CMS ML KNOWLEDGE GROUP

NEWSLETTER

v1, October 2024

INSIDE THIS ISSUE

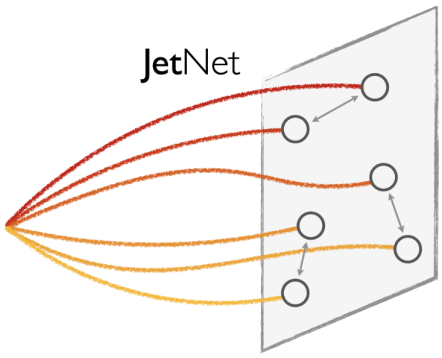
ML NEWS

WHAT YOU NEED TO KNOW

Welcome to the first edition of the CMS ML Knowledge Newsletter brought to you by the ML Knowledge Group. Read on to find out about the latest happenings in the CMS ML world.

OPEN DATA

Are you hungry for free open access data? Here are some of the latest datasets free to use:



JetNet - a project aimed at enhancing accessibility & reproducibility in jet-based machine learning

- Github Repo: https://github.com/jet-net/JetNet
- Zenodo: https://zenodo.org/records/6975118

Top Tagging Benchmark Dataset - a set of MC simulated training/testing events for the evaluation of top quark tagging architectures

- Zenodo: https://zenodo.org/records/2603256

ADC 2021 Dataset - benchmarking samples for development of trigger-level anomaly detection algorithms

- Challenge website: https://mpp-hep.github.io/ADC2021/

To learn more: https://cms-ml.github.io/documentation/resources/dataset_resources/index.html

DATASET RELEASE

Do you have a dataset that you would like to release? Contact us for more information on the steps to get your dataset released publicly.

OTHER DATA SOURCES

- CMS Open Data Portal
- Testbeam data on zenodo: HGCAL electrons, pions
- Recent ML Forum on: ML Dataset Release (restricted link)

SCHOLAR SPOTLIGHT

We interviewed Farouk Mokhtar, a graduate student at UCSD currently residing at CERN. Read on to learn more!

Q. Could you tell us a little more about your background and how you got into particle physics?

A. I studied physics in my undergrad at the University of Science and Technology, at Zewail City, in Cairo, Egypt. My entry point to particle physics was when I got a chance to participate in the DESY summer school 2019 as a summer student. I was assigned a project to develop a top tagger for a search at the LHC and I got a chance to learn about Machine Learning and how to apply it to LHC physics. This gave me the motivation and excitement to pursue graduate school in particle physics and to learn more about the research being done at the LHC experiment.

Q. Could you tell us a little more about your research?

A. I work on developing machine learning models to solve challenging problems at the LHC. A big part of this work is under the umbrella of reconstruction, where we develop graph neural networks and transformer models to learn to reconstruct high level event objects (e.g. particles, and jets) from low level input (e.g. detector level hits, and energy deposits). Another part of the work is in developing jet taggers, essentially classifiers to tag jets of different flavors, which are crucial to identify fundamental particles like the top quark, or the Higgs, which are very important to understanding the Standard Model.

Q. A lot of us use centrally provided ML taggers and frameworks. But, you work in designing something brand new. What is a challenge you have faced with your research and how have you overcome it?

A. The challenge when doing something that no one has done before (at least in the same fashion) is that the possibilities are endless. There are so many interesting ideas that could all help tackle the problem but you need to make choices on which ideas to pursue because of lack of time/personpower to explore all of them. There are also choices to be made in how you define the problem itself, and how to make sure the baseline algorithm that you are competing against is optimized with respect to that problem and to the dataset that you are evaluating on.

Q. Do you have any advice for researchers who may want to get involved with creating a new ML approach like you have?

A. My advice would be to be patient when developing ML models because they sometimes need time to train well enough to compete with other algorithms. In addition, if you are trying to attack a challenging particle physics problem using ML, you need to have a very good understanding of the problem at hand, and how it can best be mapped to a problem that can be tackled with ML. Try to isolate what is working from what is not. Try to dissect the big problem into smaller pieces that can be evaluated separately, so that it is easier to identify what is missing to get your ML algorithm to the next-level in terms of performance.

Q. What are you interested in doing in the future?

A. I am currently a PhD student, expected to graduate in about 1.5 years. I am interested in continuing to pursue research at the LHC. I find great fulfillment in participating in possibly the biggest science experiment in the history of humanity. The diversity that is found within the particle physics community at CERN is exciting. Interacting with people from different backgrounds and origins, and regularly exchanging ideas and thoughts is something that motivates me to pursue my research. I also find great excitement in leveraging the latest state-of-the-art machine learning tools and methods that are being developed in other fields (e.g. computer vision, and natural language processing), and adapting/integrating them into our physics research

Q. What do you do in your free time?

A. When I am not at CERN, I love playing squash in the maisonnex sports club right across from CERN, reading novels (among my favorite books are: the kite runner, and the book thief), and meeting new people from different backgrounds which is made easier by engaging in activities at CERN or in Geneva since the international community is very big.

ML CORNER

Many new ML tools are being developed within CMS to deal with the increased luminosity of the HL-LHC. One new proposed algorithm is Machine Learned Particle Flow (MLPF). We talked to one of the main developers of MLPF, Joosep Pata from KBFI .

Q. Why did you create MLPF?

A. I became interested in particle flow reconstruction during my PhD studies, when the algorithm was being upgraded for Run 2. I overheard that there was significant work in extending the algorithm and making it scale to higher pileup, but that it was not easy. Modern deep learning was just starting to be applied to b-tagging in CMS, and we had a random discussion over lunch with my supervisor about how it could conceptually be used for particle flow reconstruction. It was a bit too early to start developing in this direction, but the idea of machine learning based particle flow reconstruction seemed interesting. Later on, at Caltech, I learned about the ExaTrk.X project, where folks were developing ML-based tracking. It seemed like a good idea to extend these methods to particle flow reconstruction, as the tasks did not seem conceptually very different. I was essentially looking for a challenging and relevant problem to tackle with ML. Particle flow suited this goal well, as it can be computationally challenging, it can have a significant impact on physics performance and similar reconstruction algorithms may be useful in future detectors.

Q. What does MLPF do?

MLPF aims to reconstruct stable particles throughout the detector, based on the tracks and clusters in different layers of the detector. The idea is based on standard particle flow, which is a rule-based algorithm and which works very well in production. MLPF aims to replace the rule-based algorithm with a machine-learned version, such that the model can be tuned to new detector conditions easily. One example would be if a timing layer is added to the detector, the ML-based algorithm can in principle be retrained from scratch, without having to necessarily re-engineer the reconstruction completely. The other potential advantage is the portability to new computational hardware. Standard heuristic algorithms can be challenging to port to new systems like GPUs or other computational accelerators, often requiring extensive re-engineering or a completely different approach than for CPUs. On the other hand, ML models can run reasonably efficiently, as the underlying ML code is developed for high-performance usage in mind. However, heuristic algorithms can also be very efficient if standard computer science approaches are applicable to the problem at hand, so realistic throughput tests need to be done with fully optimized approaches before any deployment decision.

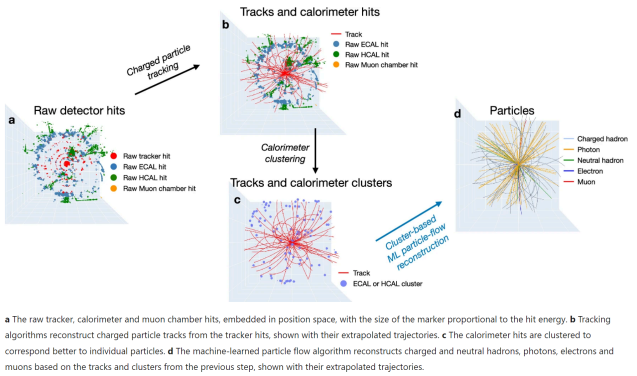
Q. What kinds of computing tools does MLPF rely on?

The training code is based on a modern python stack and efficient array-based tools developed for HEP. We rely on pytorch for model development, and tools from scikit-hep such as uproot, awkward-array, fastjet, matplotlib-hep for data manipulation and visualization. To be useful for CMS, the model needs to be interfaced to CMSSW - for this we currently use ONNX, although we have also had good results with inference as a service. To train the model, we use HPC centers, for example the LUMI HPC center in Finland, the Julich HPC center in Germany or the Voyager platform at UCSD. We had to ensure the training code works seamlessly on AMD, Nvidia and Intel Habana deep learning platforms, which was fairly straightforward with standard pytorch. We have reached a convergence point in HEP ML, where we can mostly use industry standard open source software. This is useful for training relevant skills in students and being able to quickly ramp up new collaborations.

Q. Anything else interesting about MLPF you would like to add?

Throughout the project, it's been very useful to collaborate with experts within CMS, but also with folks from outside of CMS. For this, we have set up the project such that besides CMS internal datasets, we can also work with realistic open datasets that are relevant for future collider studies. In ML projects such as this, besides the specific neural network architecture and training, a lot of effort goes into the basics that come before ML: generating and debugging the dataset, the definition of the problem and the target for the model, the metrics used to define success, and the engineering work in making the approach scale on different platforms and with large dataset sizes. We've aimed to develop a sustainable project where different ideas for ML-based reconstruction can

be tested, rather than focusing on a specific modeling approach. If people are interested to contribute, we would welcome all kinds of contributions, e.g. ML engineering, data engineering, documentation work, as well as questions/comments that may help to see the problem from a different light. It's pretty unlikely that we have reached the global minimum in terms of what can be achieved. Ultimately, for such models to be useful, they have to be extensively tested on data, and CMS is uniquely placed for that. Even if this specific approach will not end up in production, such research may inform future developments in HEP data analysis and reconstruction



a The raw tracker, calorimeter and muon chamber hits, embedded in position space, with the size of the marker proportional to the hit energy. **b** Tracking algorithms reconstruct charged particle tracks from the tracker hits, shown with their extrapolated trajectories. **c** The calorimeter hits are clustered to correspond better to individual particles. **d** The machine-learned particle flow algorithm reconstructs charged and neutral hadrons, photons, electrons and muons based on the tracks and clusters from the previous step, shown with their extrapolated trajectories.

image from Pata, J., Wulff, E., Mokhtar, F. *et al.* Improved particle-flow event reconstruction with scalable neural networks for current and future particle detectors. *Commun Phys* 7, 124 (2024). https://doi.org/10.1038/s42005-024-01599-5. image via: Creative Commons Attribution 4.0 International License.

CMS TOWN HALL

The 2024 ML Town Hall & Foundation Model MiniWorkshop was held on Sept 30 - Oct. 2, 2024. It featured contributed talks, overviews & status of forthcoming publications, discussions, and contributions from external experts in the community. To learn more, click here (note: restricted to CMS members)

EVENTS ON OUR RADAR

FAST ML: OCT 15-18
CHEP24: OCT 19-20
ML4JETS: NOV 4-8
NEURIPS 2024: DEC 9-15

EPR

Do you need EPR? We are looking for contributors. See our most recent list of tasks here (note CMS internal web only)

JOIN US

This newsletter was brought to you by the CMS ML Knowledge Group. We meet every three weeks, and welcome new members! Most recent indico here: https://indico.cern.ch/event/1465462/ (restricted).

Join our mattermost! We have a Knowledge sub-channel under the channel "CMS Machine Learning Channel." Join the CMS Machine Learning Channel first, then you can join subchannels.

If you've enjoyed this newsletter, please let us know. Also, if you have an idea for something you would like to see in this newsletter, would like to nominate someone for an interview or ML corner spotlight, please let us know! Email: Melissa Quinnan & Jieun Yoo: cms-conveners-ml-knowledge@cern.ch