PhD Proposition in the OGCID Project

Title : Optimal graph convolution neural network for efficient particle identification

Supervisor : Frédéric Magniette, Laboratoire Leprince-Ringuet (LLR), CNRS/Ecole polytechnique, frederic.magniette@llr.in2p3.fr

Funding : project OGCID ANR-21-CE31-0030, 3 years

Location : LLR, Ecole Polytechnique, 91128 Palaiseau Cedex – France

Context

Most of the recent discoveries in particle physics are linked to the increase of the detector volume and / or granularity to observe complex phenomena that were inaccessible previously due to a lack of precision. This approach increases the available statistics and precision by multiple orders of magnitude which facilitate the detection of rare events, at the price of a significant increase of the number of channels. The challenge is that most of the standard techniques for reconstruction and triggering are not operative in such a context. For example, the energy threshold-based triggers fail to handle the complexity of the high pile-up collisions. The neural network methods are known to handle well the noisy and complex data inputs to deliver high level classification and regression. In particular, the convolution techniques [1] have allowed outstanding improvement in the computer vision field. Unfortunately, they do not cope with the very peculiar topologies of the particle detectors and the irregular distribution of their sensors. Alternatives have been discovered to obtain the same classification power in that kind of non-euclidean environment, for example, the spatial graph convolution [2] which applies adapted convolution kernels to the data represented as an undirected graph labeled by the sensor measurements. These techniques have proven to give excellent results on the particle detector data at LHC [3] but also for neutrinos experiments [4]. They allow particle identification and continuous parameter regression, but also segmentation of entangled data which is a typical concern in secondary particle showers.

The operations that transform the data into a graph are often very computationally expensive. In particular, all the techniques in which this operation is based on learned parameters (in the sense of machine learning) prevent the system from being used in a context where the computational time or latency are constrained (any triggering electronics, real-time data monitoring systems or even offline systems with a too big data volume). For example, in the Super-Kamiokande neutrino experiment, a complex shape identifier would advantageously replace the current energy cut during the reconstruction phase that rejects many low energy events despite their physical interest. Another example is the future high-granularity endcap calorimeter (HGCal) of CMS [5] for which it becomes crucial to be able to extract high level trigger primitives directly from the electronics to handle the complexity of the high luminosity collisions and take accurate triggering decisions. This is why, it is of utmost importance to design high-performance versions of these algorithms, which can increase the performance in all the constrained situations and allow their realization in the detectors.

The objective of this project is to develop and implement a new efficient selection algorithms for constrained computational environments by combining three main ideas

- Reducing the graph construction complexity by developing algorithms based on pre-calculated graph connectivity which would allow obtaining a linear complexity for the online part by exploiting intrinsic parallelism of the problem. This is made possible by the fixed positionning of the sensors in the particle detectors.
- Developing segmented version of graph convolution, allowing to distribute it over multiple computational unit (CPU, GPU, FPGA…). A typical segmentation is induced by the distribution of the data in the off-detector trigger devices and must be handled by partial graph convolution followed by a concentration phase.
- Optimizing the size and the nature of the convolution networks with advanced techniques of derivative-free optimization and adaptation to the electronic implementation.

These objectives will be declined in the three experiment contexts: Offline HGCal reconstruction, Online HGCal level 1 trigger and Super-Kamiokande reconstruction of the Diffused Supernova Neutrinos Background (DSNB).

HGCal is the future front-end calorimeter of CMS. It is a Silicon-based very high granularity sampling calorimeter aiming at measuring energy, position and time-stamp of particles with 6.5 millions of channels from Si-based sensors [5]. It is a very ambitious project in terms of detection technology but also in terms of analysis software, because it generates huge amount of data describing very entangled events. These events has to be selected in the level 1 trigger with a very low latency (few micro-seconds).

The offline HGCal reconstruction is done by a complex system named TICL [11] which is based on different layers of identification applied iteratively to recognize different kinds of particles (unambiguous electromagnetic showers, hadrons, minimum ionization particle…). These different iterations are implemented in a plugin system that to ease the test of alternative reconstruction techniques.

The current system is a demonstrator based on legacy techniques, but not much effort could be dedicated to the performance. In particular, its discrimination power is not sufficient to implement particle flow algorithms. This is why it is necessary to try new techniques, including the graph convolution.

Even if the time constraint is not as important as in online systems, the collected data volume is so big with the new high granularity sub-detectors that it is of major importance to fully optimize this algorithms and this is why this project is completely in line with the need of the HGCal project. Another reason is that it would be very interesting to implement also this kind of techniques in the High Level Trigger (HLT) in which the latency is a major issue. There is a lot of advantages to use the same technique in both HLT and offline because it allows cross-developments and simplifies the downstream determination of the performance of the online algorithms.

The HGCal trigger is composed of different layers. In particular, the Level 1 (L1) trigger is implemented in the electronics and is providing trigger primitives to upper software layers (Central sub-detector triggers). The L1 trigger is composed itself of two layers of FPGA boards. The first layer is implementing a data reformating and send it to the second layer.

There, a seeding operation is performed to identify the hot spots around which hits are related to the same particle. The hits are then clustered by radius around the seed. From theses hits collections, a particle identification and a regression on its energy are performed. This part is done with energy counting algorithms, with energy correction to compensate energy losses and pile-up contamination. This algorithm will become less efficient if the pile-up becomes too high. It has been demonstrated in [3] that the Graph Convolution Neural Network (GCN) techniques would be really better in terms of identification but also that their actual implementation is not compatible with the resource available in a FPGA but also not compatible with the available latency in the triggering process [10].

This is why it could be very valuable to fully optimize the algorithmical part of the convolution as well as the topology of the neural network implemented in the FPGA. An alternative would be to split the convolution in parallel parts on different FPGA. The current architecture of the HGCal trigger does no allow this but it is interesting to study it and provide pertinent arguments for an evolution.

One of the objectives of the Super-Kamiokande experiment is to observe the Diffuse Supernova Neutrino Background (DSNB) [8]. It is a flux of neutrinos cumulatively originating from all of the supernovae events which have occurred throughout the Universe. This is a very important cosmological probe, linked to the time evolution of the star production rate. Until now, it has never been observed because it is presently much smaller than the background. As shown in Figure 1, the DSNB spectrum competes with three different backgrounds: reactor production (dashed red line on the left) which is irreducible, cosmic muon spallation in pink and atmospheric neutrinos in blue. The spallation and the atmospheric neutrinos have very different signature than DSNB and could be eliminated by a shape detection technique, possibly GCN.

Indeed, the detection reaction in Super-Kamiokande involves an anti-neutrino interacting on a proton, producing a positron and a neutron. After thermalization, the neutron is captured by an atom (with 20% of efficiency in water, and currently 50% thanks to the addition of gadolinium). This capture on hydrogen or gadolinium induces the emission of a photon at 2.2 MeV or 8 MeV, interacting with electrons of the water or creating a pair electron/positron. This induces a particular signature in the image collected by the detector. If we can train a GCN to learn this particular shape, we could reduce the background significantly and thus reduce the energy threshold which is presently used. With these improvements, we hope to be able to measure the DSNB in the narrow but significant window colored in light green in Figure 1.

The complete project (funded by ANR) is described in the website https://llrogcid.in2p3.fr

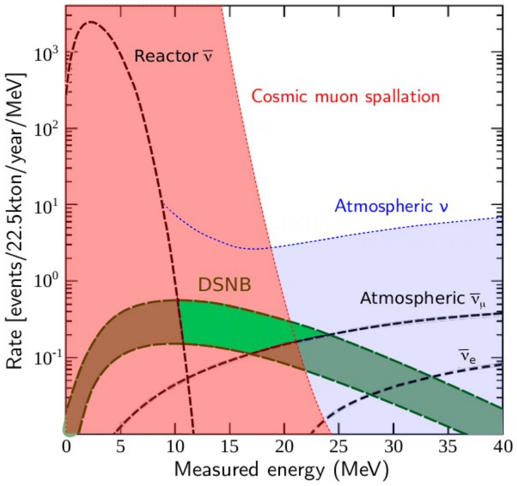

Figure 1: DSNB and background spectrum

Detailed Project

The PhD project is divided into 4 different work-packages

Work Package 1 (WP1): Data graph construction, particle classification and regression

The representation of the data is essential to extract the relevant information from our data-sets.

The first goal of the WP1 is to develop algorithms to build graphs from data with low computational complexity. The techniques of dynamic construction based on neural architecture are unusable in a real-time constrained environment or when the data volume is very high. A way to reduce the complexity of this operation is to pre-calculate a part of the potential graph (the distances between sensors) and to store it, allowing reduction of the dynamic part of the algorithm, hoping to reach a linear complexity. A complete study of the different distances has to be performed and tested on simulation to validate the approach.

The second goal of the WP1 consists in applying the graph convolution to these pre-processed inputs and to compare them to the reference KNN algorithm generated graphs. The first application will be to classify different types of particles from simulations, with different initial parameters (energy, incidence angle, first hit position…). The second application will be to estimate with the same convolution these initial parameters themselves. These code should be written as reference benchmarks to test different graph generation techniques.

Of course, multiple iterations between these two goals must be performed to identify simultaneously the best graph construction and the best graph convolution topologies to carry out classification and regression. These tasks will be performed on simulated data-sets, that will be produced specifically. In particular, we need to simulate single particle interactions with high granularity sampling calorimeter (electrons, pions, photons, muons...) and neutrino interactions in Kamiokande experiments (signal / background).

The WP1 is divided in 5 tasks

T1: Algorithms selection for HGCal offline

- Design graph generation algorithms based on detector symmetry (for both Super-Kamiokande and HGCal).
    - Implementation in Python.
    - Test of properties of the differents algorithms (Mean arity, mean connectivity...)
- Test GCN on the graphs.
    - Implementation with Pytorch and Pytorch-geometric.
- Compare the obtained efficiency with standard methods.

T2: Algorithms selection for HGCal online

- Use T2 implementation on HGCal data to select the best algorithm.

T3: Algorithms selection for Super-Kamiokande

- Use T2 implementation on Super-K data to select the best algorithm.

T4: CMS TICL plugin implementations

- Implement a TICL plugin based on the selected algorithms
- Compare the performance with the standard method

T5: Publication of the comparison studies

- HGCal offline reconstruction efficiency study based on simplified detector simulations
- HGCal offline reconstruction efficiency study based on data from CMSSoftware (i.e. with real detector and noise) with DQM plots
- HGCal online trigger efficiency study and rate reduction study without consideration of optimization
- Super-Kamiokande background rejection efficiency study

Work package 2 (WP2): Semantic segmentation

The goal of the WP2 is to extend the previous classification to the case where the energy deposits of different particles overlap and it is desirable to split the events into different components. It has been demonstrated in [3] that it is possible to evaluate a probability of attribution of each hit to a secondary particle. Different levels of recognition will be explored, from the simple top particle identification to a more complete identification of the intermediate particles. These clusters can then be classified using an architecture like those developed in the WP1 to refine its nature and its parameters. The training of the different networks will be done on data-sets built from mixing single particles (the data set used for WP1) or more complicated simulations identifying all the intermediate secondary particles. This segmentation can not be applied to the online chain because its needs in computational resource are far beyond what is available in the current architecture.

This kind of algorithm can be of major importance in the context of a calorimeter where the different objects (electromagnetic showers, hadron jets…) are intricate and must be disentangled for reconstruction analysis to increase significantly the signal-background jet discrimination. Another interesting application would be the identification of the different components of a low energy neutrino event in Super-Kamiokande: the Cherenkov signal due to a low energy neutrino is so weak that only few photomultipliers receive the Cherenkov light making the identification of a Cherenkov ring impossible whereas natural background processes create noisy weak signals in many photo-multipliers at the same time.

This work package is divided in 4 tasks

T1 : Data production by single HGCal event merging

T2 : HGCal offline segmentation implementation and performance study

- Pytorch implementation based on WP1/T2 best choice for HGCal

T3 : TICL plugin implementation

T4 : Super-Kamiokande segmentation implementation and performance study

- Pytorch implementation based on WP1/T2 best choice for Super-K

Work Package 3 (WP3): Optimization and Quantization

In order to implement the GCN analysis in constrained environment, and in particular, the HGCal L1 trigger chain, it is interesting to measure what could be the minimal size for such an analysis system in terms of topology. To achieve this goal, we plan to use different optimization techniques, including the Bayesian optimization for continuous parameters or bandit optimization for categorical optimization. These techniques make it possible to explore the huge space described by the hyper-parameters (The parameters that are not learnt during the training process) of the neural network while minimizing the number of trainings. Such studies have already been performed at LLR on simplified neural architecture and they have lead to optimized architecture leading to perform analysis with very few resources. In this project, we plan to make such an optimization of the HGCal online GCN obtained in the WP1 to see if it could be synthetised for the current Trigger FPGA.

Another adaptation of the original architecture that is mandatory to obtain a synthetizable version is the integration of quantization. Indeed, the neural implementation in the FPGA are based on Digital Signal Processor (DSP) with limited precision compared to GPUs. Thus, it is necessary to train the network with quantized neurons that reproduces the exact behaviour available in the target electronics. This can be achieved by using quantized version of the neural network which is available in both Keras (Qkeras extension) and Pytorch (native). There are multiple models giving various results, in particular, binary and ternary [7] neural networks. A quantized version of the graph convolution will be tested extensively to study the loss of accuracy induced by this mechanism.

Finally, a study of implementability will be performed on this quantized and optimized network. If the resource available in the FPGA is sufficient, a block IP will be generated with HLS4ML and Vivado HLS, the high level description generation tool developped by the HEP community for neural network synthesis.

This work package is divided in two tasks:

T1: Optimization/Quantization of the network

- Implementation of quantized version based on Pytorch quantization capability.
- Performance study of the quantized implementation
- Optimization of the topology using Bayesian Optimization Library (LLR)

T2: Study of implementability and if possible synthesis of an IP block in collaboration with an expert in electronics

- Conversion of the network using HLS4ML framework
- Synthesis with Vivado HLS
- Test of the IP Block on the FPGA test-bench

Work Package 4 (WP4): Graph convolution parallelization

In the offline context, the parallelization of the GCN analysis is straightforward. The events are independent from each other and whatever the computational time needed to perform the analysis, another one can be run on another computing unit. This is not the case in the online trigger chain where the number of computational units is constrained. In the case insufficient resources are available on the FPGA to implement a full GCN analysis, it could be interesting to split this analysis. Thus, the last goal of this project is to develop a parallelized version of the online algorithm that can be distributed over different computational units. Typically, only a part of the data is available on each of them and a part of the computation is performed locally before a concentrating phase where the local results are combined to obtained a classification or a regression. This kind of architecture has been used successfully for desease detection [6] and can be adapted to the particularities of the particle detectors.

Note that this study will probably not be directly applicable to the HGCal online Trigger chain because the type of FPGA, type of links and numbers of boards will be decided at the end of 2021, with probably little flexibility afterwards. Nevertheless, this work is intended to be an element of reflexion for future implementation.

A model of geometric data distribution will be defined (typically a sector of the endcap in its whole depth). Then an architecture will be defined with a GCN implementation applied to every chunk of data and a concentrating neural network. A study of performance will be performed on simulated data to evaluate the loss of performance induced by the parallelization, compared to the non parallel implementation developed in the WP1.

It is composed of only one task

T1 : design of the parallel version and performance study

- Implementation using Pytorch
- Performance test of the parallel implementation

Bibliography

[1] Z. Li, W. Yang, S. Peng, F. Liu, A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects, abs/2004.02806, 2020

[2] Z. Zhang, P. Cui, W. Zhu, Deep Learning on Graphs: A Survey, abs/1812.04202, 2020

[3] S.R. Qasim, J. Kieseler, Y. Iiyama, and M. Pierini, Learning representations of irregular particle-detector geometry with distance-weighted graph networks. The European Physical Journal C, 79(7), abs/1902.07987, 2019.

[4] N. Choma, F. Monti, L. Gerhardt, T. Palczewski, Z. Ronaghi, M. Prabhat, W. Bhimji, M.M. Bronstein, S.R. Klein, and J. Bruna, Graph neural networks for Icecube signal classification. CoRR, abs/1809.06166, 2018.

[5] CMS Collaboration, The Phase-2 Upgrade of the CMS Endcap Calorimeter. Technical Report CERN-LHCC-2017-023. CMS-TDR-019, 2017

[6] A. Kazi, S. Albarqouni, K. Kortuem, N. Navab, Multi Layered-Parallel Graph Convolutional Network (ML-PGCN) for Disease Prediction. abs/1804.10776, 2018

[7] H. Alemdar, V. Leroy, A. Prost-Boucle, F. Pétrot, Ternary Neural Networks for Resource-Efficient AI Applications, abs/1609.00222, 2016

[8] John F. Beacom, The Diffuse Supernova Neutrino Background, Ann.Rev.Nucl.Part.Sci.60:439-462, abs/1004.3311, 2010

[9] Super-Kamiokande collaboration, Supernova Relic Neutrino search with neutron tagging at Super-Kamiokande-IV , Astroparticle Physics 60, 41–46, abs/1311.3738, 2015

[10] Y. Iiyama et al, Distance-Weighted Graph Neural Networks on FPGAs for Real-Time Particle Reconstruction in High Energy Physics, Frontiers in Big Data 3, 44, abs/2008.03601, 2021

[11] A. Di Pilato et al, Reconstruction in an imaging calorimeter for HL-LHC, JINST 15 06, C06023, abs/2004.10027, 2020