Activity Report

Dr. N. van der Kolk

Pattern recognition and Machine learning for imaging calorimeters

1. Project goals

Future lepton colliders with center-of-mass energies of around 1TeV will play a key role in understanding the origin of electroweak symmetry breaking.

New generation detectors for these lepton colliders will include high-resolution imaging calorimeters to precisely measure the trajectory and the energy of particles produced in lepton-lepton collisions. Such calorimeters will produce 4D data (three spatial dimensions and deposited energy) that will allow the determination of the topology of hadronic showers to unprecedented detail.

The main objective of the project is to improve the jet energy resolution of linear collider detectors by exploiting the wealth of information provided by high-resolution imaging calorimeters and by using state-of-the-art statistical and pattern recognition methods. The goal is to investigate the feasibility of using advanced machine learning and pattern recognition techniques to model hadronic showers and to automatically extract high level signal features of hadronic showers. Inferring parameters of hadronic showers can also be used for fine-tuning simulators used in detector studies and in astro-particle physics.

Two approaches are proposed to infer properties of hadronic showers; the generative approach and the direct approach. The formalization of a probabilistic generative model for a single hadronic shower will allow, using inference and machine learning techniques, to infer the generating parameters. One of the challenges will then be to distinguish the parameters of overlapping signals that are superpositions of several hadronic showers. The size and the dimensionality of the data and the complexity of hadron jets make the inference problem challenging both from the statistical and the computational point of view. In the direct approach, the goal is to find functions that infer the generative parameters directly, without going through the generative model. The main challenge is to deduce signal features that are informative about the generating parameters.

2. Description of work achieved

I have started on this project in October 2012 with an analysis of simulated data and test beam data recorded with the CALICE silicon-tungsten electromagnetic calorimeter prototype (Si-W ECAL). The CALICE collaboration is an international collaboration of about 280 physicists with the aim to design, construct and test highly granular calorimeters prototypes, exploring all available detector technologies.

The data for this analysis has been recorded at the Fermilab test beam facility in 2008. This analysis was started by Philippe Doublet during his PhD work and the preliminary results were reported in the form of a CALICE Analysis Note. I have revised, improved and extended the analysis with the aim to publish it in a peer reviewed journal.

The analysis determines the interaction point of negatively charged pions in the Si-W ECAL; approximately half of them are expected to interact within the Si-W ECAL volume. The interaction point is determined based on the increase in the energy deposition in subsequent layers of the detector. I then review the radial and longitudinal hit and energy distributions of different classes of events. These distributions are compared to predictions from Monte Carlo events generated within the simulation toolkit Geant4. I have made many changes compared to the preliminary analysis, e.g. improved quality and selection cuts, increased the number of studied distributions and added systematic errors. Especially the study to determine the systematic errors associated with each of the observables has been time consuming.

Due to the complicated nature of hadronic interactions a precise description of hadronic showers in simulations is difficult to achieve. In the simulation toolkit Geant4 several phenomenological hadronic interaction models are available and these are constantly being improved based on available real data. The CALICE test beam data with highly granular calorimeters is very important in that respect. In order for the final publication of the analysis to have significance with respect to the comparison with the Monte Carlo models, I have updated the whole analysis to the (at that moment) most recent version of Geant4. This has yielded interesting results, as some models had undergone a major revision compared to the previous version, and, as a result, do not describe the energy deposition in the Si-W anymore. The Geant4 developers are very interested in these results.

The analysis publication draft is at the moment under internal review in the CALICE collaboration. I am expecting to submit the publication to a journal at the end of this month (May 2014).

In parallel I have worked on extracting features of hadronic showers manually, such as the energy deposition per layer, the gradient of the energy and the radial extent of the shower per layer. These features served as input in a supervised learning algorithm as implemented by the AppStat group at LAL in the boosting software package MultiBoost. In supervised learning one trains the algorithm on a set of features and known classes, after which it can assign classes on the features in a second test set.

In the spring of 2013 two Master students have joined the effort on pattern recognition for a few months. Franck Dubard, under the supervision of Balàzs Kégl, worked on the implementation of deep learning algorithms. In contrast to supervised learning, deep learning will look at the raw data and determine the most successful features for classifying that data. These kinds of algorithms are very successful in image recognition. He has applied deep learning to the Si-W ECAL pion data successfully. Yaroslav Nikolaiko, under supervision of Roman Pöschl and myself, has been looking at the information that is available from Geant4 from which manual features can be constructed. In standard simulations all this detailed information is not saved, so the generator program has to be adjusted.

This spring, starting from February, again two Master students are part of the team, with the same set-up as the year before. Mehdi Cherti, under supervision of Balàzs Kégl, will continue on the deep learning algorithms. Sviatoslav Bilokin, under supervision of Roman Pöschl and myself, will continue with the exploration of manual features, that can be used in supervised learning algorithms. He will do this by exploring the full Monte Carlo information offered by Geant4.

I am developing a probabilistic model of hadronic interactions for the generative approach of the project. This generative model can be used with inference and machine learning techniques in the classification of hadronic interactions. This high-level model will also be a tool of communication between researchers of different formal backgrounds.

I have set up regular meetings between all the groups involved in the project, the ILC group at LAL, the ILC group at LLR and the AppStat group at LAL, in which we discuss progress and ongoing issues and next steps to be taken.

At the regular CALICE Collaboration meetings, twice a year, I have reported on my progress.

3. Publications

The publication "*Interactions of Pions in the CALICE Silicon-Tungsten Calorimeter Prototype*" is currently under review within the CALICE collaboration. This internal review will be finalized at the end of this month (May 2014). The publication will then be submitted to a peer reviewed journal, most likely the Journal of Instrumentation or Nuclear Instruments and Methods in Physics.

CALICE collaboration papers:

Validation of GEANT4 Monte Carlo Models with a Highly Granular Scintillator-Steel Hadron Calorimeter, arXiv:1306.3037, June 2013

Track segments in hadronic showers in a highly granular scintillator-steel hadron calorimeter, Xiv:1305.7027, July 2013

Shower development of particles with momenta from 1 to 10 GeV in the CALICE Scintillator-Tungsten HCAL, arXiv:1311.3505, January 2014

4. Relevance of the project within P2IO

This project falls in the P2IO theme "Symmetries in the subatomic world" and contributes to the technological goals by exploring data and simulations of novel detector techniques using sophisticated analysis tools. The multi-disciplinary character of the project requires a strong collaborative effort on both the physics and the computer science/statistics side. The project promotes collaboration between the P2IO institutes LAL and LLR as well as interdisciplinary collaboration at the intersection of experimental physics and computational statistics between the ILC group (LAL and LLR) and the Applied Statistics group at LAL.

The project is carried out within the CALICE collaboration, an international collaboration consisting of about 280 members. Regular reports on the progress of the project within this collaboration and the publication of results contribute to a high international visibility of P2IO.

7. Position after P2IO

I aim to continue to work on hadronic showers within the CALICE collaboration and am currently applying to relevant posts throughout Europe.