigger)
- **Detector Monitoring:** Monitor detector for problems and problems, identify run time problems
- **Calibration:** Data calibration in real time. Reduce time spent on pass-0 cooking

---

SIMULATION

- **Simulation:** Replace slow parts of GEANT simulation with AI (GANs), speed up calorimeter shower generation.
- **Fast Monte-Carlo:** AI can be used to process input event and produce an output event with acceptance and resolutions already applied.

Offline Projects

- **Track Classification:**
  - In CLAS12 many combinations of tracks are considered based on the segments in the drift chamber. Combination classification significantly improves processing speed (~x6) and efficiency is 99.5%.
  - Hall A's SBS program relies on GEM trackers that will see high occupancies (~20-30%). The tracking regions are essentially field free. Identifying clusters of GEM strips as well as finding (straight) tracks in the tracker systems would undoubtedly benefit from AI pattern recognition techniques.
  - Hall A's SoLID program also plans to use GEM trackers. These will generally see somewhat lower occupancies than SBS, but tracks will not be straight, but rather helical in the field of the solenoidal magnet. AI techniques will be beneficial here, too.
- **Track Segment Predictions:**
  - Using LSTM (Long-Short Term Memory) Networks it is possible to predict where segments in region 1 should be based on segments from regions 2&3. This will help to do clustering in high luminosity runs.
- **Track Parameters Predictions:**
  - Based on segments in the DC our Neural Network was able to give track momentum with accuracy of ~3%.
  - This will impact number of iterations needed for the Kalman-Filter to converge. Improvements to speed are estimated to factor of 3.
  - Further improvements to the network can be done to infer also angles of the track and other initial parameters.
- **Calorimetry in the Offline:**
  - Clustering for calorimeters, splitting clusters (hall-D)
- **Particle Identification:**
  - Particle classification based on responses from detectors. Both Hall-B and Hall-D need it.

Online Projects

- **Level 3 Trigger:**
  - Some tests were performed to identify events where electron was reconstructed in the calorimeter, based on RAW hits patterns in DC and RAW hits pattern in Electromagnetic Calorimeter.
  - Initial tests provide inference rate of 85 KHz (on CPU)
  - This method can be extended to be used in Streaming Readout, with unsupervised learning.
  - Some of the SoLID configurations will probably require a 3rd level trigger where AI might help classify signal vs. noise.
- **Detector Monitoring:**
  - Neural Networks can be used to monitor detector health during running, and to classify runs based on histograms processed from each file.
  - Successful program was developed for Hall-D (Hydra) for online detector monitoring, we can collaborate with GlueX to implement it in Hall-B also.
  - This will also be important for streaming readout, where it will be useful to find faults before reconstructing the events.
- **Online Data Calibration:**
  - Neural Networks can be used also to calibrate (maybe event gain match) detector components while experiment is running.
  - Significantly will reduce time spend on cooking and calibrating detector component in the offline.

Simulation Projects

- **Simulation Speed up through Detector Response modeling:**
  - Some detector responses generation (such as Electromagnetic Calorimeter shower generation) take a very long time in GEANT4. They take about ~70% of simulation time.
  - There are already successful projects that use GANs to simulate calorimetry response that promise speed up up to ~1000x.
  - These methods can be used to in Hall-B, GlueX, Hall-A/C to provide much faster detector simulations.
  - Both SBS and SoLID run fairly compute-intensive simulations and digitization. Both experiments rely heavily on calorimeters (both for electrons and for hadrons), but neither experiment currently runs full calorimeter simulations due to their prohibitive CPU requirements. AI-assisted calorimeter response simulations, even if less accurate than a full simulation, would almost certainly be an improvement here.
- **Fast Monte-Carlo:**
  - Fast Monte Carlo is used to assess detector acceptances and rates for upcoming experiments.
  - Parametrization of detector sensitive volume is used, along with parametrization of detector resolutions.
  - AI can be used to process input event and produce an output event with acceptance and resolutions already applied.

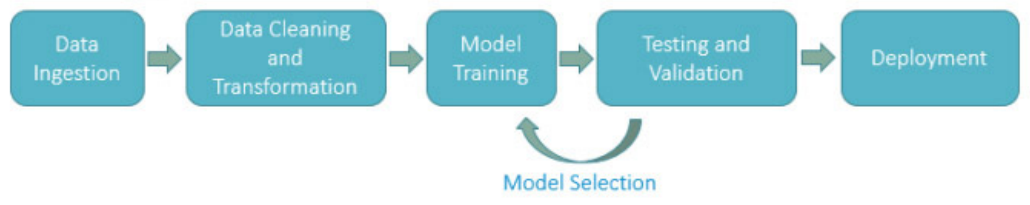
Existing AI technologies

- **AI technologies:**
  - Industry standard solutions to AI problems are Tenser-Flow, Keras, Sci-kit Learn
  - These packages are used in Python, and can be installed on any machine, including CUE.
  - Work on CPU/GPU and can potentially be ported to FPGA.

- **JLAB AI Environment:**
  - JLAB recently installed JUPYTER LAB, that can be launched on farm machines and run jobs interactively.
  - Several Machines with Nvidia-2080 RTX GPUs are available for Machine Learning Projects.

Existing AI technologies

- **Reconstruction Software:**
  - CLAS12 Reconstruction software is primarily in JAVA.
  - There are several Java Libraries that provide Machine Learning, including:
    - Deep Learning 4j
    - Apache-Spark MLib
    - Neuroph
    - Weka
  - Some Java libraries are based on Native C++ libraries compiled for local platform and have same performance as Python counterpart.
  - DeepLearning4J also supports GPU.
  - I tried all of them and found the DL4J to me best suited for our needs. Interface is more friendly and has most of networks useful for us.
- **Online Application:**
  - Java based AI can be easily integrated with existing CLAS12 tools for calibration and detector monitoring.



JAVA AI/ML/DL

- **PRO:**
  - The final product comes as one JAR file, runnable anywhere. (seriously, just one file), python installation installs thousands of files.
  - Most important network types are implemented, such as:
    - Multi-Layer Perceptron (MLP)
    - Convolutional Neural Networks (CNN)
    - Recurrent Neural Networks (RNN, LSTM, GRU)
  - Can run on GPU
  - Can run on Jupyter Lab with JAVA kernels, which are available on JLAB Jupyter.
  - Easily plugged in into CLAS12 workflow
  - Work on native C++ libraries, as fast as python libraries
  - Can run networks that were trained in python and saved as HDF5 file (some restrictions apply), but in most part works.

- **CONS:**
  - Documentation and examples are not as abundant as for Python libraries.
  - Have to switch Languages.
  - New Developments in AI are not ported very quickly into the library.
  - Not all network Types are available in the library, like GANs, but they will probably be ported as well.
  - Some data manipulation tools have to be implemented to match panda from python.
  - Have to program in Java.

Work Plan

- Work plan and its success depends on the available resources and gender neutral power.
- Everyone from collaboration is welcome to participate (I think this can qualify as service work)

Project Description	Timeline (Approximate)
Track candidate finding from DC clusters (continue the development)	Dec 2020
Extend track finding to LSTM for cluster finding in high luminosity runs	Aug 2021
Collaborate with Hall A/C to test our networks with their tracking applications	July 2021
Start development of Level-3 trigger AI & test it online	July 2021
Detector online monitoring using AI	July 2021
Investigate GANs for electromagnetic calorimeter shower generation	Dec 2021
Particle identification improvements using AI (post reconstruction)	Dec 2021

Estimate Cost and Identify Resources

- **Track Reconstruction Applications:**
  - During past year we have formed a collaboration with ODU CRTC department to assist us with Neural Networks training and software implementations.
  - CRTC has big resources for running training on latest NVIDIA cards on their farm. No Harare investments are needed.
  - Past year we paid on Grad student salary to ODU to work on Hall-B project, we can extend it to have 2 Grad Students to assist as with above mentioned projects with total cost of ~$50K.
  - From JLAB staff some supervision will be required (based on my previous experience about 0.5 FTE)
- **Online Monitoring and Triggering:**
  - At ODU they used our data to construct unsupervised train-less networks, mainly done for track path predictions, but can be used in Level 3 , Monitoring and Calibration applications.
  - With collaboration with computer scientists JLAB supervision will be required too (0.5 FTE)
- **Simulation AI:**
  - We were not able to identify resources and cost for this work yet.

Summary

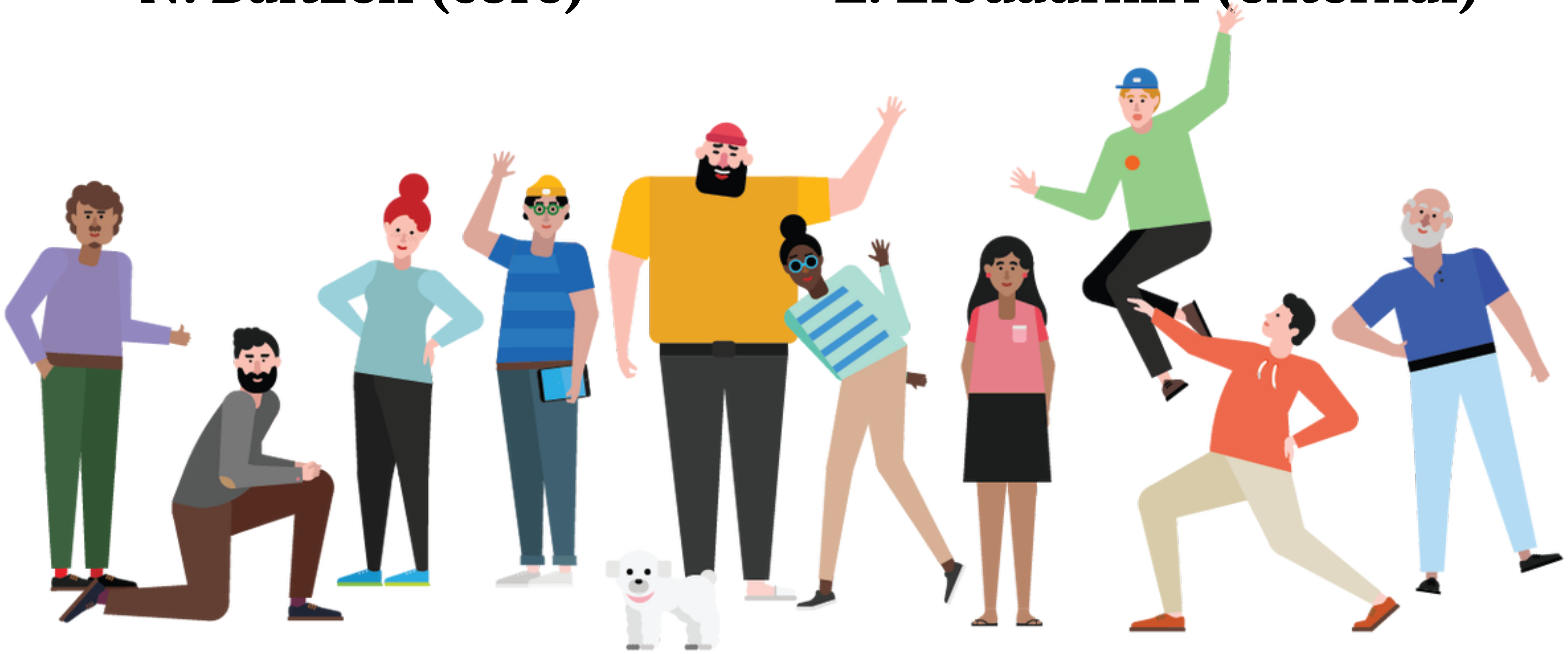
- **Areas of application:**
  - **Offline**:
    - Already positive results from track candidate identification (code speedup x6), will be integrated into workflow soon.
    - Continue this effort to improve track reconstruction efficiency (predicting track trajectory with LSTMs)
    - Collaborate with with other Halls to exchange experience and implementations.
  - **Online**:
    - Implement and test Neural Networks for Level-3 trigger, assess efficiency and speed, includes track identification from raw DC hits and clusters from raw EC hits, and matching them for electron trigger.
    - Implement online detector monitoring, collaborate with GlueX, they have established framework that can be trained on any data and can provide fault feedback.
    - Investigate how to do online data calibration, no work has been done in this direction at the LAB, we have to find out what HEP community is doing.
  - **Simulation**:
    - Look into the projects that already implemented shower generation in calorimeter (https://arxiv.org/abs/1712.10321)
- **Synergies:**
  - Collaborate with Hall A/C to test our AI tracking approach in their tracking applications
  - Collaborate with GlueX to implement their online monitoring AI software (Hydra) in Hall-B
  - Calorimeter clustering developments (needed for Level-3 trigger) can be shared with GlueX
  - Electromagnetic Calorimeter shower generation with GANs will be useful development for Hall A/C and GlueX
- **How to get comfortable with AI:**
  - Read more articles online about machine learning and deep learning, take some classes online,
  - attend AI Lunch seminars (Wednesdays 12:00PM).
  - There are many talks and tutorials on YouTube.

Thanks

Thanks to the team for contributions and advice !

Gagik Gavalian (PI)
D. Lawrence (PI)
V. Ziegler (core)
T. Britton (core)
N. Baltzell (core)

O. Hansen (core)
C. Fanelli (core)
M. Ungaro (external)
D. Heddle (external)
L. Elouadrhiri (external)



Backup