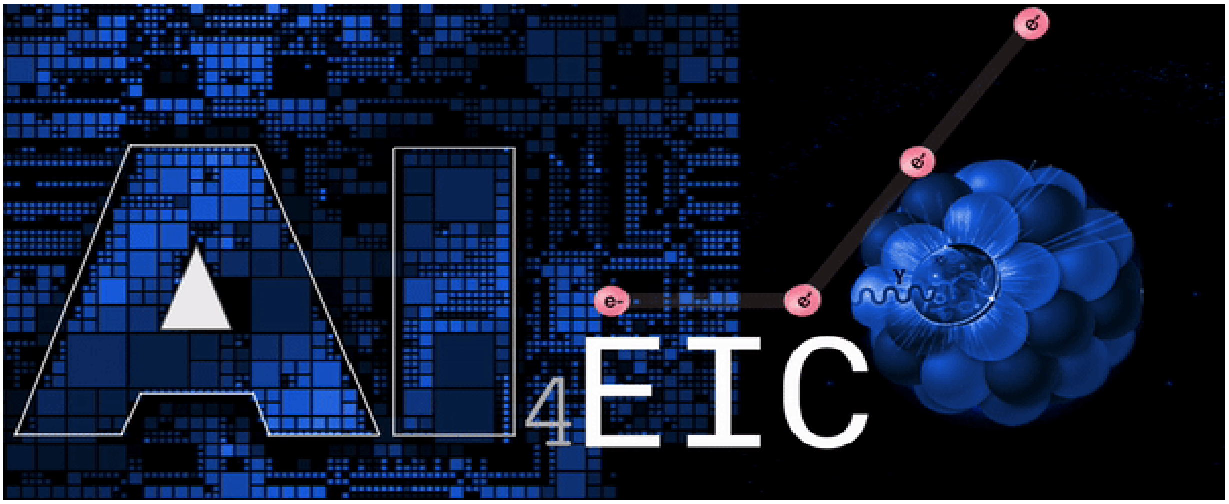

3rd AI4EIC workshop Nov 28 - Dec 1, 2023

Catholic University of America

Live Document:

Artificial Intelligence for the Electron Ion Collider

Timetable

This is the live meeting notes (Q&A) document for the second workshop dedicated to Artificial Intelligence for the Electron Ion Collider, which will take place at the Catholic University of America, Washington, D.C., from **November 28 to December 1, 2023**. The workshop will include sessions on:

- AI/ML for ePIC and Beyond (Nov 28, morning)
- Calibration, Monitoring, and Experimental Control in Streaming Environments (Nov 28, afternoon)
- AI/ML for Accelerators (Nov 29, morning)
- AI/ML for Data Analysis and Theory (Nov 29, afternoon)
- Foundation Models and Trends in Data Science (Nov 30, morning)
- AI/ML in Production, Distributed ML (Nov 30, afternoon)

NOTES — only drafted, please edit

November 28: Morning Session (AI/ML for ePIC and Beyond)

Talk: ePIC Overview
Speaker: John Lajoie

Notes

Text

Questions on talk

- **(Manouchehr Farkhondeh , DOE)**
- Comment :Nice to see AI incorporated from the beginning

- **(David Lawrence, Jlab)**
- Question: do you have institutions that are just datascience? Or DS focus?
- Answer: Yes, e.g. Brunel. Also want to make the point that this is a chance for new paradigms, focus on importance of DS for physics experiment

Talk: Performance Optimization for a Glass Electromagnetic Calorimeter at EIC
Speaker: Dmitrii Kalinkin

Notes

Text

Questions on talk

- **(Your name, Institution)**
- Question: Can the optimal solution be manufactured? (precision, etc.)
- Answer: Yes, you need to evaluate the uncertainties, e.g. effect of misalignment on optimal solution. Follow up question? Can this be incorporated easily in the framework? Not a software or math problem, just have to implement benchmarks with imperfections/variations.

Talk: Object Condensation for Track Building in a Backward Electron Tagger at the EIC
Speaker: Simon Gardner

Questions on talk

- **(Your name, Institution)** How robust against missing hits? What physics analysis/observables would validate on real-data when this will be operated in a realistic scenario?
- Question:text

Answer: The initial tests on missing hits showed the reconstruction was still very successful. There are a lot more investigations that need to be carried out to verify this, it might be the case that only hits in the final layer are being used as condensation points, this will naturally have the points from different hits easiest to discriminate between. However if the hit in the final layer is missing on a relatively rare occasion it won't be classed as a track. These things all still need to be checked and the training sample tuned to make sure it is robust to these situations.

Talk: The Optimal Use of Segmentation for Sampling Calorimeters
Speaker: Fernando Torales Acosta

Questions on talk

- **(Your name, Institution)**
- Question:text

Answer: text

Talk: AID(2)E AI-Assisted Design at EIC
Speaker: Cristiano Fanelli

Questions on talk

- **(David Lawrence, JLab)**
- Question: can this be used by a common collaborator or does it need to be coordinated

Answer: want to provide a framework that can be used by the experiment. Plan to assess the effectivity/portability of the methods

- **(Casey Morean, CUA)**
- Question: In the demonstration is this inherent in the simulation or something that needs to be designed specifically? [CHECK]

Answer: What is needed are the design points. Utilize the most effective tools in the design phase to get relevant insights and adapt the tools for following phases [CHECK]

Flash Talk: Pi0s with BDTs
Speaker: Gregory Matousek

Questions on talk

- ***(Gagik Gavalian: JLab)***
- Question: Why gradient boosted trees?
- Answer: speed and convenience/performance. NNs as good

Flash Talk: ML on FPGA
Speaker: Dmitry Romanov

Questions on talk

- ***()***
- Couldn't hear…
-

November 28: Afternoon Session (Calibration, Monitoring, and Experimental Control in Streaming Environments)

Talk: LHCb Calibration/Alignment
Speaker: Biljana Mitreska

Questions on talk

- **(Your name, Institution)**

- Question:text
- Answer: text

Talk: ML-based Calibration and Control of the GlueX Central Drift Chamber
Speaker: David Lawrence

Questions on talk

- **(Nuwan Chaminda, University of Virginia)**
- Question: When the HV of the Drift Chambers changes, how the leak currents of the CDC gets changed? Would the HV trip if it has not been HV trained when HV changes, depending on the temperature, pressure?
- Answer: The control software has a ramp rate built into it, so the change in HV is not instantaneous, it takes a few seconds, long enough that the HV does not trip. Also, the maximum change in HV due to the tuning would be from 2110V to 2140V or vice versa, so this is a fairly small change (30V) around a large potential (2125V). Most of the voltage changes due to pressure or temperature are +/- 1V. This is because we have an alarm system set up to tell us to start a new run if the pressure has changed by 0.1kPa.

Talk: Towards Fast Calibration with the ePIC Barrel Hadronic Calorimeter
Speaker: Derek Anderson

Questions on talk

- **(Your name, Institution)**
- Question:text
- Answer: text

Talk: Hydra: Computer Vision for Data Quality Monitoring
Speaker: Thomas Britton

Questions on talk

- **(Your name, Institution)**

- Question:text
- Answer: text

Talk: Fast 2D Bicephalous Convolutional Autoencoder for Compressing 3D Time Projection Chamber Data
Speaker: Yi Huang

Questions on talk

- **(Gagik , Institution)**
- Question: what kind of CPU did you use to support the claim it is too slow in your slides?
- Answer: We did not really use any CPU…we “assumed” it is too slow for this

Talk: Autonomous selection of physics events: A RHIC demonstrator for EIC physics
Speaker: Cameron Dean

Questions on talk

- **(David Lawrence, JLab)**
- Question:Very impressive. No real question.
- Answer: text

November 29: Morning Session (AI/ML for accelerators)

Tutorial: Continual Learning

Questions on talk

- **(Casey Morean, CUA/JLab)**
- Question: would you be able and willing to post the slides on the indigo page, please?
- Answer: I don’t think I have access to the page but I sent the slides to the organizers.

- **Casey Morean, CUA/JLab)**
- Question: Often, NP workflows are spikey in terms of cpu/memory resource utilization. Otherwise, the computers are sitting idle. Are there existing packages for continual learning that aim to utilize these compute resources and stop quickly when high priority tasks start? Are there ways to trigger continual learning on specific events to reduce the dependence on a human operator? As an example, say an alarm goes off, flag that period in time and try to identify possible ramp up to event. Similar to root cause analysis.
- Answer: At the moment there are no ready-to-use continual learning packages to do that.The main way to go would be to integrate existing continual learning solutions into the tools that monitor the system and/or interact through API if they provide such possibility.
- **Casey Morean, CUA/JLab)**
- Question: In the NP analysis, there is an effort to archive the code, environment, and dependencies for future re-analysis and publications. In production, are any of the packages shown able to go back to a previous trained state to ensure repeatability?
- Answer: Yes, for example in Avalanche you can easily use plugins to save previous states of the model and go back to that state (usually, this is called checkpointing).

Session AI/ML for Accelerators 29 Nov 2023, 10:00

Summary

The accelerator session aim was to give overviews of successful efforts on the use of ML for various accelerator applications. In general, the techniques described have been applied either to operations or in beam experiments. There is tremendous growth in the use of ML in accelerators, which is evident from conference papers (see IPAC’23, ICALEPCS’23, etc.) as well as peer reviewed papers (a search in PRAB will show a long listing). These talks represent the cornerstones to applications of ML that will be useful in the EIC.
Kevin Brown, 12/4/23

Ryan Roussel Title: Bayesian Optimization Techniques for Accelerator Control and Characterization

Questions on talk

(Name, Institution)

- Question:
- Answer:

Lucy Lin Machine learning for digital twin development and polarization optimization at BNL hadron injectors

Questions on talk

(Name, Institution)

- Question:
- Answer:

Matt Kilpatrick Beam Condition Forecasting with non-destructive measurements at FACET-II

Questions on talk

(Name, Institution)

- Question:
- Answer:

Xiaofeng Gu, Machine Learning applications for collider luminosity maximization

Questions on talk

(Name, Institution)

- Question:
- Answer:

Daniel Ratner, Anomaly detection at an X-ray FEL

Questions on talk

(Name, Institution)

- Question:
- Answer:

Malachi Schram, Uncertainty estimation and RL applications at JLab

Questions on talk

(Name, Institution)

- Question:
- Answer:

Frederik Van der Veken, Using Machine Learning to Improve Dynamic Aperture Estimates

Questions on talk

(Name, Institution)

- Question: Are the produced DA estimates further used for the optimization?
- Answer: It is a common procedure to find e.g. tune settings to improve the DA, currently there is also work ongoing to apply Bayesian Optimization for DA optimization - specifically for this case, the presented techniques to obtain a fast DA estimate using MI is very useful, to make the optimization less time and power consuming.

Notes on discussion:

- Question on the collaboration with industry partners e.g. on software development of ML
  Answer: parts of the fundings on ML related projects can be reserved for the collaborators. Collaborators such RadiaSoft are interested to understand better the use cases and the needs in the accelerators domain.

Session AI/ML for Data Analysis and Theory 29 Nov 2023, 14:00

Simonetta Liuti: The EXCLAIM collaboration approach to deeply virtual exclusive processes

- Questions: How is using AI different from a standard fit?
- Answer: Compton form factors as an example: Fit based methods disregard the correlation when a point by point analysis is performed, which can be included with the AI/ML. Additionally, you can investigate the latent space to identify structures that have physics information.

Yaohang Li: Decoding inverse problems in QCD with ML-based algorithms

- Questions: What documentation do you recommend we read to reproduce the toy inverse problems presented?
- Answer: https://github.com/alanaziyasir/VAIM

Huey Wen: What can AI do for lattice-QCD parton distribution calculations?

- Question:
- Answer:

Charles Hughes: Interpretable Machine Learning applications to Jet Background Subtraction

- Question:Computing constants with symbolic regression. With different bins in R and pT? How many models do you have to train?
- Answer: In R, training across individual bins. In pT you are training across the entire pT range.
- Question: Derive function using symbolic regression, what if you had another variable? What's the next one?
- Answer: slide 22, these are the best 8 you can get. The next one down in violet has 3 constants and looks like the one above. If you look at 9, it has a higher complexity score but doesn't do much better for the overall score for how well it performs. d ln(loss)/d complexity. None perform better than the chosen metric. Best performing model while being least complex. Not a lot of optimization done on this study. There was preprocessing, but tweak these things and see if it comes out differently.

Benjamin Nachman: High Dimensional Unfolding using Machine Learning

- Question**:** One issue I had when implementing OmniFold for my analysis was that the number of reco-data events needed to equal the number of Monte Carlo reco/truth events. i.e., Although we had several times the amount of Monte Carlo to use for training, we needed to truncate it to match the size of our data. Is this a limitation of OmniFold and if so what would be the best way to get around it?
- Answer:
- Question: OmniFold was one of the first methods. What were the challenges, especially with the diffusion models, to bringing this to experimental facilities?
- Answer: Sociological (we have to really show that we know what's going on at all levels) and technical. Technical - we need to achieve precision. Challenge because when people do binned methods, teh response matrix is a constant. But now imagine the response matrix is a NN. For many applications in NP you train a classifier, and the small jitter from training isn't a huge problem, but here it is. Imagine that your XSX changes by

1%, and systematic is 1%, if you have 10 systematic uncertainties so you have to add up all of these in quadrature. We have 2 options, beat this down with ensembling, or developing more robust methods. This is true for classifier based methods and generative methods. Closer you start towards the right answer the more robust it will be.
- Question: Slide 32, what do you mean by unbinned and benefits?
- Answer: The actual measurement is done in a smooth way. If I wanted to change the binning it's 30s of python versus rerunning the whole analysis. The underlying result is a NN and the NN is a smooth function, there's no binning. So you can re-run on the fly. If I made the binning arbitrarily fine, then you would find that the covariance matrix is very correlated between bins, so best to try and get as close to the binning with the experimental measurements.

Manuel Szewc: Hadronization Models Using Machine Learning

- Question: It would be great to put spin in there.
- Answer: I agree completely, we are also thinking about how to do this. Trying to tackle the problem in a modular way from simple to more complex. All of the subtleties should be taken into consideration.
- Question: Translate Normalizing flow in slide 9 to the LHS of slide 10?
- Answer: Learn the individual hadronization function, take an initial string and break it. Do this iteratively until we run out of energy. Threshold is 20 GeV. Singlet, break it and subtract energy from string. Boost to the string center of mass. Simplified version of what Pythia does. We can also use the user hooks part of Pythia to do this as well, this is just faster.

Discussion points

- You can't learn without making some kind of assumptions. Observable, tailor algorithm or process to that observable. For given observables, there won't be the same algorithm from one to the other.
- There are 3 things that I (Brandon) think have been identified:
    - Learning the underlying theory of QCD in a human readable way
    - Interpretability - can we separate spurious correlations from actual learned physics?
    - UQ - separating aleatoric, epistemic, and OOD sampling. We need to know what our models do/do not know.
- Workshop title called benchmarks to understand what is a ML model that is useful for the EIC.
- LHC there are a lot of community challenges to benchmark
- We need the infrastructure to incorporate ML into experimental facilities. What do we need to focus on to take advantage of the work?
    - Going towards near real time analysis and streaming, especially at the LHC.
    - High quality data that we eventually want to extract the physics online.

Tutorial on Reinforcement Learning – Haipeng Chen

- Question (David Lawrence, JLab): It wasn't clear to me how the RL methods include time-series information. For example, when teaching a model to play Atari games, the state is given by the image (single frame). However that often does not contain info on motion. (e.g. which direction and at what speed is the ball traveling in Pong?) Is the model itself recording this like in LSTM, or do you need to include this as part of the state by providing some number of previous frames?
- Answer: Good question. In short, yes time-series information is indeed captured. Both of your thoughts are reasonable methods – in the original version of DQN of the Nature paper, m (e.g., m=4) most recent frames are stacked and fed as input to the CNNs, in another paper, LSTM was used to extract that information. More recently, there is this Decision Transformer paper that models the motion information using more advanced sequence models like Transformer. This is also one of the initial papers of a line of research that use RL+LLMs-empowered agents for decision-making.

Session Foundation Models and Trends for Data Science 30 Nov 2023, 10:00

Kazuhiro Terao: Introduction and Overview of Foundation Models

- Question Vinicius Mikuni (National Energy Research Scientific Computing Center): The question is about the correctness. For example, in ChatGPT, some questions can be checked by their correctness. The image shown in the slide is on Amplitude. How to verify its correctness?
- Answer: I don't know. We refer to the PI to answer this question.

Rick Stevens: The Trillion Parameter Consortium

- Question (Daniel Murnane, LBNL): A question about using LLMs to consume structured data - isn't there a more natural way to consume such data.
- Answer: Yes, it may be a short-term solution, but it's a functioning way to solve the problem. Other groups are encouraged to try to solve this in different ways. But this templating technique can interface with a database of structured data

- Question (Karthik Suresh, W&M): Elaborate AGI for Science?
- Answer: Basically the same definition as industry. Need academy, lab community and industry to work together towards it.

Karthik Suresh: A Large Language Model-based Assistant for the Electron Ion Collider

- Question: (Manuel Szewc): Slide 8, How much engineering is done?
- Answer: No recipe in doing that. Depending on the underlying model. Few iterations until matching exactly. But start with putting in assertive statements.

Daniel Murnane: Chatlas - the ATLAS AI Assistant

- Question: (Karthik Suresh, W&M) Is the Flask server scalable?

- Answer: Yes it is. It can handle thousands of requests.
- Comment: (Karthik Suresh, W&M) Flask API-based solution langserve to serve locally and a remote database that can connect to the API.

Xiangyang Ju: TrackBERT - Generalist Learning of Trackers for Downstream Tasks

- Question: (Gagik) Do you use straight track, i.e., predicting the next from the previous one?
- Answer: Yes, that is our current approach. Our ultimate goal is to use all tracks.
- Question: (Yaohang Li, ODU) Should the scaling plot be related to 3D: Loss, number of parameters, and data size?
- Answer: Yes. But in our plot, we assume almost infinite model size and trained for almost infinite time.

Eric Yeats: Adversarial Methods for Generative Data Understanding

- Question: (Yaohang Li, ODU) For detangling the latent space, your algorithm depends on the hyperparameters. Is there any guideline to select the appropriate hyperparameters?
- Answer: Sure. One way to do that is to run multiple models with respect to different parameters. Then we can calculate the MSEs correspondingly. If there is a sharp decline, it is a good indicator of the appropriate parameter.

Tareq Alghamdi: Toward a generative modeling analysis of CLAS exclusive 2π photoproduction

- Question
- Answer

Discussions:

Session AI/ML in Production, Distributed ML 30 Nov 2023

Luca Giommi

- Q:

Maja Karwowska

- Q:

Wen Guan

- Q:

Xiangyang Ju

- Q: